The Computer Analysis of Facial Expressions: On the Example of Depression and Anxiety

Gordon James McIntyre

A thesis submitted for the degree of Doctor of Philosophy

of the Australian National University

May 2010

School of Engineering

College of Engineering and Computer Science

The Australian National University

Canberra, Australia

Declaration

This thesis describes the results of research undertaken in the School of Engineering, College of Engineering and Computer Science, The Australian National University, Canberra. This research was supported by a scholarship from The Australian National University.

The results and analyses presented in this thesis are my own original work, accomplished under the supervision of Doctor Roland Göcke, Doctor Bruce Millar and Doctor Antonio Robles-Kelly, except where otherwise acknowledged. This thesis has not been submitted for any other degree.

Gordon McIntyre
School of Engineering
College of Engineering and Computer Science
The Australian National University
Canberra, Australia
10 May 2010

Acknowledgements

First of all I would like to thank the members of my supervisory panel Doctor Roland Göcke, Doctor Bruce Millar and Doctor Antonio Robles-Kelly. They have added invaluable insight from their respective areas which has been a big help in this multi-disciplinary project.

Roland, thank you for being an excellent supervisor and driving force, especially at times when it all seemed a bit too hard. You abound with positive energy and are blessed with shrewdness and patience well beyond your years. Bruce, I would like to thank you for your constructive criticism and the benefit of the wisdom that you accumulated over a distinguished career.

This PhD project would not have been as enjoyable without the support of staff and fellow students at the College of Engineering and Computer Science and my colleagues at the Centre for Mental Health Research. Thank you to all of you! I would like to also thank the administrative and support staff for putting up with my inane questions and providing prompt and professional assistance.

My gratitude goes to the Black Dog Institute in Sydney, it was the experience of a lifetime to be a part of such a multi-disciplinary team in an incredibly innovative organisation. It helped me to get an appreciation of the fantastic work they do in such a complex field.

In the course of this project, I have made many friends from a diverse range of backgrounds and my life is all the more richer for it. There are some special people that I need to acknowledge. To Dot, your spirit and determination is always an inspiration to me. To Judy, thank you for being so supportive and understanding. Lastly, to my children who were, oftentimes, deprived of my attention but nevertheless seemed to show an interest in my work - thank you!

Abstract

Significant advances have been made in the field of computer vision in the last few years. The mathematical underpinnings have evolved in conjunction with increases in computer processing speed. Many researchers have attempted to apply these improvements to the field of Facial Expression Recognition (FER).

In the typical FER approach, once an image has been acquired, possibly from capturing frames in a video, the face is detected and local information is extracted from the facial region in the image. One popular approach is to build a database of the raw feature data, and then use statistical measures to group the data into representations that correspond to facial expressions. Newly acquired images are then subjected to the same feature extraction process, and the resulting feature data compared to that in the database for matching facial expressions.

Academic studies tend to make use of freely available, annotated sets of images. These community databases, used for training and testing, are usually built from acted or posed expressions [Kanade 00, Wallhoff] of primary or prototypical emotion expressions such as fear, anger and happiness. Thus, making use of video or images captured in a natural setting is less common, and fewer studies attempt to apply the techniques to more subtle and pervasive moods and emotional states, such as boredom, arousal, anxiety and depression.

Overall, this tends to give an oversimplified picture of emotional expression. This is not to say that current research is oversimplifying the problem. It is acknowledgement that FER is a difficult problem. It should not, however, preclude investigation into the use of recently developed techniques that could be applied to non-primary, emotional expressions.

Of particular interest in this dissertation is the recent evolution in computer vision of the Active Appearance Model (AAM). These are used to locate fiduciary, or landmark, points, around a face in an image. If the landmark points can be reliably and consistently found within an image then the collective "shape" for the points, together with the pixel information, can be used to build representations of facial expressions.

This dissertation aims to test whether state-of-the-art developments in the field of computer vision can be successfully applied in a practical situation involving non-primary FER. The functional requirements of a system that can perform full lifecycle, video analysis of vocal and facial expressions are outlined. These have been used to build a fully-functional prototype system that incorporates AAMs. The system has been integral to supporting the experimental aspects of this dissertation.

Two experiments were undertaken. The first experiment investigated whether FER practices could be applied to sense for anxious expressions. A second experiment was conducted, analysing the facial activity and expressions displayed by patients diagnosed with Major Depressive Disorder (MDD).

Finally, the practical limitations of the statistical approach to FER are considered along with strategies for overcoming those limitations.

List of Publications

- G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear. *An Approach for Automatically Measuring Facial Activity in Depressed Subjects*. In 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, September 2009. DOI 10.1109/ACII.2009.5349593
- G. McIntyre and R. Göcke. *A Composite Framework for Affective Sensing*. In Proceedings of Interspeech 2008, pages 2767–2770. ISCA, 22-26 September 2008
- G. McIntyre and R. Göcke. Affect and Emotion in Human-Computer Interactions, chapter The Composite Sensing of Affect, pages 104–115. Lecture Notes in Computer Science LNCS 4868. Springer, August 2008
- G. Mcintyre and R. Göcke. *Towards Affective Sensing*. In Proceedings of the 12th International Conference on Human-Computer Interaction HCII2007, Volume 3 of *Lecture Notes in Computer Science LNCS 4552*, pages 411–420, Beijing, China, July 2007. Springer
- G. McIntyre and R. Göcke. *Researching Emotions in Speech*. In 11th Australasian International Conference on Speech Science and Technology, pages 264–369, Auckland, New Zealand, December 2006. ASSTA

Abbreviations

AAM Active Appearance Model

ANN Artificial Neural Network

ASM Active Shape Model

ASR Automatic Speech Recognition

AU Action Unit

AVI Audio Video Interleave

CA Classification Accuracy

CBT Cognitive Behaviour Therapy

DDL Description Definition Language

DFT Discrete Fourier Transform

EARL Emotion Annotation and Representation Language

EMG Electromyography

FACS Facial Action Coding System

FDP Facial Definition Parameters

FER Facial Expression Recognition

fMRI functional Magnetic Resonance Imaging

FR Face Recognition

FT Fourier Transform

GAD Generalised Anxiety Disorder

GSR Galvanic Skin Response

HMM Hidden Markov Model

HUMAINE Human-Machine Interaction Network on Emotion

IAPS International Affective Picture System

IEBM Iterative Error Bound Minimisation

k-NN k-Nearest Neighbour

LDA Linear Discriminant Analysis

MDD Major Depressive Disorder

MDS Multimedia Description Schemes

MPEG Moving Picture Experts Group

NXS Any Expression Recognition System

OCD Obsessive Compulsive Disorder

OO Object Oriented

PCA Principal Component Analysis

POIC Project-Out Inverse Compositional

PTSD Post-Traumatic Stress Disorder

ROC Receiver Operating Curve

RU Region Unit

SIC Simultaneous Inverse Compositional

SML Statistical Machine Learning

STFT Short Time Fourier Transform

SVD Singular Value Decomposition

SVM Support Vector Machine

VXL Vision *something* Libraries

XML Extensible Markup Language

Contents

List of Figures

List of Tables

A man's face as a rule says more, and more interesting things, than his mouth, for it is a compendium of everything his mouth will ever say, in that it is the monogram of all this man's thoughts and aspirations.

Arthur Schopenhauer

1 Introduction

Significant advances have been made in the field of computer vision in the last few years. The mathematical underpinnings have evolved, in conjunction with increases in computer processing speed. Many researchers have attempted to apply these improvements to the field of FER in a process, which typically resembles Figure 1.1.

Once an image has been acquired, possibly from capturing frames in a video, the face is detected and local information is extracted from the facial region in the image. One popular approach is to build a database of the raw feature data, and then use statistical measures to group the data into representations that correspond to facial expressions. Newly acquired images are then subjected to the same feature extraction

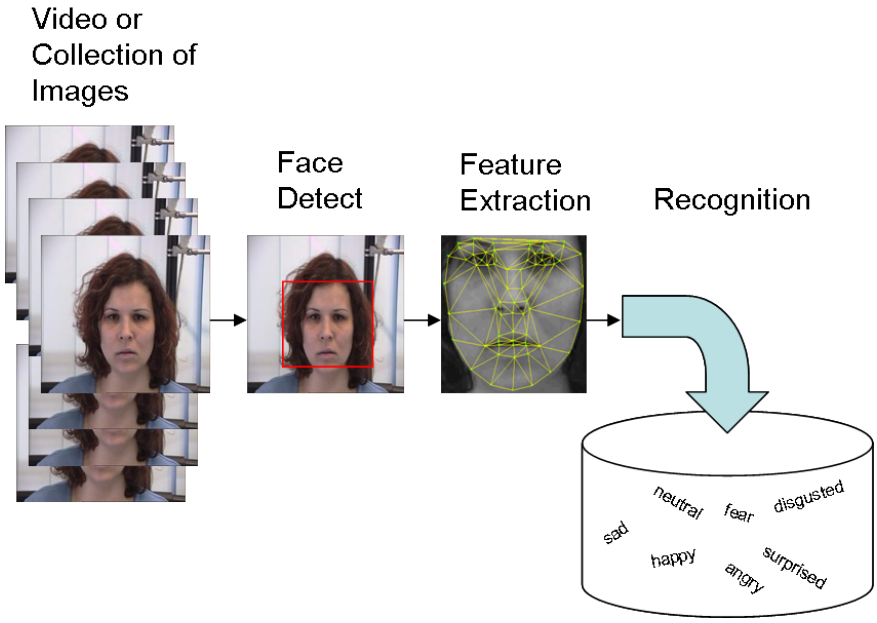


Figure 1.1: Affective sensing system

process, and the resulting feature data compared to that in the database for matching facial expressions.

Academic studies tend to make use of freely available, annotated sets of images. These community databases used for training and testing are usually built from acted or posed expressions [Kanade 00, Wallhoff], and this tends to give an oversimplified picture of emotional expression. Making use of video or images captured in a natural setting is less common.

Most studies aim to recognise prototypical emotion expressions such as fear, anger and happiness. Fewer attempt to apply the same techniques to more subtle and pervasive moods and emotional states, such as boredom, arousal, anxiety and depression.

The limitations placed on FER studies are quite understandable. [Ekman 82] has shown that the facial expressions of anger, disgust, fear, joy, sadness, and surprise are universal across human cultures (although in [Ekman 99] he did expand the list to include amusement, contempt, contentment, embarrassment, excitement, guilt, pride, relief, satisfaction, sensory pleasure and shame). Outside of the “unbidden” [Ekman 82]

emotions, the display rules of facial expressions vary with factors such as culture, context, personality type.

1.1 Motivation

The difficulties outlined above, however, should not preclude research into how the use of recently developed techniques could be applied to non-primary, emotional expressions. Several studies have confirmed characteristics such as speech, face, gesture, Galvanic Skin Response (GSR) and body temperature, as being useful in the diagnosis and evaluation of therapy for anxiety, depression and psychomotor retardation [Flint 93, Moore 08]. Vocal indicators have been shown to be of use in the detection of mood change in depression [Ellgring 96]. Some studies have suggested linking certain syndromes by comparing parameters from modalities, such as speech and motor, to discriminate different groups. In [Chen 03], the eye blink rate in adults was used in an attempt to diagnose Parkinson's disease. [Alvinoa 07] have attempted to link the computerised measurement of facial expression to emotions in patients with schizophrenia.

Thus, even if the results are to be used in conjunction with measurements of other modal expressions, e.g. vocal, eye-blink or gaze, there is good reason to explore the use of the recent developments in computer vision. One obvious incentive is to provide low-cost and unobtrusive ways to sense for disorders.

The motivation behind this dissertation is to test whether recent improvements in the field of computer vision could be used to verify the existence of states such as anxiety and depression.

1.2 Objectives

Of particular interest in this dissertation is the recent evolution in computer vision of the AAM [Edwards 98b]. These are used to locate fiduciary, or landmark, points around a face in an image. An example of the before and after "fitting" of the landmark points is shown in Figure 1.2.

(a) Original image (b) Image with landmark points

Figure 1.2: Original image on left and "fitted" image on right

If the landmark points can be reliably and consistent found within an image then the collective "shape" for the points, together with the pixel information, can be used to build representations of facial expressions.

The major objectives in this dissertation are outlined as follows:

1. Explore, through the construction of a prototype system that incorporates AAMs, what would be required in order to build a fully-functional FER system, i.e. one where the system could be trained and then used to recognise new and previously unseen video or images;

2. Investigate whether FER practices could be applied to non-primary emotional expression such as anxiety;

3. Examine whether FER practices could be applied to non-primary emotional expressions, such as those displayed by someone suffering from a MDD; and

4. Identify avenues for improvement in the emotional expression recognition process.

1.3 Thesis Outline

This dissertation comprises eight chapters, including this introduction. A brief outline of each of the other chapters is as follows:

Chapter 2 - Literature Review In the literature review, the controversial topic of defining emotions is discussed, followed by an explanation of some of the schemes used to describe emotions. This is followed by a brief description of the physiology contributing to emotional display. Next, the Facial Action Coding System (FACS), used ubiquitously to describe facial musculature activity, is introduced. The chapter concludes with a discussion of how dysphoric conditions, i.e. anxiety and depression, might affect emotional expression.

Chapter 3 - Emotional Expression Recognition by Machines *Affective sensing* is reviewed in Chapter 3. Schemes used to elicit speech samples (audio and video) are discussed along with their relative strengths and weaknesses. The chapter then narrows the focus to facial expression analysis, and a detailed view of the state-of-the-art components for analysing them in image and video is presented. The remainder of the chapter is devoted to the theoretical bases of the components, real-life applications, and their strengths and weaknesses.

Chapter 4 - Expression Analysis in Practice The practical exploration and experimental contribution in this dissertation is presented, firstly, with a discussion of

some of the capabilities that a full lifecycle, real-world, multi-modal affective sensing system might require, before turning, briefly, to describe the NXS system.

Chapter 5 - Sensing for Anxiety The experimental contribution of this thesis begins in Chapter 5. Using anxiety as an example, the exercise serves as a proof of concept for the techniques presented in Chapters 3 and 4, and to test whether more subtle expressions could possibly be tracked using these concepts. The experiments were conducted on the Cohn-Kanade [Kanade 00] database which is available for academic use. Each of the experiments involves different aspects and degrees of difficulty.

Chapter 6 - Depression Analysis Using Computer Vision Chapter 6 builds on the work from the previous chapter and describes an experiment that is currently incorporated in a collaborative project at the Black Dog Institute, Sydney, Australia.[1] The contribution is an exploration into applying the state-of-the-art, low-cost, unobtrusive techniques to measure facial activity and expression, and then applying them to a real-life application

Chapter 7 - Semantics in Expression Recognition With the results from Chapters 5 and 6 at hand, the reality and limitations of the statistical approach to facial expression analysis are considered, along with strategies to improve the field and overcome some of the limitations.

Chapter 8 - Conclusions Finally, the conclusions and a summary of the contributions of this dissertation are presented. Open issues and future directions for ongoing research are discussed.

[1]The Black Dog Institute is a not-for-profit, educational, research, clinical and community-oriented facility in Sydney, Australia, offering specialist expertise in depression and bipolar disorder. Available at `http://www.blackdoginstitute.org.au/`, last accessed 23 May 2010.

Even before I open my mouth to speak, the culture into which I've been born has entered and suffused it. My place of birth and the country where I've been raised, along with my mother tongue, all help regulate the setting of my jaw, the laxity of my lips, my most comfortable pitch.

Anne Karpf

2

Literature Review

2.1 Introduction

This chapter begins with a broad discussion of emotional expression. The accepted practices for its elicitation and description are discussed, before, narrowing the focus, and reviewing the research of expressions, and how they relate, more specifically, to anxiety and depression.

In Section 2.2, the somewhat difficult topic of defining emotions is broached, followed by an introduction to the annotation schemes commonly used. This is followed, in Section 2.3, by a brief description of the physiology contributing to emotional dis-

play. Next, in Section 2.4, FACS, the ubiquitous system for describing facial muscle activity, is introduced. The chapter concludes with a discussion of how dysphoric conditions, i.e. anxiety and depression, might affect emotional expression.

2.2 Describing Emotions

Defining emotion is a bit like trying to define "knowledge". It has deep ontological significance; it goes to the heart of human existence; yet there is no universal agreement on its definition. Whilst it is possible to observe physical symptoms arising from our internal state, attaining agreement on emotion definitions and categories is challenging. Taxonomies vary across disciplines, and [Cowie 03] point out that psychology, biology and ecology have different stances. There are qualitative and quantitative approaches.

Definitions in philosophy such as, "Emotion is evolution's way of giving meaning to our lives", might advance a philosophical view but do not translate neatly into computer science. Some authors suggest that nostalgia, jealousy and disappointment are emotions [Bower 92], while others propose a taxonomy of around five basic categories, e.g. Love, Joy, Anger, Sadness, and Fear, with all other emotions as subcategories. Many of the cross-disciplinary articles and books agree on a small set of emotions, e.g. fear, anger, sadness, happiness, sometimes disgust, sometimes surprise, and some include neutral. [Bower 92] "weed" the 600-plus items in the affective lexicon, removing behavioural responses, physical and body states, and short-hand expressions. Whissel [Whissell 89], on the other hand, has created a dictionary of thousands of affective language words that has been used to rate the affective content of the Bible, the Quran, Shakespeare, Dickens, Beowulf, and the works of several poets.

Complicating the picture even further is the question of moods. We know that moods are accompanied by physiological changes and affect our decisions, but are they emotions? The common view in the literature is that they are the long-term of

the affective state [Picard 97]. Emotions are seen as reactions [Bower 92, Cowie 03], having a cause or stimulus and a brief experience associated with them, whereas mood is seen as lingering and less specific. Moods have valance but less intensity, and emotions and mood can exist at the same time. Intuitively, one would think that they could affect one another, and, presumably, affect the valence of the emotion.

2.2.1 Emotion Classification Schemes

Traditional emotion theory is a large amalgam of approaches to the study of emotion, mostly based on cognitive psychology but with contributions from learning theory, physiological psychology, clinical psychology and other disciplines, including philosophy. Classification schemes have traditionally been sourced from these areas.

[Cowie 03] present a good review of emotional classification regimes. Two approaches to describing emotions are dominant. The first, to define categories, is the more common technique. The second approach uses dimensions. A third, less dominant, approach makes use of the appraisal theory [Sander 05].

The various classification schemes are discussed in more detail in the following subsections.

Category Approach

The most popular grouping of emotions, referred to as the “big six”, comprises fear, anger, happiness, sadness, surprise and disgust [Cornelius 96, Ekman 99]. These are regarded as full-blown emotions [Scherer 99] and are evolutionary, developmental and cross-cultural in nature [Ekman 82, Ekman 99]. However, there are many alternative groupings both across disciplines and within disciplines. Some studies concentrate on only one or two select categories, others employ schemes using more than twenty emotional archetypes. Thus, one of the difficulties in comparing results from studies

into emotions is that the choice of categories used between studies is not consistent and will often depend on the application that the researcher has in mind. If the focus of the research is to understand full-blown emotions, then the big-six or a subset might be adequate. However, if the objective is to study the less dramatic emotional states in everyday life, with all the shades and nuances that we know distinguish them, then the choice of categories is much more difficult [Schröder 05].

The main criticisms of the category approach are that:

- there are no agreed number of categories or definitions;
- a large number of descriptors exist, with overlapping meanings;
- there is a lack of consistency of words across languages; and
- the use of categories in research is inconsistent.

Dimensional Approach

Another way to label the affective state is to use dimensions. Instead of choosing discrete labels, one or more continuous scales, such as pleasant/unpleasant, attention/rejection or simple/complicated are used. Two common scales are valence (negative/positive) and arousal (calm/excited). Valence describes the degree of positivity or negativity of an emotion or mood; arousal describes the level of activation or emotional excitement. Sometimes a third dimension, control or attention, is used to address the internal or external source of emotion. [Cowie 03] have developed a software application called FEELTRACE to assist in the continuous tracking of emotional state on two dimensions [Cowie 03].

Appraisal Approach

Scherer has extensively studied the assessment process in humans and suggests that people affectively appraise events with respect to novelty, intrinsic pleasantness, goal/need significance, coping, and norm/self compatibility [Scherer 99].

It is not yet clear how to implement this approach in practice although there is at least one quite complex paper describing a practical application of the model [Sander 05].

Discussion of Description Schemes

Both the categorical and dimensional approaches, whilst practical, suffer from being highly subjective. This is not only due to complexity, but also because of the differences in the efficiency of listeners' physiology and the fact that the listener's own affective state influences their judgment of the speaker's affective state. Hence, there is a need for a model that includes listener attributes.

Describing More Subtle Emotions

Analysis of the Belfast Naturalistic Database, a corpus of emotional utterances, has shown that full-blown emotions occur more rarely in natural speech than is commonly assumed [Cowie 03].

A study by the Human-Machine Interaction Network on Emotion [HUMAINE 06] group proposed that emotions be dichotomised into *episodic* and *pervasive* categories. Episodic emotions are more like the traditional set of full-blown emotions or that which Scherer would describe as emotions. Pervasive are the everyday emotions such as grief/sorrow, sarcasm/irony and surprise/astonishment, but could include the more subtle states of anxiety and depression. Regardless of the method employed to describe emotions, adding more subtle emotions to any exercise, greatly complicates the task of describing emotional content in speech samples.

A recent study by [Devillers 05] has included labelling of blended and secondary emotions in a corpus of medical emergency call centre dialogues, as well as including task-specific context annotation to one of the corpora.

2.3 The Physiology of Emotional Display

Apart from the obvious, speech carries a great deal more information than just the verbal message. It can tell us about the speaker, their background and their emotional state. Age, gender, culture, social setting, personality and well-being all play their part in suffusing our communication apparatus even before we begin to speak. Studies by [Koike 98, Shigeno 98] have shown that subjects more easily identify an emotion in a speaker from their own culture, and that people will predominantly use visual information to identify the emotion. Everyday expressions such as "lump in the throat", "stiff upper lip", "plumb in the mouth", point to our awareness of the physiological changes that emotions have on the voice.

Early work from James [James 90] contended that emotions could be equated with awareness of a visceral response; in other words, the contention that emotions follow physical stimuli. That may be true for fast primary emotions, however, the twentieth century view is that emotions are antecedent, and can be more often detected from physiological measurements. For example, your heart rate goes up when you discover that you have won lotto, you think that you have lost your ATM card, or you realise that you have forgotten an important birthday.

The neurobiological explanation of human emotion is that it is a pleasant or unpleasant mental state organised in the limbic system. Recent studies establish that emotional stimuli is given priority, or a privileged status, within the brain [Davidson 04]. *Primary emotions* such as fear use the limbic system circuitry along with the amygdala and anterior cingulate gyrus. *Secondary emotions*, take a slightly different path to take

in memory. The stimulus may still be processed directly via the amygdala, but is now also analysed in the thought process before processing.

Changes in brain patterns result in modulations in our major anatomical systems. Stress tenses the laryngeal muscles, in turn, tightening the vocal folds. The result is that more pressure is required to produce sound. Consequently, the fundamental frequency and amplitude, particularly with regard to the ratio of the open to the closed phase of the cycle, varies the larynx wave. The harmonics of the larynx wave vary according to the specific balance of mass, length and tension that is set up to produce a given frequency [Fry 79].

Some affective states like anxiety can influence breathing, resulting in variations in sub-glottal pressure. Drying of the mucus membrane causes shrinking of the voice. Rapid breath alters the tempo of the voice and relaxation tends to deepen the breath and lower the voice. Changes in facial expression can also alter the sound of the voice.

Figure 2.1 represents the typical cues to the six most common emotion categories [Murray 93].

Darwin raised the issue of whether or not it was possible to inhibit emotional expression [Ekman 03]. This is an important question in human emotion recognition and in emotion recognition by computers. Intentional or not, the voice and face are used in everyday life, to judge verisimilitude in speakers. Many studies [Anolli 97] [Hirschberg 05] have investigated the detection of deception in the voice.

2.4 Describing Facial Activity

Although some studies have made use of MPEG-4 compliant Facial Definition Parameters (FDP) [Cowie 05b], the most ubiquitous and versatile method of describing facial behaviour, pioneered by Ekman [Ekman 75, Ekman 82, Ekman 97, Ekman 99, Ekman 03], is FACS. The goal of FACS is to provide an accurate description of facial activity based

	fear	**anger**	**sorrow**	**joy**	**disgust**	**surprise**
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	Breathy, blaring	grumble chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	slurring	normal	normal	

Figure 2.1: The effect of emotion on the human voice [Murray 93]

on musculature, and to lead to a system of consistent and objective description of facial expression.

Despite being based on musculature, FACS measurement units are AUs, not muscles. One muscle can be represented by a single AU. Conversely, the appearance changes produced by one muscle can sometimes appear as two or more relatively independent actions attributable to different parts of the muscle. A FACS coder decomposes an observed expression into the specific AUs that produced the movement, recording the list of AUs that produced it. For example, during a smile, the *zygomaticus major* muscle is activated, corresponding to AU12. During a spontaneous or Duchenne smile, the *orbicularis oculi* muscle is recruited, corresponding to AU6. The FACS coder records the AUs, and if needed, the duration, intensity, and asymmetry. Figure 2.2 and Table 2.1, together, are included to give a description of the facial muscles and an exam-

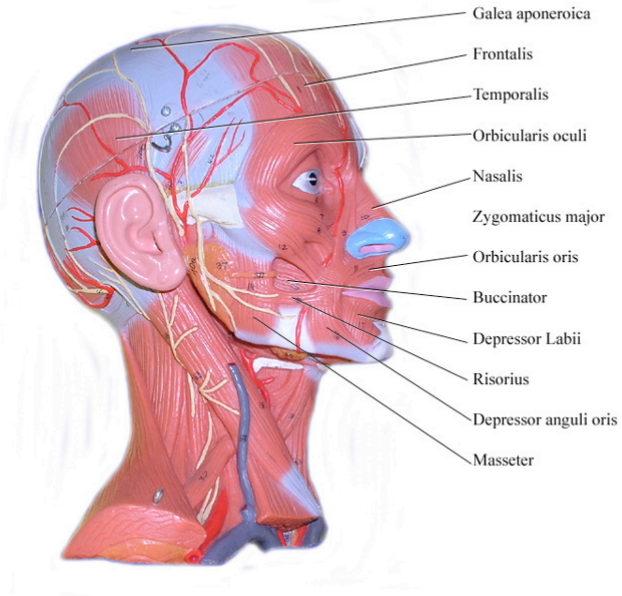


Figure 2.2: Facial muscles [Muscles]

ple of some common AUs. FACS scores, in themselves, do not provide interpretations. EMFACS, as mentioned previously, deals only with emotionally relevant facial action units.

2.5 Anxious Expression

2.5.1 Anxiety Disorders

There are many types of inflictions that are labelled as "anxiety", e.g. test anxiety (a type of performance anxiety), death anxiety and stage fright. Typically, these have short-term impact and, depending on the level of arousal, may or may not affect a person's performance. It is when anxiety begins to affect someone's day to day life, that it is classed as a *disorder*. An *anxiety disorder* is an umbrella term used to cover (at least) six types of disorder [beyondblue , NIMH]:

AU	Description
1	Inner Brow Raiser – Frontalis (pars medialis)
2	Outer Brow Raiser – Frontalis (pars lateralis)
4	Brow Lowerer – Corrugator and Depressor supercilii
5	Upper Lid Raiser – Levator palpebrae superioris
6	Cheek Raiser – Orbicularis oculi (pars orbitalis)
7	Lid Tightener – Orbicularis oculi (pars palpebralis)
9	Nose Wrinkler – Levator labii superioris alaeque nasi
10	Upper Lip Raiser – Levator labii superioris
11	Nasolabial Deepener – Zygomaticus minor
12	Lip Corner Puller – Zygomaticus major
13	Cheek Puffer – Levator anguli oris
14	Dimpler – Buccinator
15	Lip Corner Depressor – Depressor anguli oris
16	Lower Lip Depressor – Depressor labii inferioris
17	Chin Raiser – Mentalis

Table 2.1: Facial Action Coding System - Sample AUs

- Generalised Anxiety Disorder (GAD);
- social anxiety disorder;
- phobia;
- Obsessive Compulsive Disorder (OCD);
- Post-Traumatic Stress Disorder (PTSD); and
- panic disorder.

In the discussion that follows, the social anxiety disorder has been included along with the phobia disorder.

GAD

GAD is usually diagnosed with reference to an instrument such as the *Diagnostic and Statistical Manual of Mental Disorders, 4th ed.* (DSM IV) [Kvaal 05]. Broadly speaking, someone who has felt anxious for at least six months and it is adversely affecting

their life, will meet the criteria. The anxiety might be associated with issues such as finances, illness or family problems. The adverse impact might include factors such as insomnia, missed work days or fatigue.

[beyondblue] report that GAD affects approximately 5 per cent of people in Australia at some time in their lives.[1] Diagnosing GAD can be difficult for a clinician as the symptoms are shared with other types of anxiety and it often coexists with other psychiatric disorders, e.g. depression or dysthymia (chronic, "low-grade" depression).

The symptoms of GAD are so broad that it would be difficult to imagine it ever being capable of detection by machine.

Phobia

The most common phobias are:

- acrophobia - fear of heights;
- agoraphobia - fear of open spaces such as parks and big shopping centres;
- claustrophobia - fear of small spaces such as lifts, aeroplanes and crowded rooms;
- mysophobia - fear of dirt and germs in places such as toilets and kitchens.;
- social phobia or *social anxiety* - fear of social situations such as parties and meetings; and
- zoophobia - fear of animals.

beyondblue report that, "...approximately 9 per cent of people in Australia experience a phobia at some time in their lives. Phobias are twice as common in women as in men and can start at any age." [beyondblue]"

[1] *beyondblue* is a national, independent, not-for-profit organisation working to address issues associated with depression, anxiety and related substance misuse disorders in Australia.

Anxious episodes of the types listed above are the easiest to induce (ethically and practically) and are the most common candidates for facial expression recognition experiments.

OCD

OCD symptoms are characterised by uncontrollable, ongoing ritualistic and intrusive thoughts and behaviors that the sufferer feels compelled to perform. The thoughts and behaviors are irrational but, even so, become obsessions that affect the person's everyday life [beyondblue].

Typical compulsions include:

- cleaning or hand-washing;
- checking things repeatedly, e.g. that appliances are turned off or that doors and windows are locked;
- constantly counting or checking the order or symmetry of objects;
- superstitions about colours or numbers; and
- hoarding items such as newspapers, books, or clothes.

Performing the rituals gives the sufferer short term relief from the anxiety. OCD affects 2 to 3 per cent of people in Australia at some time in their lives [beyondblue].

The varied and sometimes serious nature of the compulsions and situations make OCD an unlikely candidate for facial expression recognition.

PTSD

PTSD is the emotional trauma following a very stressful event such as war, accident, assault or an illness [beyondblue]. The symptoms of PTSD can include:

- flashbacks and nightmares;
- insomnia;
- loss of interest or enjoyment in life;
- difficulty concentrating; and
- amnesia.

Approximately 8 per cent of people in Australia are affected by PTSD at some time in their lives [beyondblue]. PTSD has been the subject of some interesting virtual reality applications for rehabilitation. In 2010, the U.S. Army began a four-year study to track the results of using virtual reality therapy to treat Iraq and Afghanistan war veterans suffering PTSD.[2]

The serious nature of this type of anxiety would mean that there would be some important ethical and patient-care considerations to be met before any study is undertaken.

Panic Disorder

A panic attack is an episode of such an intense feeling of anxiety that it seems like it cannot be brought under control. Panic attacks can occur in short bursts when someone is stressed, or an initial attack can lead to fears of other attacks at a later stage. This can result in a vicious cycle where the person is constantly worried about the next attack.

Technically, if a person experiences a panic attack at least four times a month, they may be diagnosed as having a panic disorder. Around 3 per cent of the Australian population has experienced a panic disorder [beyondblue].

In many panic attack situations, sufferers believe that they are suffering a heart-attack or nervous breakdown. This type of anxiety would make an interesting research

[2] `http://www.army.mil/`, last accessed 10 April 2010

topic, but one would think that there would be some quite restrictive ethical considerations.

2.5.2 Anxious Facial Expressions

Numerous studies in the past twenty years have confirmed characteristics such as speech, face, gesture, galvanic skin response (GSR) and body temperature, as being useful in the diagnosis and evaluation of therapy for anxiety, depressions and psychomotor retardation [Flint 93]. Some earlier studies have suggested linking certain syndromes by comparing parameters from modalities such as speech and motor to discriminate different groups. [Chen 03] attempted to link eye blink rate in adults to Parkinsons disease.

Anxiety is sometimes confused with fear, which is a reaction normally commensurate with some form of imminent threat. In the case of anxiety, the perceived threat is usually in the future and the reaction tends to be irrational or out of proportion to the threat. The facial expressions of fear and anxiety, however, are similar [Harrigan 96].

Confounding the problem is that the facial expression of surprise is similar to that of fear. One key difference between the fearful and the surprise expression is the mouth movement - a fearful expression involves a stretching of the lips in the horizontal direction rather than opening of the mouth. The AUs for fearful expressions are shown in Table 3.2.

Relatively little research has been conducted to definitively map which action units are associated with anxiety. While available research generally supports the efficacy of human ability to judge anxiety from facial expressions [Harrigan 96, Harrigan 97, Harrigan 04, Ladouceur 06], understandably, due to logistical considerations, much of the work has been conducted within the confines of social anxieties and in specific situations such as dental treatment [Buchanan 02], examinations, public speaking, children

Emotion	Prototype	Major Variants
Fear	1 +2+4+5*+20*+25 1+2+4+5*+25	1+2+4+5*+L or R20*+25, 26, or 27 1+2+4+5* 1+2+5Z, with or without 25, 26, 27 5*+20* with or without 25, 26, 27

* means in this combination the AU may be at any level of intensity
L - Left, R - Right
Action Unit 1 (Inner Brow Raiser), Action Unit 2 (Outer Brow Raiser)
Action Unit 5 (Upper Lid Raiser)
Action Unit 20 (Lip Stretcher)
Action Unit 25 (Lips part Jaw Drop)

Table 2.2: Action units for fear expressions [Kanade 00]

receiving immunisations, medical examinations [Buchheim 07] and human-computer interactions [Kaiser 98]. [Harrigan 96] reports:

> "First, people reveal feelings of anxiety facially in the form of actions composing the expression of fear rather than other affects thought to compose anxiety (distress, interest). These actions were not the full fear expression of widened, tense eyes, raised and drawn brows, and horizontal mouth stretch typically displayed in more intense situations described as fearful [Ekman 71]. Rather, partial actions of the fear expression involving the mouth or brows were exhibited and corresponded to the degree of fear-anxiety experienced by the participants (i.e. moderate anxiety)."

Of importance, [Harrigan 96] found that the brow movement exhibited when fear is experienced, i.e. brows raised and drawn together, was displayed by the participants in the study, but less often than the mouth movement for fear. They summarise:

> "The most predominant fear element was the horizontal mouth stretch movement. This horizontal pulling movement of the mouth and brief tension of the lips was clearly visible on the videotapes and could not be confused with movements required in verbalization, smiling, or other facial action units. The brow movement exhibited when fear is experienced,

> brows raised and drawn together, was displayed by the participants in this study, but less often than the mouth movement for fear."

The finding of [Kaiser 98], while studying human-computer interactions, was that AU20 (lip stretcher) is found more often only in fear. [Ellgring 05] noted that actor portrayals provide little evidence for an abundance of distinctive AUs or AU combinations, that are specific for basic emotions. [Ellgring 05] reports that there is quite a distribution of AUs through each emotion - including anxiety. Muddying the waters is the fact that there can be large variations in the way individuals react to stressful stimuli. Genetics, personality and biochemistry factors all play a part in the propensity to display an anxious expression.

Trait anxiety is a longer-term, predisposition to anxiety whereas *state anxiety* is the short-term anxiety induced by some recent event. These are defined in the clinical practitioner's Spielberger State-Trait Anxiety Inventory [Kvaal 05]. In a meta-analysis of 46 state anxiety studies and 34 studies on trait anxiety, [Harrigan 04] conclude that state anxiety was recognised by observers with greater accuracy than was trait anxiety, but the modality was important. State anxiety was identified best from auditory signals, whereas trait anxiety was identified best from visual (video) signals. Intuitively, this makes some sense.

[Lazarus 91] concludes that a cognitive appraisal of threat is a prerequisite for the experience of this type of emotion. If this is the case, then this does not augur well for the ability to detect anxiety in a system trained from acted expressions. That is, one would question how well actors could portray anxious expressions without a threat stimulus.

2.6 Non-verbal Communication in Depression

As in the previous section, this chapter outlines the affects on facial expressions. Unlike the previous section, it is expanded to include the processing of emotional content by patients with MDD. This is in order to provide a background to the experimental work in Chapter 6.

Early attempts to link facial activity with depression used broad measurements such as cries, smiles, frowns [Grinker 61]. Some studies have used Electromyography (EMG) to measure muscle response, notwithstanding the somewhat intrusive and constraining nature of the equipment [Fridlund 83]. The more recent trend is to use the FACS to add rigour and objectivity to the process [Reed 07, Renneberg 05].

The difficulty is that capturing and recording measurements of facial activity manually, requires time, effort, training, and regime for maintaining objectivity. Such manual work is tedious and is prone to errors. However, even the abridged version of FACS, EMFACS [EMFACS], which deals only with emotionally relevant facial action units, requires a scoring time of approximately 10 minutes of measurement for one minute of facial behaviour. Further, only people who have passed the FACS final test are eligible to download EMFACS for use.

2.6.1 Ellgring's Study

In [Ellgring 08], the levels of facial activity, before and after treatment, of endogenous and neurotic depressives were measured through several key indicators. [Ellgring 08] hypothesised that facial activity and the repertoire of its elements will be reduced during depression and will expand with improvement of subjective wellbeing. Facial behaviour was analysed by applying EMFACS to the videotapes of 40 clinical interviews of 20 endogenous depressed patients. After analysing a frequency distribution of all of the AU observations across all of the interviews, 13 AUs or groups of AUs were

then used to complete the study. AUs that nearly always occur together, e.g. AU1 and AU2 (inner and outer brow raiser), AU6 and AU12 (cheek raiser and lip corner puller) were considered as one AU group. Activity, repertoire and patterns were defined as parameters of facial activity and can be summarised as the following three measurements:

- **General facial activity:** The first measurement, total number of AUs in a specific interval, counts the number of single, significant AUs and groups of closely related AUs that occur within a 5 minute interval.
- **Specific facial activity:** The second, frequency of specific, major AUs, defines the number of major AU combinations, e.g. AU6+12 in the case of a spontaneous smile, occurring within a 5 minute interval.
- **Repertoire:** The third, repertoire of AUs, is the number of distinct AUs occurring more than twice within a 5 minute interval.

2.6.2 Processing of Emotional Content

Depressed subjects have been shown to respond differently to images of negative and positive content, when compared with non-depressed subjects. The underlying cause could be the impaired inhibition of negative affect, which has been found in depressed patients across several studies [Goeleven 06, Lee 07]. In turn, altered patterns of facial activity have been reported in those patients suffering MDD [Reed 07, Renneberg 05]. If this "affective facial processing loop", as shown in Figure 2, is reliable, then 1) objective observations could be of clinical importance in diagnosis; and 2) measurements of facial activity could possibly predict response to treatments such as pharmacotherapy and Cognitive Behaviour Therapy (CBT).

As stated previously, it is commonly hypothesised that depression is characterised by dysfunctional inhibition toward negative content. The reason for this has been pos-

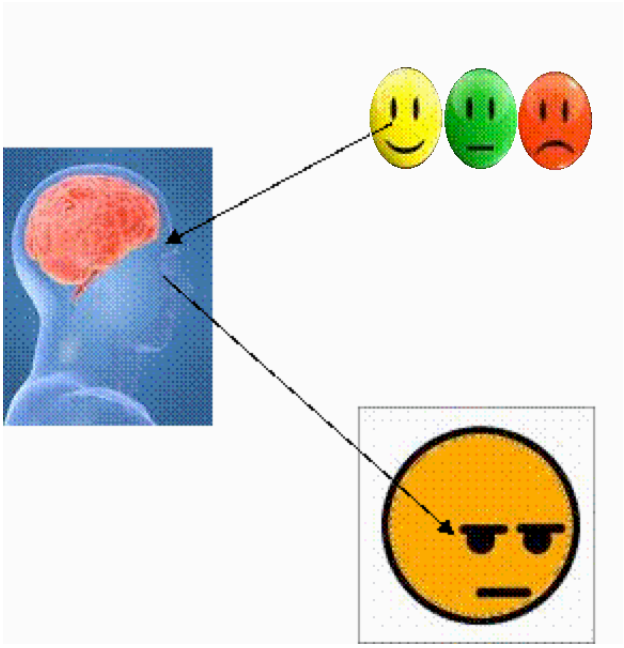


Figure 2.3: The affective facial processing loop

tulated that depressed subjects show lowered activation in the regions responsible for gaining attentional control over emotional interference. In [Goeleven 06], inhibition to positive and negative stimuli was studied across subjects including hospitalised depressed patients, formerly depressed, and never-depressed control subjects. They report that depressed patients show a specific failure to inhibit negative information whereas positive information was unaffected. Surprisingly, they report that formerly depressed subjects display impairment to inhibiting negative content.

[Joormann 07] similarly found attentional bias was evident even after individuals had recovered from a depressive episode. In that study, the attentional biases in the processing of emotional faces in currently and formerly depressed participants and healthy controls were examined. Faces expressing happy or sad emotions paired with neutral faces were presented (in a dot-probe task). Currently and formerly depressed participants selectively attended to the sad faces, the control participants selectively avoided the sad faces and oriented toward the happy faces. They also report that a

positive bias that was not observed for either of the depressed groups.

In an evaluation of the evidence in studies which have used modified Stroop and visual probe test, [Mogg 05] has found that the "inhibition theory" only holds true if the material is relevant to their "negative self-concept" and the stimulus is presented for longer durations. [Joormann 06] found that depressed participants required significantly greater intensity of emotion to correctly identify happy expressions, and less intensity to identify sad than angry expressions.

Medical imaging studies also find impairment of neural processing of depressed subjects. [Fu 08] conducted a comparison of 16 participants who had suffered from acute unipolar major depression and 16 healthy volunteers. The patients received 16 weeks of CBT. functional Magnetic Resonance Imaging (fMRI) scans were undertaken at weeks 0 and 16 while the participants viewed facial stimuli displaying varying degrees of sadness. Although there are some limitations to the study due to sample size, there seems to be evidence that excessive *amygdala* activity correlated to the processing of sad faces during episodes of acute depression.

This impairment may extend to the offspring of parents with MDD. [Monk 08] found (small volume corrected) greater *amygdala* and *nucleus accumbens* activation to fearful faces, and lower *nucleus accumbens* activation to happy faces, in high-risk subjects when attention was unconstrained.

2.6.3 Facial Feedback in Depression

In a study of 116 participants (30 men, 86 women), some with a history of MDD and individuals with no psychopathological history, the smile response to a comedy clip was recorded. Participants were asked to rate a short film clip. FACS coding was applied to 11 seconds of the clips - long enough to allow for the 4-6 second spontaneous smile [Frank 93]. Those with a history of MDD and current depression symptoms were

more likely to control smiles than were the asymptomatic group [Reed 07].

There is always something ridiculous about the emotions of people whom one has ceased to love.

Oscar Wilde

3

Emotional Expression Recognition by Machines

3.1 Introduction

This section provides a summary of the more recent approaches and developments in the field of emotional expression recognition. The scope is limited to FER, although it would be incomplete without some reference to recognising vocal expression. The objective of the chapter is to provide a theoretical grounding for later, more practical chapters.

This chapter is organised as follows:

Section 3.2 introduces the broad concepts found in affective sensing systems. This is quite a large section and encompasses the elicitation of training data, an overview of the typical processing and a comparison of approaches to feature extraction. Section 3.3 describes the computer vision techniques that are of particular interest in this dissertation.

3.2 Affective Sensing Systems

Affect is emotional feeling, tone, and mood attached to a thought, including its external manifestations. *Affective communication* is the, often complex, multimodal interplay of affect between communication parties (which potentially includes non-humans). Much of our daily dose of affective communication constitutes small talk, or phatic speech. Recognising emotions from the modulations in another person's voice and facial expressions is perhaps one of our most important human abilities. Yet, it is one of the greatest challenges in order for machines to become more human-like. *Affective sensing* is an attempt to map manifestations or measurable physical responses to affective states.

Historically, attempts at *Affective Sensing* or emotional expression recognition systems were based on the audio modality and had their origins in Automatic Speech Recognition (ASR). The usual tact is to first acquire samples of vocal or facial expressions, use them to train some form of system and then to test the system's recognition against some newly introduced samples. In practice, for development purposes, the training and the testing samples are often taken from the same set and recognition performance is measured using a *K-fold* or *leave-one-out* cross-validation method over the sample collection [Dellaert 96, Kohavi 95, Yacoub 03].

Whether it be a vocal or a facial expression recognition system, the conceptual

system usually resembles Figure 3.1. Whilst systems may use different types of classifiers, e.g. k-NN [Cover 67], SVM [Vapnik 95], AdaBoost [Freund 99], the main differences surround the feature extraction process, i.e. the number and type of features used; and, whether they incorporate rule-based logic such as that presented by Pantic and Rothkrantz [Pantic 00].

One major difference between vocal and facial expression is that speech signals are inherently one-dimensional, whereas facial signals can be 2D or 3D - although the use of 3D processing has only become popular in the last few years [Lucey 06]. In the case of facial expression recognition, an additional differentiating factor is whether holistic (spanning the whole or a large part of the face) features or, what Pantic terms, "analytic" (sub-regions) of the face are used [Pantic 07].

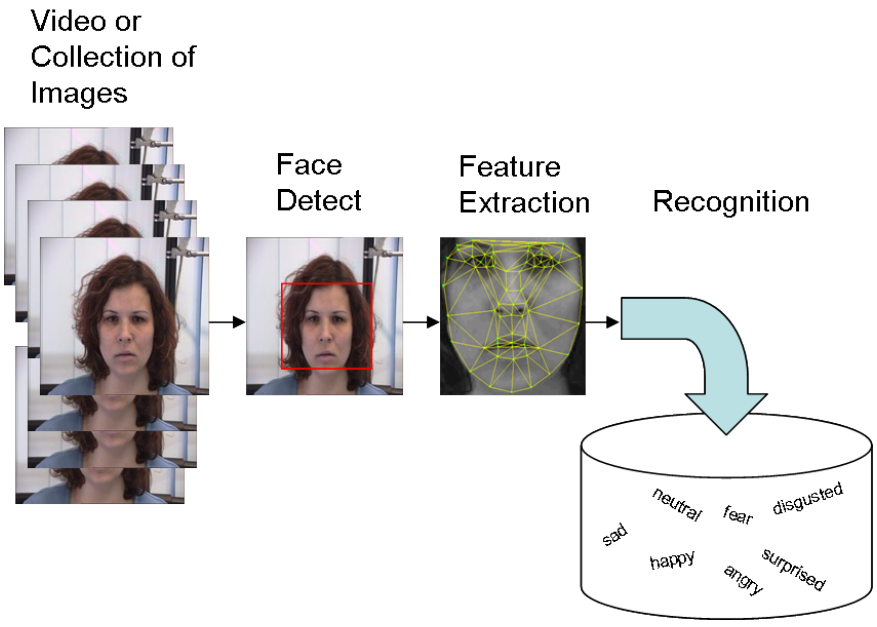


Figure 3.1: Affective sensing system

3.2.1 Eliciting Training Data

[ten Bosch 00] and [Schröder 05] have explored what is possible in extending the ASR framework to emotion recognition. The common ASR approach is to train probabilistic models from extracted speech features and then to use pattern matching to perform recognition.

Most studies into affective communication begin with the collection of audio and/or video communication samples. The topic of collection of emotional speech has been well covered by other reviews [Cowie 03, Cowie 05a, Scherer 03], so it is only briefly summarised here.

Naturally Occurring Speech

To date, call centre recordings, recordings of pilot conversations, and news readings have provided sensible sources of data to research emotions in speech. Samples of this nature have the highest ecological validity. However, aside from the copyright and privacy issues, it is very difficult to construct a database of emotional speech from this kind of naturally occurring emotional data. In audio samples, there are the complications of background noise and overlapping utterances. In video, there are difficulties in detecting moving faces and facial expressions. A further complication is the suppression of emotional behaviour by the speaker who is aware of being recorded.

Induced Emotional Speech

One technique introduced by Velten [Velten 68], is to have subjects read emotive texts and passages which, in turn, induce emotional states in the speaker. Other techniques include the use of Wizard of Oz setups where, for example, a dialog between a human and a computer is controlled without the knowledge of the human. This method has the benefit of providing a degree of control over the dialogue and can simulate a natural

setting [Hajdinjak 03].

The principal shortcoming of these methods is that the response to stimuli may induce different emotional states in different people.

Acted Emotional Speech

A popular method is to engage actors to portray emotions. This technique provides for a lot of experimental control over a range of emotions and like the previous method provides for a degree of control over the ambient conditions.

One problem with this approach is that acted speech elicits how emotions should be portrayed, not necessarily how they are portrayed. The other serious drawback is that acted emotions are unlikely to derive from emotions in the way that [Scherer 04] describe them, i.e. episodes of massive, synchronised recruitment of mental and somatic resources to adapt or cope with a stimulus event subjectively appraised as being highly pertinent to the needs, goals and values of the individual.

3.2.2 Approaches to Affective Sensing

The affective sensing of facial expression is more commonly known as automatic FER. FER is somewhat similar in approach to face recognition but the objectives of the two are quite different. In the former, representation of expressions is sought in sets of images, possibly from different people, whereas in the latter, discriminating features are sought, which will distinguish one face from a set of faces. FER works better if there is large variation among the different expressions generated by a given face, but small variation in how a given expression is generated amongst different faces [Daugman]. Nevertheless, both endeavours have common techniques.

Broadly speaking, there are two approaches to automatic facial expression recognition. The first is an *holistic* one, in which a set of raw features extracted from an image are matched to an emotional facial expression such as happy, sad, anger, pain, or possibly to some gestural emblem such as a wink, stare or eye-roll [Ashraf 09, Liu 06, Martin 08, Okada 09, Saatci 06, Sung 08]. The second approach, an *analytic* one, is more fine-grained and the face is divided into regions. The most popular scheme for annotating regions is through the use of the FACS [Ekman 75, Ekman 82, Ekman 97, Ekman 99, Ekman 03], in which surface or musculature movements are tracked using an analytical framework.

Computer vision techniques [Cootes 95, Lucey 06, Nixon 01, Sebe 03, Sebe 05] are used to detect features and build evidence of FACS AUs [Bartlett 99, Bartlett 02, Bartlett 03, Bartlett 06, Lien 98, Lucey 06, Tian 01, Valstar 06b]. The theoretical benefit of this approach relates to its purported reusability. For example, if a system can accurately detect 20 major FACS AUs then, in theory, any facial expression that has a repertoire using a combination of the 20 AUs can be detected.

However, in practice, this is not as straightforward as it might seem. For instance, the AU combinations for surprise and fear are very similar and the AUs are shown in Figures 3.1 and 3.2:

Emotion	Prototype	Major Variants
Surprise	1+2+5B+26	1+2+5B
	1+2+5B+27	1+2+26
		1+2+27
		5B+26
		5B+27

Table 3.1: Action units for surprise expressions [Kanade 00]

So which approach is correct? Regardless of the approach, testing and comparing

Fear Prototype	Major Variants	
Fear	1 +2+4+5*+20*+25 1+2+4+5*+25	1+2+4+5*+L or R20*+25, 26, or 27 1+2+4+5* 1+2+5Z, with or without 25, 26, 27 5*+20* with or without 25, 26, 27

* means in this combination the AU may be at any level of intensity.

Table 3.2: Action units for fear expressions [Kanade 00]

the claims of FER research reports is difficult. It is not like Face Recognition (FR) which tends to be a binary classification problem, and a Receiver Operating Curve (ROC) makes visual comparison of reports quite simple (even though it might not be the best instrument). FR vendors can even evaluate their software against the US government "Face Recognition Vendor Test" [FRV]. FER, on the other hand, has a lot more subjectivity and variability with regards to the expressions being recognised. Even human judges may not agree or be able to correctly identify a facial expression.

One of the problems with the *holistic* approach in comparing and validating results is that each published report tends to use a different set or subset of classes or expressions. Most experiments use a closed set of expressions and it is unknown how well the reported systems would work, if at all, when an extraneous expression, or indeed non-expression, is introduced. For example, if the training and testing database consists of 2,000 images split evenly with happy, sad, angry and neutral expressions, the article might report a 93% accuracy rate. However, if the database is then interspersed with 500 other expressions (perhaps surprise, fear or yawn), it is likely that the system in question will try to find the best match to one of the classes.

That is not to say that the *analytic* approach is an automatic choice either. Accredited FACS coders do not always achieve consensus on AU displays. Not all AUs are easily detectable in an image, and the ability of the systems are impacted by the quality of the recordings. *Holistic* classification might have an advantage where fea-

tures are occluded; where, for instance, the subject has a beard, wears spectacles or is in a non-frontal pose. However, if reliable detection of the muscle movements could be guaranteed, one would think, that the *analytic* approach would be a better option to advance research.

There is a lacuna in objective reporting instruments. To address the study of comparison and validation problems and to advance the field of research, a standard reporting system needs to be introduced.

3.2.3 Description of Process

Whilst there are some strategies in recognition systems, e.g. some systems operate on profile poses [Pantic 04a, Pantic 04b] and in [Pantic 04a] the system that incorporates case-based reasoning, most follow the fairly generic model as depicted in Figure 3.2.

If the system processes video then the first stage is to capture some or all of the images from the video at predefined intervals. Next, faces are usually detected and then segmented from the images and this is certainly the case with the *AAM* approach discussed later.

The most common method of face detection is that formulated by Viola and Jones [Viola 01], which is implemented in the popular "openCV" software [OpenCV] from Intel. [Bartlett 05] report an improved face detection implementation using the GentleBoost algorithm.

The next stage, facial feature extraction, is where most variation between systems arises. Optical Flow, Particle Filters and Gabor Filters, and more recently, Active Appearance Models are some of the choices, with the latter two the most widely reported in recent years. Since Gabor filters and Active Appearance Models are used in this research work, they are discussed in more detail in Section 3.3 onwards. Sometimes the techniques are combined as in [Gao 09] and in several instances, especially

in the use of Gabor wavelets, dimension reduction is attempted through some pre-processing such as Boosting, Principal Component Analysis (PCA) or SVM classification [Chen 07, Shen 07]. It is this stage where most research activity is taking place, as it is critical to the success of facial expression recognition.

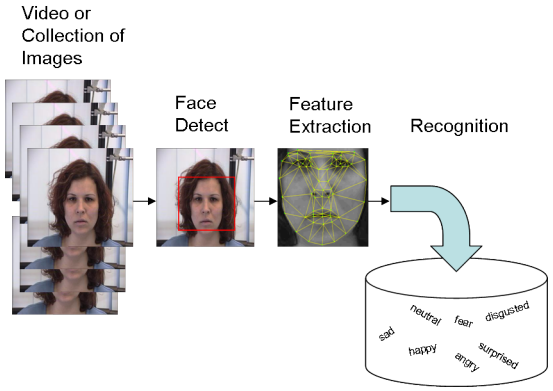


Figure 3.2: Processing of facial activity measurements

Finally, once the features have been extracted, the next stage is to adopt some means of classification such as k-Nearest Neighbour (k-NN), Artificial Neural Network (ANN), SVM, or AdaBoost. If the temporal patterns are to be classified, or the facial patterns combined with other signals, e.g. vocal speech, then the solution may involve ensembles of classifiers or Hidden Markov Model (HMM).

3.3 Computer Vision Techniques

Whilst there have been attempts to provide a top-down categorisation of automatic facial feature extraction methods, the lines between the categories are quite blurred. Pantic and Rothkrantz [Pantic 00] attempt to summarise the approaches, on one hand, into “analysis from static facial images” and “analysis from facial image sequences”. On the other hand, they also split the methods into “Holistic approach”, “Analytic approach” and “Hybrid approach”. They introduce further terms “Template-based” meth-

ods, which align with the "Holistic approach", and "Feature-based" methods, which align with the "analytic approach".

Fasel and Luettin [Fasel 03] provide an appealing dichotomy of facial feature extraction methods into *Deformation Extraction*, either image or model-based, and *Motion Extraction*. From a different viewpoint, they follow on from Pantic and Rothkrantz [Pantic 00] and dichotomise into "holistic methods", where the face is processed in its entirety, or "local methods" (similar to Pantic and Rothkrantz's "Analytic approach"), which analyse only areas of interest in the face, especially where transient facial features involved in expressions exist (as opposed to intransient features such as furrows, wrinkles and laughter lines). Of course, the selection of local feature tends to be very application dependent.

3.3.1 Gabor Filters

The *Discrete Fourier Transform (DFT)*, commonly used in speech processing, can be applied to rows and columns of pixels in an image, indexed by co-ordinates *x* and *y*. The 2D DFT of an *N* x *N* pixel image can be given by

$$\mathbf{FP}_{u,v} = \frac{1}{N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \mathbf{P}_{x,y} e^{-j(\frac{2\pi}{N})(ux+vy)} \tag{3.3.1}$$

where *u* and *v* are the horizontal and vertical dimensions of spatial frequency respectively.

However, *Fourier Transform (FT)* and DFT are not well suited to non-stationary signals, i.e. signals with time-varying spectra such as spikes. Both perform decimation in frequency *across the entire image* in the forward transform and decimation in time in the inverse transform. This can be overcome to some extend through the *Short Time Fourier Transform (STFT)*, which uses a window function to divide the signal

into segments.[1] However, the inherent problem with STFT is in choosing the width of the window. Too narrow a window will give good time resolution but poor frequency resolution, and too wide a window will give good frequency resolution but poor time resolution. If the spectral components in the input signal are already well separated from each other, then a narrow window will possibly provide a satisfactory frequency resolution. In the case of the frequency components being tightly packed, then a narrow window will be needed, resulting in good time resolution but poor frequency resolution.

Wavelet transformations offer a solution to this resolution problem by delivering decimation in frequency and space simultaneously. Although they are performed in a similar fashion to STFT, i.e. the function (*wavelet*) is applied to different segments of the time domain signal, the signal is analysed at different resolutions, i.e. the width of the window is computed for every frequency.

Invented by Dennis Gabor in 1946, *Gabor wavelets* have found many applications including speech analysis, handwriting, fingerprint, face and facial expression recognition. One reason for their popularity in computer vision is that, in their 2-D form and primed with the appropriate values, their filter response resembles the neural responses of the mammalian primary visual cortex [Daugman 85, Bhuiyan 07, Lee 96, Gao 09]. Just as bandpass filter bank's ability to approximate cochlear processing and ANN ability to analogue human neural processing appealed to researchers, this biological resemblance has also attracted much attention.

A spatial 2D complex Gabor function is given by the formula

$$g(x, y) = s(x, y)w_r(x, y) \tag{3.3.2}$$

[1] *window* is the term used when referring to the *continuous-time* STFT, whereas, *frame* is used in *discrete-time* STFT. For explanatory purposes, only continuous-time STFT is referred to. However, the same principles apply.

where $s(x, y)$ is a complex sinusoidal referred to as the *carrier*, and $w_r(x, y)$ is a Gaussian-shaped function, known as the *envelope*, modulated by the sinusoidal.

The *carrier* is defined as

$$s(x, y) = \exp(j(2\pi(u_0 x + v_0 y) + P)) \tag{3.3.3}$$

where (u_0, v_0) and P define the spatial frequency and the phase of the sinusoidal, respectively [Movellan 08].

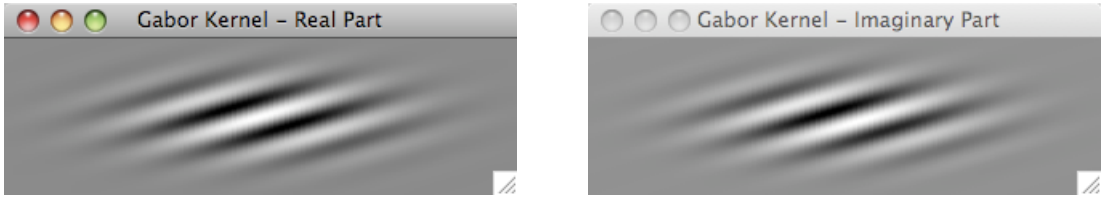
The complex sinusoidal can be split into its real and imaginary parts as

$$\mathrm{Re}\,(s(x, y)) = \cos(2\pi(u_0 x + v_0 y) + P) \tag{3.3.4}$$

and

$$\mathrm{Im}\,(s(x, y)) = \sin(2\pi(u_0 x + v_0 y) + P) \tag{3.3.5}$$

The real and imaginary parts of a Gabor with an orientation of $\pi/4$ and a scale of 4 is shown in Figure 5.6.



(a) Real part of wavelet (b) Imaginary part of wavelet

Figure 3.3: Real and imaginary part of a Gabor wavelet

This spatial frequency can be expressed in polar coordinates as magnitude F_0 and direction ω_0

$$s(x, y) = exp(j(2\pi F_0(x\cos\omega_0 + y\sin\omega_0) + P)) \tag{3.3.6}$$

The *envelope* is defined as

$$w(x,y) = Kexp(-\pi(a^2(x-x_0)_r^2 + b^2(y-y_0)_r^2)) \tag{3.3.7}$$

where K scales the magnitude of the Gaussian envelope, (x_0, y_0) is the peak of the function, *a* and *b* are scaling parameters, and the *r* subscript is a rotation operation such that

$$(x-x_0)_r = (x-x_0)\cos\theta + (y-y_0)\sin\theta \tag{3.3.8}$$

and

$$(y-y_0)_r = -(x-x_0)\sin\theta + (y-y_0)\cos\theta \tag{3.3.9}$$

The response of a Gabor filter to an image is obtained by a 2D convolution operation. Let $I(x,y)$ denote the image and $G(x,y,\theta,\phi)$ denote the response of a Gabor filter with frequency and orientation to an image at point (x,y) on the image plane. $G(.)$ is obtained as

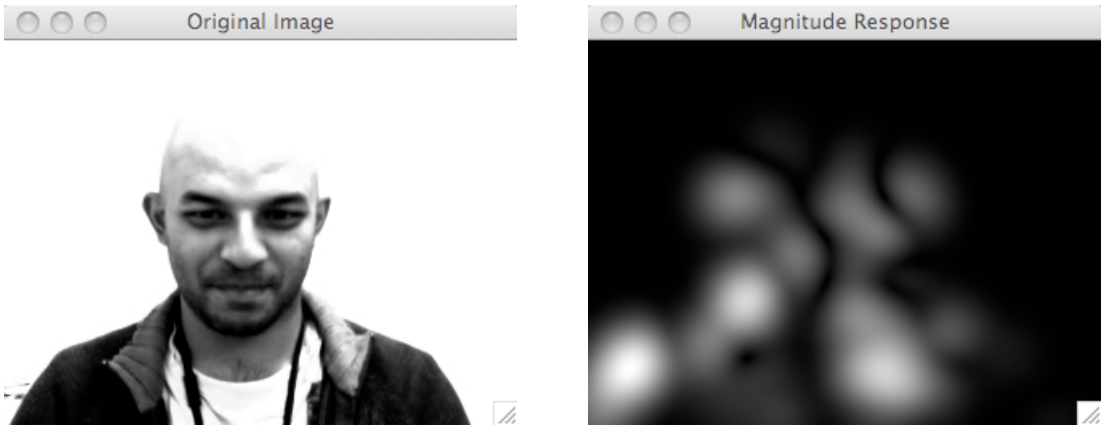
$$G(x,y,\theta,\phi) = \int\int I(p,q)g(x-p, y-q), \theta, \phi)\, dp\, dq \tag{3.3.10}$$

Figure 3.4 shows the original image on the left and the magnitude response image on the right.

3.3.2 Active Appearance Models (AAM)

3.3.3 Introduction

In recent years, a powerful deformable model technique, known as the *AAM* [Edwards 98b], has become very popular for real-time face and facial expression recognition. The literature is not very clear in what exactly AAMs are and how they should be differentiated from other approaches that represent the appearance of an object by shape and texture



(a) Original image (b) Magnitude after transform

Figure 3.4: Original image with Gabor magnitude

subspaces, hence some explanation follows.

According to [Saragih 08], AAMs are examples of Linear Deformable Model (LDM) which also includes Active Shape Models (ASM) [Cootes 92], AAMs [Cootes 98] and 3D Morphable Models (3DMM) [Blanz 99]. According to [Matthews 04], AAMs, together with the closely related concepts of Morphable Models and Active Blobs, are "generative models of a certain visual phenomenon" and, "are just one instance in a large class of closely related linear shape and appearance models and their associated fitting algorithms". In [Gross 05], they are defined as, "generative parametric models commonly used to model faces".

In the AAM approach, the non-rigid shape and visual texture (intensity and colour) of an object (a face in an image perhaps) are statistically modelled using a low dimensional representation obtained by applying PCA to a set of labelled training data. After the models have been created, they can be parameterised to *fit* a new object (of similar properties), which might vary in shape or texture or both.

Usually, the AAMs are pre-trained on static images using a method such as the Simultaneous Inverse Compositional (SIC) algorithm [Baker 01, Baker 03b] (discussed in Subsection 3.3.5) and then, when ready to be applied, the model will be *fitted* to one

or more images that were not present in the training set.

For the better understanding of the following sections, some terms are introduced and explained now.

Shape All the geometrical information that remains when location, scale and rotational effects are filtered out from an object - invariant to Euclidian similarity transformations [Stegmann 02].

Landmark point Point of correspondence on each object that matches between and within populations.

Shape space Set of all possible shapes of the object in question.

Texture The pattern of intensity or colour across the region of the object or image patch [Cootes 01].

Image Registration This is the process of finding the optimal transformation between a set of images in order to get them into one coordinate system.

Image Segmentation Segmentation is used to separate an object in an image from the background. In practice, it is the process of separating objects, as represented by sets of pixels, in an image.

Image Warping Image warping is a type of geometric manipulation of an image or region of an image. It is the *non-uniform* mapping of a set of points in one shape to a set of points in another shape.

Fitting An efficient scheme for adjusting the model parameters so that a synthetic example is generated, which matches the image as closely as possible.

Model A model consists of the mean positions of the points and a number of vectors describing the modes of variation [Cootes 01].

3.3.4 Building an AAM

AAMs usually refer to two things:

1. A statistical model of shape and appearance, trained from a set of images, each of which has a set of corresponding landmark points (usually manually annotated); and once built,

2. A method or algorithm for fitting the model to new and previously unseen images, i.e. images that were not in the set of images that were used to train the model;

although sometimes they are used simply to mean the statistical model. It should be noted that there are actually two types of AAM:

1. *independent* shape and appearance models, where the shape and appearance are modelled separately; and

2. *combined* shape and appearance models, which use a single set of parameters to describe shape and appearance [Matthews 04]

The reason for the different build strategies is to do with variations in the algorithms that are used to "fit" the model to an image (discussed in Section 3.3.5).

The steps to building an AAM are covered in the following sections.

Annotate the Landmark Points

Although there have been some attempts at automatically annotating images [Asthana 09, Tong 09], there is no completely automated method for facial feature extraction and at least one image has to be manually marked up as shown in Figure 5.4.

For each image, a corresponding set of points (each point denoting the x, y co-ordindates of a single point) exists as shown in table 3.3.

(a) Original image (b) Image with landmark points

Figure 3.5: Original image with transform

n_points: 69
{
249.809 274.693
249.785 297.994
259.769 361.231
...
328.853 393.34
305.365 280.317
431.243 281.514
}

Table 3.3: Sample point file with x, y co-ordinates of landmark points

Align the Set of Points

A generalised Procrustes analysis is shown in Algorithm 3. The mean shape, if using the *Procrustes mean*, is used is given by:

$$\overline{x} = 1/N \sum_{i=1}^{N} x_i \tag{3.3.11}$$

Algorithm 1: Aligning the training set

foreach *image in the training set* **do**

Translate so that its centre of gravity is at the origin

Choose one example as an initial estimate of the mean shape and scale it so that $|\overline{x}| = 1$

Record the first estimate as $\overline{x}_0$ to define the default reference frame

repeat

Align all the shapes with the current estimate of the mean shape

Re-estimate mean from aligned shapes

Apply constraints on the current estimate of the mean by aligning it with $\overline{x}_0$ and scaling so that $|\overline{x}| = 1$

until *convergence, i.e. until the estimate of the mean* $\leq \epsilon$*, where* ϵ *is some suitable threshold*

However, one of the undesirable side effects of the Procrustes analysis, due to the scaling and normalisation process, is that aligned shapes, or their shape vectors, will now lie on the curved surface of a hypersphere, which can introduce non-linearities. A popular approach to counter this problem is to transform or modify the shape vectors into a *tangent space* to form a hyper plane [Cootes 01]. That way, linearity is assumed and, the next step, "modelling the shape variation" is simplified and calculation performance improved.

Model the Shape Variation

At this point, a set of points exist that are aligned to a common co-ordinate frame. What remains to be done is to somehow model the point distributions, so that new and plausible shapes can be generated. It is best to start with a reduction in dimensionality and this is most commonly performed using PCA, which will derive a set of t eigenvectors, Φ, corresponding to the largest eigenvalues that best explain the data (spread

of the landmark points). The normal procedure is to first derive the mean shape as in 3.3.11.

The next step is then to compute the covariance matrix

$$\Sigma_x = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T \tag{3.3.12}$$

Since computer vision applications are typically resource intensive, alternative ways to derive the eigenvectors, Φ, and corresponding eigenvalues, λ, of the covariance matrix have been devised and the choice depends on whether there are more samples, s, than dimensions, n, in the feature vectors. If $s > n$ then Singular Value Decomposition (SVD) is can be used.

Any set of points, x, can then be approximated by

$$x \approx \overline{x} + \Phi b \tag{3.3.13}$$

where $\Phi = (\phi_1|, \phi_2, ..., \phi_t)$ and b is a t dimensional vector given by

$$b = \Phi^T(x - \overline{x}) \tag{3.3.14}$$

The total variation in the training samples can be explained by the sum of all eigenvalues. Limiting the number of eigenvectors, e.g. $\pm 3\sqrt{\lambda_i}$ of the parameter b_i will determine how close generated shapes will be to the training set.

All that remains is to find a way to model the spread of the variation or distribution around the points. A Gaussian is a reasonable starting point but for facial feature processing, where there are likely to be non-linear shape variations due, for example, to pitch, yaw and head roll, a better solution is required. A Gaussian mixture is a reasonable approach.

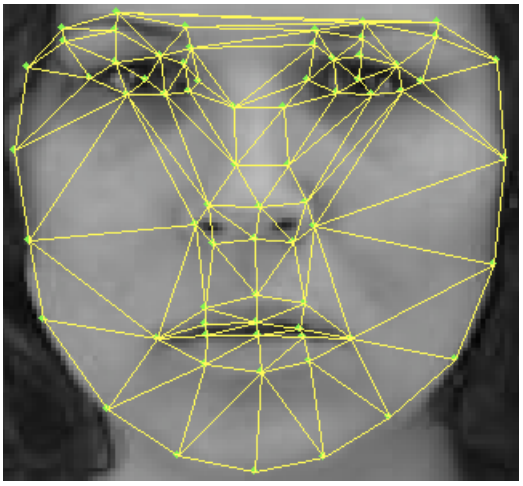


Figure 3.6: Face mesh used to build AAM

Build a Combined Appearance Model

Having already obtained a mean shape, each image texture is now warped to the mean shape, obtaining what is described as a "shape-free" patch or "canonical frame". This is done by ensuring that each image's control points match the mean shape. A triangulation algorithm that creates a mesh, such as that shown in Figure 3.6 is used. Next, the image texture over the warped image is sampled to obtain a vector of texture, $\mathbf{g}_{im}$.

The texture vectors are normalised by applying a linear transformation (not shown here but discussed in [Cootes 01]). Applying PCA to the normalised training samples produces a linear model of texture

$$\mathbf{g} = \overline{\mathbf{g}} + P_g b_g \tag{3.3.15}$$

where $\overline{\mathbf{g}}$ is the mean normalised grey-level vector, P_g is a set of orthogonal models of variation and b_g is a set of grey-level parameters.

Shape and texture can now be expressed in terms of its shape-level parameters b_s (Equation 3.3.14) and its grey-level parameters b_g. However, a further PCA is applied to account for correlations between shape and grey-level texture and finally appearance

model shape and texture can be controlled by a vector of parameters c

$$x = \overline{x} + Q_s c \tag{3.3.16}$$

and

$$\mathbf{g} = \overline{\mathbf{g}} + Q_g c \tag{3.3.17}$$

where c is a vector of appearance parameters controlling both the shape and the grey-level texture, and Q_s and Q_g are matrices describing modes of variation.

Reconstruction of an image is almost the reverse process of building the appearance model. The texture is generated within the mean-shaped patch and then warped to suit some image points that have been generated by applying a transformation (scale, rotation and translation) from the image frame.

3.3.5 Model Fitting Schemes

Two fitting schemes are described, since they are used in this research work. Until now, the building of a statistical model of shape and texture has been discussed. Although there has been some notion of generating shapes and texture, the actual process of *fitting* or adjusting model parameters to build a synthetic model, in an attempt to match it to a new and previously unseen image, i.e. one not in the model training set, has not been covered. Although it is a generalisation, state-of-the-art recognition systems are heavily dependent on improvements in the area of model *fitting*, as it is crucial to achieving the performance required for a real-time FER.

There are many variations and performance improvements (some application specific) that can be made to AAMs. Searching schemes can employ shape only, appearance only, or combined shape and appearance.

Simultaneous Project-out Inverse Compositional Method

[Baker 01, Baker 02, Baker 03a, Baker 03b, Baker 04a, Baker 04b] treat the AAM search process as an image alignment problem. The original image alignment algorithm was formulated by Lucas and Kanade in 1981 [Lucas 81]. The goal is to minimise the difference between an image and a template image $T(x)$ by minimising

$$\sum_{x}[I(W(w;p)) - T(x)]^2 \tag{3.3.18}$$

where I is the image, W is the warp, x are the pixels in the template and p is a vector of parameters.

The fine details go beyond the scope of this thesis. However, a concise explanation of [Baker 01] is that the formulation to solve Equation 3.3.18 requires minimising

$$\sum_{x}[I(W(w;p+\Delta p)) - T(x)]^2 \tag{3.3.19}$$

using the algorithm

Algorithm 2: Algorithm to minimise Equation 3.3.19

repeat

- Warp I with $W(x;p)$ to compute $I(W(x;p))$
- Compute the error image $T(x) - I(W(x;p))$
- Warp the gradient of image I to compute $\Delta I)$
- Evaluate the Jacobian $\frac{\delta W}{\delta p}$
- Compute the Hessian matrix
- Compute Δp
- Update the parameters $p \leftarrow p + \Delta p$

until *convergence*

As can be seen, algorithm 2 necessitates the re-evaluation of a Hessian and a Jacobian at every iteration until convergence, which is a very expensive operation. [Baker 01, Baker 02] propose an analytically-derived gradient descent algorithm called the *inverse compositional* algorithm, in which the roles of the image and the template are reversed and which allows both the Hessian and the Jacobian to be pre-computed and then held constant when iterating the remainder of the algorithm.

[Matthews 04] provide a further efficiency improvement by "projecting out" appearance variation. Using this technique implies a major difference in the way that the AAMs are initially built - the shape and the appearance parameters need to be modelled separately - *independent AAMs* previously mentioned in Subsection 3.3.4.

The SIC algorithm [Baker 03a] is an another adaptation of inverse compositional image alignment for AAM fitting that addresses the problem of significant shape and texture variability by finding the optimal shape and texture parameters simultaneously. Rather than re-evaluating the linear update model at every iteration using the current estimate of appearance parameters, it can be approximated by evaluating it at the mean appearance parameters, allowing the update model to be pre-computed, which has significantly more computational efficiency.

Iterative Error Bound Minimisation (IEBM)

This is a form of what [Saragih 08] describes as "Iterative Discriminative Fitting". Attempts at improving the "fitting" efficiency of AAMs, such as those discussed at Subsection 3.3.5, involve various techniques to streamline the original Lucas-Kanade algorithm [Lucas 81], and essentially aim to minimise a least squares error function over the texture. [Matthews 04] provide experimental evidence to show that the *Project-Out Inverse Compositional (POIC)* fitting method provides very fast fitting. However, when there is a lot of shape and appearance variation, it comes at the expense of poor generalisability, i.e. in the case of facial expression recognition the model becomes

very person-specific. Changing the algorithm to improve generalisability then impacts performance.

To overcome these problems, [Saragih 06, Saragih 08] pre-learn the update model by minimising the error bounds over the data, rather than minimising least squares distances. Conceptually, this is akin to *boosting* [Freund 99] where a bunch of weak classifiers are iteratively passed over a training set of data and, for each iteration, a distribution of weights is updated so that the weights of each incorrectly classified example are increased, thereby resulting in a *strong* classifier. However, to continue with the analogy, in the case of IEBM, when a new weak classifier is introduced after each iteration, as the data used in calculating the weak learners changes, it is subject to what the authors describe as "resampling" [Saragih 06]. It is this resampling process that promotes generalisability.

3.4 Classification Techniques

Machine learning classification finds its way into many aspects of research and there are many options available to take the features extracted from say, a bank of Gabor filters and classify them. The most frequently used include ANN, k-Nearest Neighbour (k-NN), SVM and AdaBoost (although a variant for a multi-class problem is Multi-Boost). Only SVM and AdaBoost are discussed in this dissertation.

3.4.1 Support Vector Machines

SVMs are a set of supervised learning methods that are suitable for classification in high or infinite dimensional space. A hyperplane or set of hyperplanes is projected through a set of point, which is to be used for classification, regression.

Although fairly recent, SVMs [Burges 98, Chen 05, Vapnik 95] are well understood and have been successfully implemented in academic and commercial data min-

ing. SVM generalization performance depends on the appropriate setting of meta-parameters C and the kernel gamma parameter γ.

3.4.2 Adaboost

AdaBoost [Freund 99] is a "meta-algorithm" that calls a set of "weak" classifiers repeatedly, eventually yielding a strong classifier. The key concept of the algorithm is that a set of weights is maintained over an input training set. Initially the weights are set equally, and after each round of classification using the "weak" learner, the weights of the misclassified examples are increased, thus forcing the weak learner to focus on the hard examples in the training set. Although intended for binary classification, AdaBoost can be extended to the multiclass problem and the implementation from [Casagrande 06] exists.

There are no facts, only interpretations.

Friedrich Nietzsche

4

Expression Analysis in Practice

4.1 Introduction and Motivation

A system capable of interpreting affect from a speaking face must recognise and fuse signals from multiple cues. Building such a system necessitates the integration of such tasks as image registration, video segmentation, facial expression analysis and recognition, speech analysis and recognition.

Without the availability of vast sums of time and money to build the components "from the ground up", this almost certainly entails re-using publicly available software. However, such components tend to be idiosyncratic, purpose-built, and driven

by scripts and peculiar configuration files. Integrating them to achieve the necessary degree of flexibility to perform full multimodal affective recognition is a serious challenge.

If one contemplates the operation of a full-lifecycle system that can be trained from audio and video samples and then used to perform multimodal affect recognition, the requirements are extensive and diverse. For example, to detect emotion in the voice the system must be capable of training, say, HMM from prosody in the speech signals.

Another requirement might be that a SVM be trained to recognise still image facial expressions, e.g. fear, anger, happiness, sadness, disgust, surprise or neutral. More complex, is the requirement to capture a sequence of video frames and, from the sequence, recognise temporal expressions. In order to perform the latter, it might be necessary to use a deformable model, e.g. an AAM [Edwards 98a] to fit to each image and provide parameters that can then, in turn, be trained using some classifier - possibly another HMM.

Other features might also be considered for input to the system, e.g. eye gaze and blink rate. Ultimately, some strategy is required to assess the overall meaning of the signals, whether it involves fusion using a combined HMM or some other technique.

From the concise consideration of the requirements it can be seen that a broad range of expertise and software is needed. It is not practical to develop the software from first principles. Software capable of the recognition of voice and facial expressions implement techniques from different areas of specialisation. ASR techniques has evolved over decades while computer vision has become practical in the last ten years, with the evolution of statistical techniques and computer processing power.

The reasons behind choosing one software product over another is not within the scope of this work. A brief overview of some of the critical components and the "composite framework" used to harness them is presented.

This chapter is organised as follows:

Section 4.2 discusses of the functional requirements of a system capable of sensing multiple variable inputs from voice, facial expression and movement, making some assessment of the signals of each, and then fusing them to provide some degree of affect recognition. Section 4.3 describes how the key requirements have been implemented in the *NXS*, which has been built to support the experimental aspects of this dissertation.

Although, due to time constraints, the main focus in this dissertation and the experimental work is on facial expression recognition, the NXS system has been designed to support multi-modal analysis and recognition.

4.2 Functional Requirements

There are several levels of sophistication that a system capable of sensing affect could provide:

1. recognition of affect from audio only;
2. recognition of affect from image only;
3. recognition of affect from video without audio; and
4. recognition of affect from video with audio.

Indeed, the system should be able to operate within each modality or across a conflation of modalities. The ultimate is to be able to recognise emotion from both audio and video inputs from a speaking face. Consider a speaking face in a real world situation. Voice expression is not necessarily continuous, there may be long pauses or sustained periods of speech. Vocal speech and facial expression may not necessarily be contemporaneous, the verisimilitude of the voiced expression might be confirmed or contradicted by the facial expressions. The face might be expressionless, hidden,

not available or occluded to some degree for certain periods of time. This implies that the system needs to be able to:

1. detect the voice and facial expressions independently;

2. operate on only one modality in some cases; and

3. weigh one signal against the other when more than one modality is available.

Lastly, the system must be flexible so that alternative software products and techniques can be substituted without a large amount of effort or re-engineering work involved. For example, to compare classification performance, it might be desirable to substitute an ANN package for an SVM package or simply compare different types of SVM implementations.

The following sections present a minimalist requirement statement of some of the key individual functional areas.

4.2.1 Audio Processing

The most common approach to recognising affect in speech is to use existing ASR software and to try to detect prosody from the energy levels and variance in the signals. There are several ASR packages freely available, some with better support than others. Whatever the choice, this capability is mandatory.

4.2.2 Image Processing

Before a facial expression in an image can be analysed it is necessary to have software capable of first detecting the face. Once found, facial features need to be extracted, e.g. by fitting an AAM to the face. This will yield parameters that can be used in the classification process.

A major issue in recognising facial expressions from video is boundary detection, i.e. the boundary between the onset and offset of each expression. One approach to this is to try to recognise either the onset or the apex of a facial expression. For example, in a simple situation a subject might begin with a neutral expression and then progress to an expression of surprise. [Lucey 06] attempted to recognise expressions from peak to peak. Whatever the temporal position in the expression, this implies training a classifier of still images and then capturing one or more video frames from a video sequence before attempting to match the expression.

4.2.3 Video Processing

Videos come in a wide variety of formats and containers. Unfortunately, not all freely available computer vision software can operate on all formats. The very popular OpenCV [OpenCV] software, commonly used to perform face detection in videos is also capable of capturing frames from a video, in principle, obviating the need for an additional specialised image capture software. However, OpenCV will only process the Audio Video Interleave (AVI) container format, introduced by Microsoft in 1992, thus constraining the solution to some extent.

Regardless of the video format, capturing image frames from the video at certain intervals is essential and, in practice, may need to be performed both manually and automatically. The frames need to be captured for training and recognition phases.

4.2.4 Classification

It is essential that the system be capable of incorporating different types of classifiers, e.g. SVM, ANN, Boosting, for individual inputs and possibly an ensemble of classifiers for fusing the individual classifications and weighting them.

4.2.5 Miscellaneous

In order to compare different techniques it must be possible to re-run training and testing phases, i.e. persisting all the available inputs, interim and final results. It is also desirable to be able to compare studies or projects and store the results separately for later comparison. The system must be capable of running in offline or real-time mode. For example, models to be used in the classification phase might be pre-built, but the system should be able to perform real-time analysis. Easy analysis of results must be possible in order to determine how best to tune and improve the system.

4.2.6 System Operational Requirements

4.2.7 Implementation Platforms

Ideally the system should have broad platform support. In practical terms, this translates to variants of Unix and Linux, Mac OS and Windows.

4.2.8 Audio and Video Formats

One of the first hurdles that one encounters, especially with video processing, is the number of different video container formats and their lack of availability on one or more operating environments. Where possible the system should be capable of supporting multiple audio and video container formats. At a minimum these should include WAV, AVI, MP4, MPEG2, and MOV. If support is not available, then there should be some simple way of converting between formats where this is possible.

Image Processing

The system needs the capability to train a classifier on a corpus of still emotional expressions. The corpus could be images of jpeg, png or some other format. Al-

ternatively, there may be no image corpus, rather, a video collection that will require significant images to be captured into a suitable format, thus creating a de facto corpus. The images will then be subjected to some recognition process.

Video Processing

A video can be hours in duration or it could simply be, as in the Cohn-Kanade database [Kanade 00], collections of short, sample expressions. Both in training and testing, the system needs to be able to capture frames from a video segment. The frames will then be subjected to a treatment similar to the image processing mentioned previously, and the resulting parameters input to, say, a HMM.

Classification

Several subsystems require some form of classification component. It is essential that the system be able to perform some data reduction of any input vectors that are large in dimensionality, e.g. PCA or Linear Discriminant Analysis (LDA).

4.2.9 System Performance

Ideally, the system will make use of multi-threading and multi-processing capabilities of the operating system. Performance is critical, as is efficient memory usage. It is preferable that the system be able to execute in online mode.

4.2.10 User Interface

The core system must be simple to use so that the effort required to integrate components and to re-run exercises is minimised.

4.3 The Any Expression Recognition System (NXS)

At the heart of NXS is the concept of a "Project". A project is a meta-object, e.g. a FER experiment containing metadata about each of the objects that are used in an experiment. However, it is not limited to FER and can contain information about audio object, as well. Each project can contain references to collections of videos, images, images sequences, audio segments, multiple AAMs and SVMs, and even other projects themselves.

The major benefit of projects is that results can be saved, experiments can be re-run with different parameters and outputs can be written to comma-delimited files for use in third-party products, e.g. Excel. AAMs and SVMs built for one project can be referenced and used in another project.

NXS is capable extracting and cataloguing frames from video, as well as performing real-time expression analysis.

4.3.1 Software Selection for the Core System

While the Windows operating system dominates the commercial and home environments, systems such as Linux and Mac OS also remain popular. In order to meet the cross-platform operating environment requirement, C++ or Java was considered suitable for development of the core system. However, given the critical performance requirements and the fact that most high-performing video processing libraries are C or C++ based, C++ was selected.

There is sometimes the misconception that the C or C++ programming languages can be used to build a system that is portable between different platforms. Without a lot of effort, this is not usually the case. Building a system using Nokia's Qt integrated development environment [Qt 09] guarantees that the computer programs, written in C++, will be cross-platform compatible. This was selected for development of the core

system and user interface.

Qt has another very attractive feature. Its MetaObject Pattern supports a programming concept known as "reflection". The benefits of reflection are realised when it comes to saving and restoring exercises. The state and values of objects and their properties are "reflected" fairly simply into, in this case, Extensible Markup Language (XML). Making use of this metaobject, it can do "round-trip" xml - serialising the objects, persisting them, and then deserialising them. In practical terms, this is the means to saving and restoring projects.

4.3.2 Software Selection for Major Functions

The Vision *something* Libraries (VXL) are used for image processing [VXL]. The libraries are written in C++ and are very efficient. OpenCV is used for simple video display and to capture images from videos [OpenCV]. SVMLIB [Chang 01] will be used for facial expression recognition.

4.3.3 Class Structure

The system has been built using design patterns as described by [Pree 95]. Qt lends itself to building with design patterns [Ezust 06]. Figure 4.1 depicts the conceptual class structure. Use is made of the serializer and composite patterns to effect the round-trip processing, mentioned in Subsection 4.3.1.

A "project" is the top level concept, created by a software factory, and is simply a collection of segments (discussed in the next section). Facades are used to abstract the details of external classes such as those used to perform AAM processing. A form factory is used to create simple dialog boxes for user input, reducing the amount of effort that would have otherwise have been required if the dialogs had been hand-crafted.

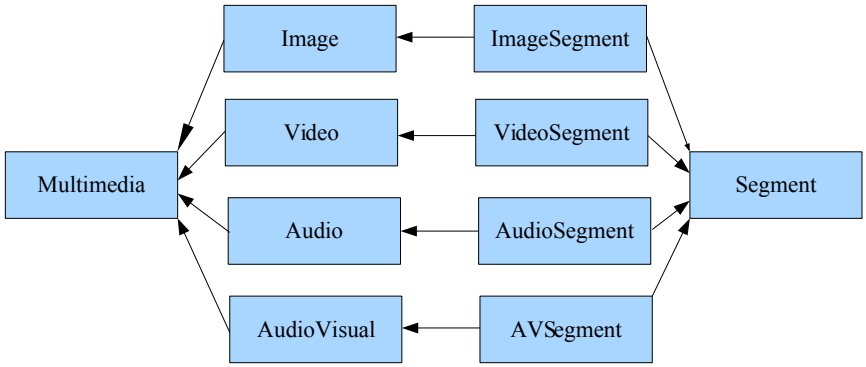


Figure 4.1: Class hierarchy

4.3.4 Segments

Central to the design of the system is the concept of segments. This borrows to some extent from, but is simpler than, MPEG-7 [Salembier 01] and its concept of segment types. This is no coincidence, MPEG-7, previously known as "Multimedia Content Description Interface", is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code. However, the implementation deviates slightly in that segment and multimedia data members and operations are combined.

The various types of segments are created by a segment factory. As can be seen in Figure 4.1, these include Image Segments, Image Sequence Segments, Image Collections, Audio Collections and Video Collections. Using factories to provide a layer of abstraction not only conceals the implementation complexity from the calling functions but simplifies the creation of new types of segments. Figure 4.2 depicts the segment factory class diagram.

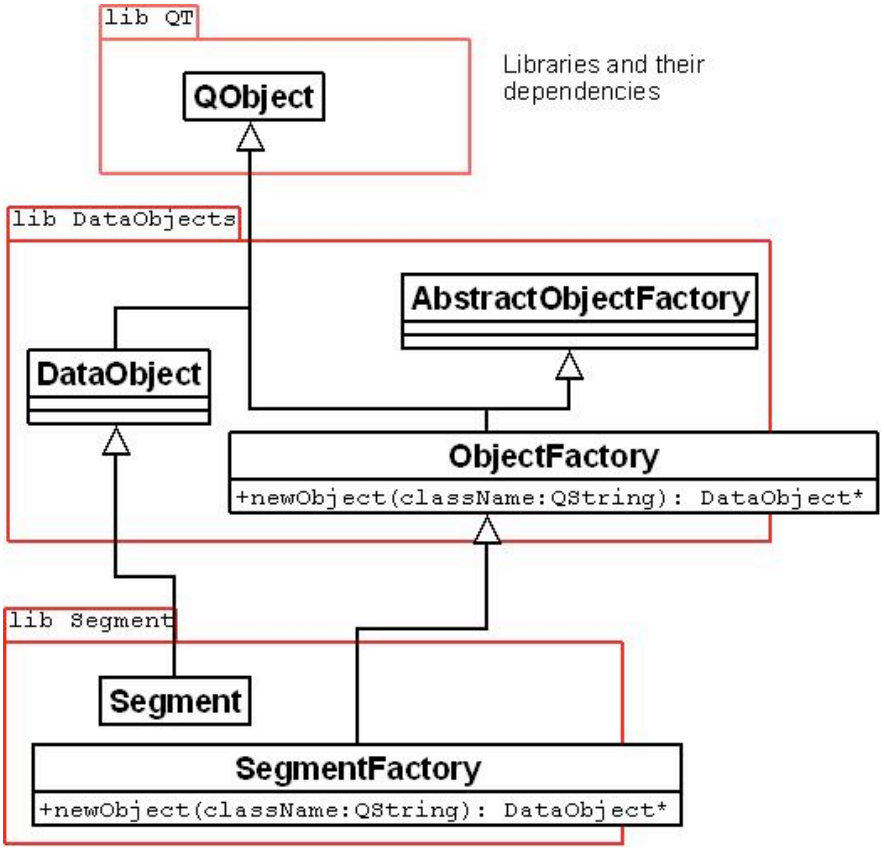


Figure 4.2: Class diagram of the Segment Factory

4.3.5 User Interface

4.3.6 Dialog Creation

User dialogs are, in keeping with the design pattern approach, created by factories. Figure 4.3 demonstrates the simplicity in creating new dialogs through the use of a form factory [Ezust 06].

```
bool ImageSegment::askQuestions() {
    FormFactory ff;
    *this << ff.newQuestion("name", "Image name", "ISEG000001");
    *this << ff.newQuestion("parent", "Parent Image Sequence Segment number", "");
    // new segment y or n
    m_fv = new FormDialog(this);
    //qDebug() << "Setting parent";
    m_fv->setWindowTitle("Image sequence details.");
    m_fv->exec();
```

Figure 4.3: Dialog creation

4.3.7 Processing Scenario

Rather than simply list each function, this is better described by a practical walk through a typical processing scenario. Most major functions are accessible by right-clicking to present a context menu as shown in Figure 4.4. The use of the product begins with the creation of a "Project" as seen in figure 4.5.

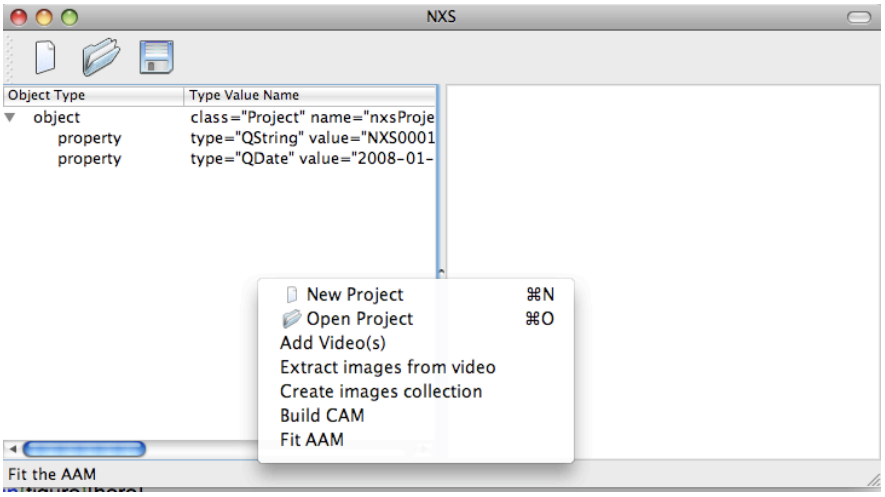


Figure 4.4: The system menu

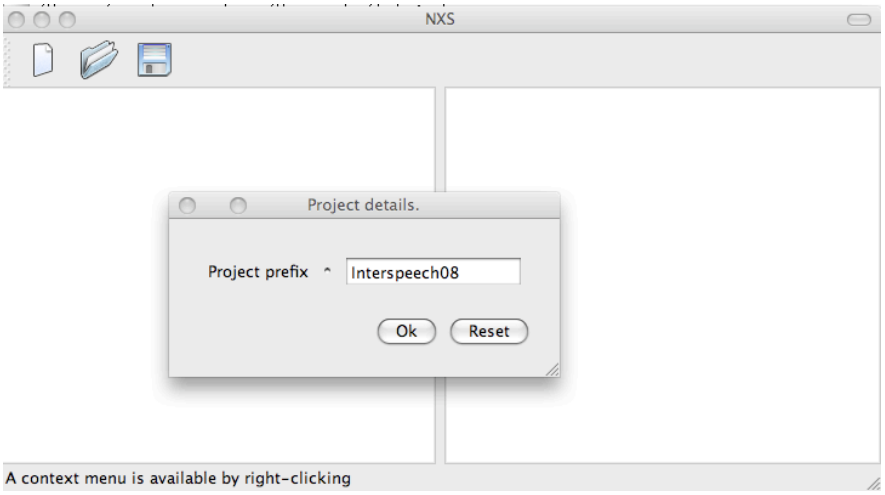


Figure 4.5: Project creation

Figure 4.6 shows the project tree structure after a project has been created, an image segment, image collection segment, video segment, and model segment have

been added to the project. The tree structure is effectively the xml that reflects the objects' states and data member values. From here the xml can be saved and reopened later.

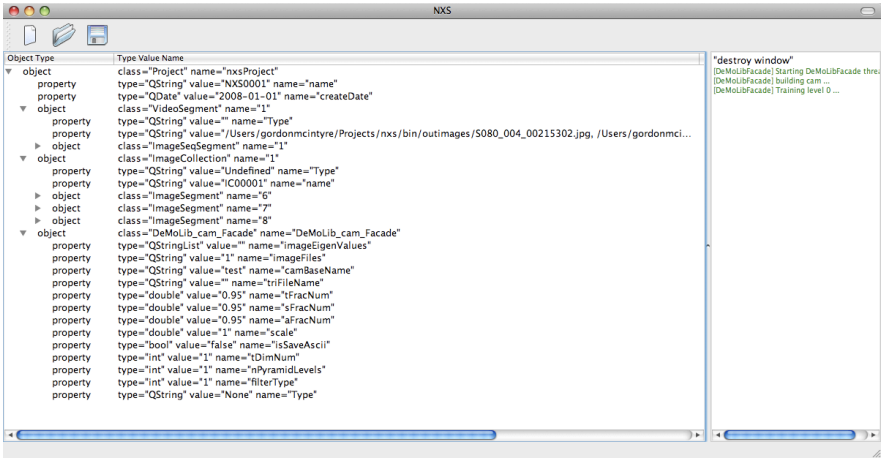


Figure 4.6: User interface

4.3.8 Measuring Facial Features

The facial image is subdivided into regions in order to track inter-region movement and a common scale is applied to the facial images. Similar to [S.Strupp 08], it is done by first creating a horizontal or transverse delineation line, mid-way between the topmost and bottommost landmark points on the facial image being examined, as shown in Figure 5.2. This simplified view is shown for illustration. This process is applied to any facial image used to train or test the system.

Within the experiments described in Chapters 5 and 6, reference to "shape" refers to normalised and scaled vectors. The process of "scaling" the data is explained in Section 5.3.2. The algorithm for the *normalisation* process is described in Algorithm 3.

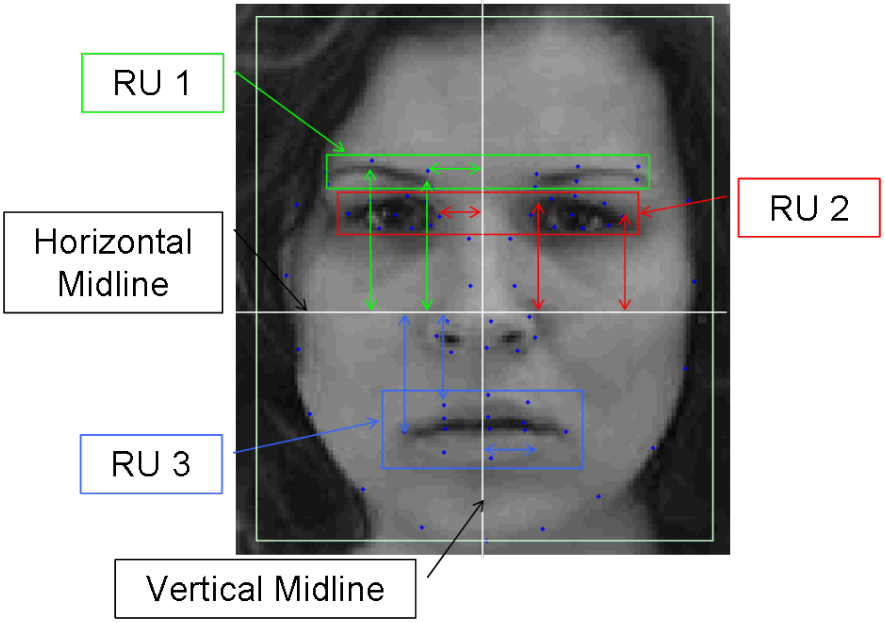


Figure 4.7: Measurements from horizontal delineations image from Feedtum database [Wallhoff]

4.3.9 Classification

Two sets of classifiers are used in this work, one for prototypical expressions, and the others specific to RUs. Those specific to RUs are trained to classify the major intra-RU patterns that would normally accompany prototypical expressions. For instance, the mouth and lip movements of fiduciary points in an image of the prototypical smile, are re-used and represent one class in the RU3 classifier. The relevant normalised measurements are used as inputs to all of the classifiers. For example, in the case of RU1, only the landmark points within RU1 are input to the classifier. The mappings are shown in Table 4.1.

Note that eye movement AUs, such as those that effect blinks, are not incorporated at present, e.g. AU43, AU45-46 and AU61-68.

Algorithm 3: Measuring and normalising the x, y distances

```
input Face x,y coordinates
xPoints (total number of x, y points/2)
yPoints (total number of x, y points/2)
xNormalised (total number of x, y points/2)
yNormalised (total number of x, y points/2)
i = 0
for total number of x, y points/2 do
    xPoints[i] = FaceXYCoordinates[2*i] ;    // group x coordinates
    yPoints[i] = FaceXYCoordinates[2*i+1] ; // group y coordinates
    i++
// get the offset from the left
// and top of the image frame
xMin = xPoints.min()
yMin = Points.min()
// get the mean
xmean = xPoints.mean()
ymean = yPoints.mean()
j = 0
for total number of x, y points/2 do
    xNormalised[j] = (xPoints[j] - xmean)
    yNormalised[j] = (yPoints[j] - ymean)
    j++
// normalize the vectors
xNormalised.normalize()
yNormalised.normalize()
```

4.3.10 System Processing

The processing sequence is depicted in Figure 4.8. First, images are captured at predefined intervals from video, 500ms in this case. Frontal face poses are then segmented from the images using the Viola and Jones [Viola 01] (or a derived) technique to determine the global pose parameters. See [Viola 01] for details. Next, AAMs, which have been prepared in advance, are fitted to new and previously unseen images to derive the local shape and texture features. The measurements obtained (see Section 4.3.8) are then used to classify the regions using SVM classifiers. Ultimately, the number and mixture of classified regions are used to report facial activity.

AU	Description	RU
1, 2, 3, 4	Brow movement	1
5, 6, 7	Eyelid activity and orbicularis oculi	2
9, 10, 12, 15, 17, 20, 25, 26, 27	Nose, mouth and lip regions	3

Table 4.1: Mapping AUs to RUs

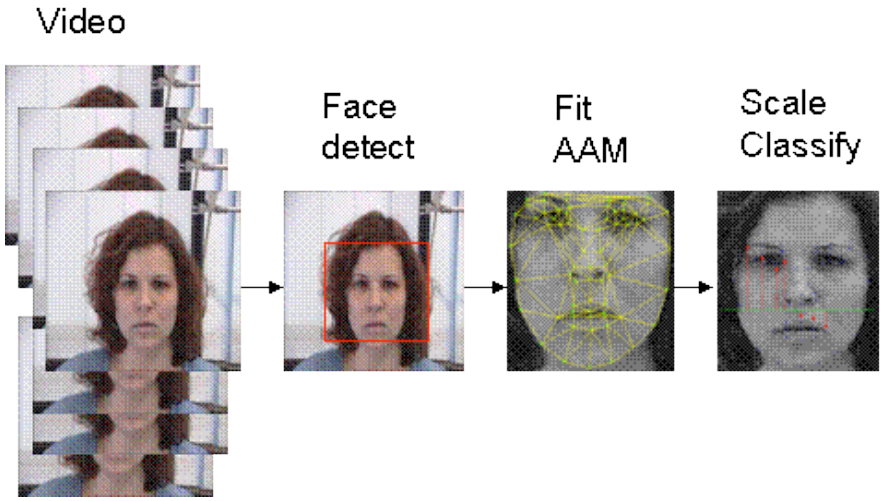


Figure 4.8: Processing of facial activity measurements

4.3.11 Active Appearance Models

To recap on chapter 3, in recent years, a powerful deformable model technique, known as the Active Appearance Model [Edwards 98b], has become very popular for real-time face and facial expression recognition. In the AAM approach, the non-rigid shape and visual texture (intensity and colour) of an object are statistically modelled using a low dimensional representation obtained by applying PCA to a set of labelled training data. After these models have been created, they can be parameterised to fit a new image of the object, which might vary in shape or texture or both.

Figure 4.9 depicts the facial triangulation mesh, based on the feature points, used to divide the facial image into regions. In this case, AAMs are pre-trained on static images using the SIC fitting method [Baker 03a]. SIC is an another adaptation of in-

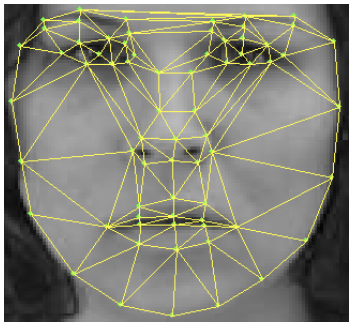


Figure 4.9: Face mesh used to group AUs

verse compositional image alignment for AAM fitting, that addresses the problem of the significant shape and texture variability, by finding the optimal shape and texture parameters simultaneously. Rather than re-computing the linear update model at every iteration using the current estimate of appearance parameters, it can be approximated by evaluating it at the mean appearance parameters, allowing the update model to be pre-computed, which has significantly more computational efficiency. See [Baker 03a] for details. The system described in section 7 involves the recording of frontal face images, while the participant views stimuli on a computer display. Under this setup, there is a reasonable tolerance to head movement and short-duration occlusion.

4.3.12 Classification Using SVM

One implementation of SVM, often incorporated within large-scale and comprehensive data mining packages, is LIBSVM [Chang 01]. It is well supported and can be implemented independently of any host product. SVM generalization performance depends on the appropriate setting of meta-parameters parameters C and the kernel gamma parameter γ and, to achieve this, LIBSVM includes a *grid search* program to establish optimum parameter settings.

4.3.13 Gabor Filter Processing

The Gabor filter processing made use of a software implementation [cvG], which had been used by [Zhou 06].

I am an old man and have known a great many troubles, but most of them never happened.

Mark Twain

5

Sensing for Anxiety

5.1 Introduction and Motivation for Experiments

5.1.1 Introduction

Anxiety is a normal reaction to everyday events and everyone experiences it at some point in time. Elevated arousal levels accompany many common activities such as examinations, public speaking and visits to the dentist. It is when anxiety starts to affect a person's everyday life, that it becomes classed as a disorder. Anxiety disorders are the most common mental disorders in Australia. Nearly 10% of the Australian population

will experience some type of anxiety disorder in any one year - around one in twelve women and one in eight men. One in four people will experience an anxiety disorder at some stage of their lives [beyondblue]. In the United States, anxiety disorders affect about forty million adults a year - around 18% of the population [Kessler 05]. Anxiety is often comorbid with depression and between them they can have serious health implications.

The term, "anxious expression", is used in everyday language and most people would give tacit acknowledgement to it's intended meaning. Thus, given the prevalence of anxiety and actual disorders, one would think that such an expression would be straightforward to define and, similarly, to automatically recognise. However, the void that occupies the literature on automatic anxious expression recognition suggests that this is not the case. This Chapter describes an attempt, through a set of novel experiments, to test the feasibility of such an exercise.

Chapter 3 introduced several state-of-the-art techniques used in the computer analysis of facial expressions. Chapter 4 described a computer system, *NXS*, that has been developed in order to support the experimental aspects of this dissertation. Although the system provides flexibility in processing video and images, incorporating implementations of AAMs, Gabor Wavelet Transforms and SVM. Its successful application is conditional on proper calibration of these functions. Indeed, even with depth of knowledge and a thorough review of the literature of similar research, some of the optimal settings are obtained through experience, and empirically, i.e. through trial and error which can appear like a *black art* at times.

The motivation for the experiments presented in this chapter is to:

- prove the concepts described in previous chapters can be applied in a practical experiment;

- understand how best to use and calibrate the NXS system; and

- taking anxiety as an example, try to establish if automatic expression recognition can be used to differentiate a subtle, non-primary emotion from other facial expressions.

This chapter is organised as follows. Section 5.2 explains the hypotheses and questions to be addressed. This is followed in Section 6.3 with a description of the methodology used in the experiments. Section 6.4 presents the data and analysis from the experiments. Finally, Section 5.5 concludes and evaluates the exercise.

5.2 Questions and Hypotheses

5.2.1 Hypotheses

On the basis of the motivation for the experiments and the literature review in Chapter 3, the following hypotheses and questions were generated: TODO fix indentation

1. Using computerised facial expression recognition techniques, anxious expressions can be differentiated[1] from fearful expressions.

2. Using computerised facial expression recognition techniques, anxious expressions can be differentiated from a larger set of prototypical expressions.

5.2.2 Questions Pertaining to the Importance of Feature Data

Figure 5.1 depicts the fiduciary or landmark points that are "fitted" to each image during analysis. The collective landmark points, referred to in this dissertation as "shape", are captured as a set of x, y Cartesian coordinates. As explained in Chapter 3, *texture*, or the spatial variation in the gray values of pixel intensities in the image, are also obtained.

[1] "Differentiated" is defined as that better than chance.

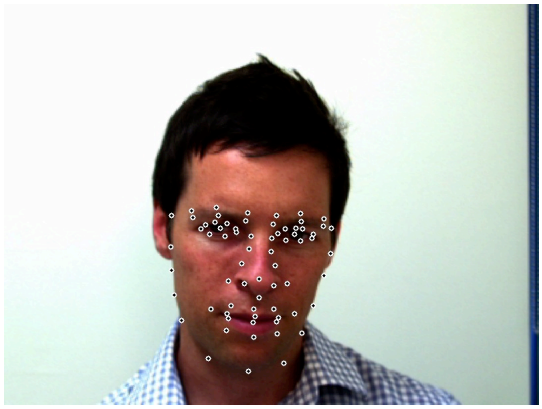


Figure 5.1: Facial landmark points in image.

Several options are available to make use of this feature data to classify expressions, e.g. using shape information only, using texture information only, or by combining the shape and texture information in some way. In keeping with the motivation of the this Section, the following set of questions were posed:
How does facial expression recognition performance, i.e. *Classification Accuracy (CA)*, vary when using:

1. the location of facial landmark points only;
2. the location of facial landmark points concatenated with Gabor magnitude;
3. the location of facial landmark points concatenated with AAM texture parameters;
4. AAM texture parameters only; and
5. Gabor magnitude only?

5.2.3 Questions Pertaining to the Relative Importance of Facial Regions

In Section 4.3.8 of Chapter 4 the subdivision of facial areas for analysis is explained. Figure 5.2 is reproduced for convenience, showing the three facial regions to be used:

- R1 - The eyebrow region;
- R2 - The eye region; and
- R3 - the mouth region

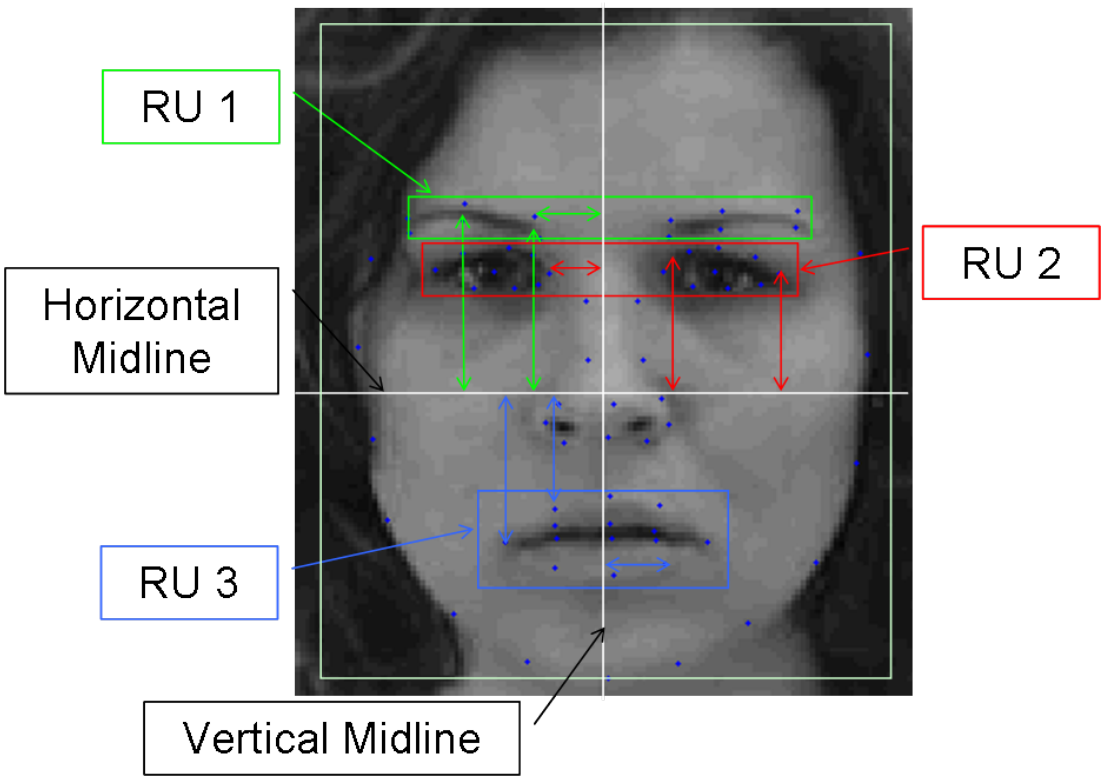


Figure 5.2: Facial region demarcation in image [Wallhoff]

It would be useful to know if one facial region is more important than another for *all* expressions or only for *specific* expressions. More formally, the following questions were of interest:

1. Is one facial region *generally* more reliable for recognition of expressions?
2. Is one facial region more reliable for recognition of a *specific* expression?

5.2.4 Question Pertaining to System Performance

Of interest is the cumulative execution time of face-detection, AAM fitting and classification of an image since these steps would apply to an online recognition system (although the face-detection step is not normally undertaken in every frame).

The following question was therefore posed:

1. Would performance be sufficient to achieve on-line recognition of facial expressions, in a video running at 30 frames per second (online recognition has the additional step of capturing the frame from video)?

5.3 Methodology

5.3.1 Experimental Setup

Experiment 1 The first experiment, as a baseline, was to determine if the NXS system could be trained to differentiate between prototypical facial expressions labelled as 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and ' Neutral' from the Cohn-Kanade database. The number of occurrences of each expression is shown at Table 5.1.

Anger	Fear	Sad	Surprise	Neutral	Happy	**Total**
33	16	17	21	36	31	**154**

Table 5.1: Experiment 1 - Number of occurrences of each expression [Kanade 00]

Experiment 2 The second experiment was to test Hypothesis 1 in Section 5.2 and determine if the NXS system could differentiate between facial expressions *nominated by human judges* as anxiety against those *nominated by human judges* as fear.

There are no freely available databases of anxious facial expressions. To improvise, a set of expressions from the Cohn-Kanade [Kanade 00] database having annotations corresponding to action units of fear and anxiety were selected. They were first analysed by the author who made a preliminary judgement, labelling the expression as either 'Anxiety' or 'Fear'. The number of occurrences of each class label from the preliminary assessment is shown at Table 5.2.

An anonymous poll was then conducted and participants asked to view each expression and judge whether they thought the expression should be labelled as 'Fear', 'Anxiety or 'Uncertain'. Those invited to participate in the poll were not given any indication of the preliminary label.

Where 50% or more of judges labelled an expression as either 'Fear' or 'Anxiety', i.e. one or the other, the image was retained for use in the exercise with the voted label attached. Note that there is no suggestion that any of the expressions *are*, in fact, anxiety. For the purpose of the experiment it is not essential that they be anxiety - merely that the computer system be trained to classify them as such, and agree with the judges' opinions on the expressions used in training.

Fear	Anxious	**Total**
29	26	**55**

Table 5.2: Initial numbers of each expression [Kanade 00]

The raw data results of the poll are shown at Appendix A. A summary of the revised numbers of each class is at Table 5.3.

Fear	Anxious	**Total**
18	16	**34**

Table 5.3: Results from poll - numbers labelled as fear and anxiety retained.

For reasons discussed later, 3 images were removed from the exercise, leaving 31 images for use in the experiment. The number of occurrences of fear and

anxiety are shown in Table 5.4

Fear	Anxious	**Total**
16	15	**31**

Table 5.4: Experiment 2 - Final number of occurrences of each expression retained.

Of these occurrences, 9 are male and 22 are female. An even split would have been ideal, but a little difficult to attain with such a small sample set when sixty-five percent of subjects in the Cohn-Kanade database are female. Individual attributes of the recorded actors are not known. We are told, however, that in the database subjects range in age from 18 to 30 years, 15 percent are African-American, and three percent are Asian or Latino.

Experiment 3 The third experiment was to test Hypothesis 2 in Section 5.2 and establish if the NXS system could differentiate facial expressions of anxiety from a larger set of emotional expressions which included, 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and ' Neutral'. The images of fearful and anxious expressions used in Experiment 2 were combined with those of Experiment 1, replacing Experiment 1's fearful expressions with Experiment 2's fearful expressions, and adding Experiment 2's anxious expressions. All images were from the Cohn-Kanade database and the numbers of each expression is shown at Table 5.5.

Anger	Fear	Sad	Surprise	Neutral	Happy	Anxious	**Total**
33	16	17	21	36	31	15	**169**

Table 5.5: Experiment 3 - Numbers of each expression from Cohn-Kanade database

Experiment 4 The objective of the fourth experiment was to test whether the classifier built from the Cohn-Kanade database of images, in Experiment 1, could be used to predict the facial expressions in the Feedtum database of images [Wallhoff], which were recorded with different subjects and under different lighting conditions. Prototypical facial expressions of 'Fear', 'Anger', 'Happy', 'Sad', 'Sur-

prise', and ' Neutral' from both the Cohn-Kanade database and the Feedtum database were selected.

As a preliminary step, a baseline classification performance test was conducted using prototypical facial expressions of 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and ' Neutral' from only the Feedtum database, so that classification performance could be compared with that found in Experiment 1, where images from the Cohn-Kanade database were used. This is referred to here as the "baseline" experiment.

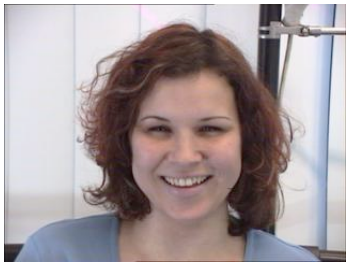
Next, an attempt was made to automatically classify expressions in images from the Feedtum database against the SVM models built in Experiment 1. The total numbers of each expression is shown at Figure 5.6.

Anger	Fear	Sad	Surprise	Neutral	Happy	**Total**
9	11	10	11	15	14	**70**

Table 5.6: Experiment 4 - Numbers of each expression from Feedtum database

A sample of an image from each database is presented at Figure 5.3 which show clearly the marked difference in the lighting conditions between the image from the Cohn-Kanade database on the left and the Feedtum database on the right.

(a) Sample image from the Cohn-Kanade database



(b) Sample image from Feedtum database

Figure 5.3: Experiment 4 - Images from different databases showing different lighting conditions

5.3.2 System Setup

The broad concept of facial expression recognition has been explained in Chapter 3 and the implementation of various associated functions in the NXS system described in Chapter 4. Although the NXS system can be used to both train and test classifiers, in this set of experiments, it was used, predominantly, to build AAMs and fit them to the set of images in the experiments. When fitting AAMs to images, NXS records:

- the Cartesian coordinates of each landmark point that makes up the face "shape". These are normalised to take into account differences in face sizes;
- the AAM texture parameters; and
- the Gabor magnitudes.

The meaning of the term, "shape", as used in this dissertation, is used broadly to mean the facial landmark points and the subdivision of facial regions into eyebrow (R1), eye (R2), and mouth (R3). This is explained in Chapter 4.

One of NXS' other functions, using the aforementioned stored face and texture features, is to scale and normalise them and output them in a format suitable for input to an external classification product such as LIBSVM [Chang 01] or RapidMiner [RapidMiner] (which uses LIBSVM).[2]

Eight output datasets were produced for each experiment. These contained feature sets of shape, shape and Gabor texture concatenated, shape and AAM texture parameters concatenated, Gabor magnitude, AAM texture, eyebrow shape (R1), eye shape (R2) and mouth shape (R3). The tuning details of each major functional area and parameter selections used in this set of experiments are explained below.

[2]LIBSVM uses an format called SVM Lite, whereas RapidMiner imports, amongst other things, tab-delimited format.

AAM Choice and Parameter Selection

The Iterative IEBM method [Saragih 06] of building the AAM and fitting the model to the image was chosen because of its fast fitting capabilities[3] which would be critical in achieving real-time fitting as required in later experiments. Depending on the parameter selection, model training time was longer than that experienced with the SIC method. The implementation of the algorithm by [Saragih 06] was used,[4] and after the recommended settings were applied, the parameters were fine-tuned by trial-and-error in consultation with the software provider.

A before and after example of "fitting" an AAM to a face in an image from the Cohn-Kanade database is given at Figure 5.4.

(a) Original image (b) Image after fitting

Figure 5.4: Original image on left and image after fitting on right

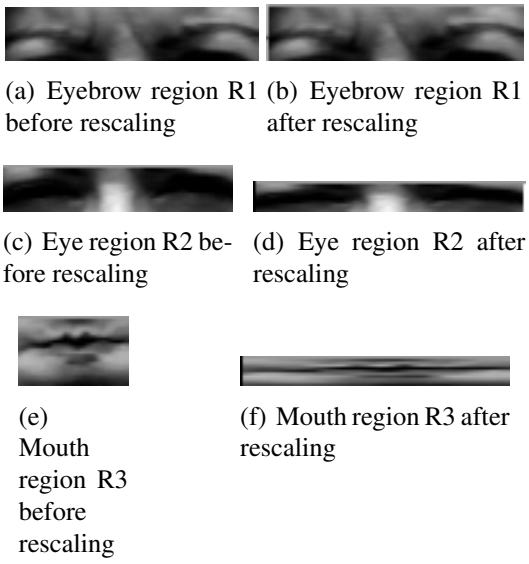
Gabor Processing and Parameter Selection

Due to the high-dimensionality of output features from Gabor filters, processing was only applied to the R1, R2 and R3 regions - not to the entire cropped face. Prior to convolving the image with the Gabor filter, the region of interest, e.g. R1, is scaled or warped using bicubic interpolation to a fixed, canonical size - 100×10 pixels for R1

[3] This had been shown in earlier informal trials [Saragih 09].

[4] The implementation is written in the C++ programming language.

and R2 and 100 x 20 pixels for R3 and written to a grayscale image. To facilitate the explanation, the before and after images are shown at Figure 5.5.



(a) Eyebrow region R1 before rescaling

(b) Eyebrow region R1 after rescaling

(c) Eye region R2 before rescaling

(d) Eye region R2 after rescaling

(e) Mouth region R3 before rescaling

(f) Mouth region R3 after rescaling

Figure 5.5: Regions before and after rescaling (3 $\times$ actual size)

The Gabor filter processing made use of a software implementation [cvG], which had been used by [Zhou 06].[5] The program implements the Gabor wavelet using the formula in [Zhou 06] and shown at Equation 5.3.1.

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}} [e^{ik_{\mu,\nu^z}} - e^{-\frac{\sigma^2}{2}}] \tag{5.3.1}$$

where $z = (x, y)$ is the point with the horizontal coordinate x and the vertical coordinate y. The parameters μ and ν define the orientation and scale of the Gabor kernel, $\|\cdot\|$ denotes the norm operator, and σ is related to the standard derivation of the Gaussian window in the kernel and determines the ratio of the Gaussian window width to the wavelength. The wave vector $k_{\mu,\nu}$ is defined as follows

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \tag{5.3.2}$$

[5]The implementation is also written in the C++ programming language.

where $k_\nu = \frac{k_{max}}{f^\nu}$ and $\phi_\mu = \frac{\pi\mu}{8}$ if 8 different orientations have been chosen. k_{max} is the maximum frequency, and f^ν is the spatial frequency between kernels in the frequency domain.

The second term in the square brackets in Equation 5.3.1 compensates for the DC value and its effect becomes negligible when the parameter σ, which determines the ratio of the Gaussian window width to wavelength, is sufficiently large.

The approach taken was analogous to the "eigenface" recognition process. In this method, each training image is "flattened" to a $1 \times D$ vector and the vector pushed onto a stack of vectors. Once all training images have been processed, dimensionality reduction takes place. The eigenvectors of the most significant eigenvalues are used to project the vectors into eigenspace and the coefficients used to train a classifier. The recognition phase processes the images in a similar manner with coefficients matched against those derived in the training phase.

In this experiment, each $N \times M$ image of the extracted facial region (R1, R2 and R3) was convolved with each Gabor filter and the resulting Gabor magnitudes are laid out end-to-end in a vector of dimension $N \times M$. Thus, after convolving an image region with 40 filters, the result is an $N \times M \times 40$ vector. After processing each image, the $N \times M \times 40$ vector is pushed into a stack of vectors of convolved images. Once all of the images are stacked, OpenCV's *cvCalcPCA* function is used to perform PCA to get the eigenvalues and eigenvectors. A truncated set of eigenvectors is then used and OpenCV's *cvProjectPCA* function is called to project the vectors into "eigenspace", finally using the resulting coefficients for recognition. Through trial and error, it was found that 20 eigenvectors explained approximately 90% of the variation in the set of Gabor magnitude responses.

By far the most challenging aspect of the experiment was to calibrate the Gabor filters. Selection of the scale parameters σ and ν, affects the width of the kernel and involves a tradeoff. Larger values are more robust to noise but less sensitive. Smaller val-

ues are more sensitive but less effective in removing noise. Finding optimum settings involved reference to a number of articles [Bhuiyan 07, Chen 07, Fasel 02, Gao 09, Kamarainen 06, Kanade 00, Lades 93, Lee 96, Liu 04, Liu 06, Movellan 08, Shen 06, Shen 07, Wiskott 97, Wu 04] and a lot of trial and error.

Many articles do not explain the reasons behind parameter selection, instead referring, if at all, to other articles that, in turn, do not explain the settings or simply refer to yet another article. The popular setting for K_{max} is $\pi/2$ and for the spatial frequency f is $\sqrt{2}$. It would seem that the setting for $K_{max} = \pi/2$ originates from [Lades 93] who noted it yielded the best results after trialing values of $3\pi/4, \pi/2, \pi/3$. [Lades 93] also seems to be responsible for the setting of the spatial frequency f value to $\sqrt{2}$ after trialing values $f = 2, f = \sqrt{2}$.

In a study into automatic coding of facial expressions displayed during posed and genuine pain, [Littlewort 07] disclose that they convolved their 96×96 images through a bank of Gabor filters eight orientations and 9 spatial frequencies (232 pixels per cycle at $1/2$ octave steps). They then pass the output magnitudes to action unit classifiers. Yet, in a study from the same laboratory, [Whitehill 09] describe an attempt to provide an optimum set of parameters in the task of performing smile detection against a real-world set of photographs. The database, GENKI, consists of pictures from thousands of different subjects, photographed by the subjects themselves. In the experiment, all images were converted to grayscale, normalised by rotating and cropping around the eye region to a canonical width of 24 pixels. The authors report that they used energy filters to model the primate visual cortex, using 8 equally spaced orientations of 22.5 degrees, but they do not explain how they arrived at the spatial frequencies with wavelengths of $1.17, 1.65, 2.33, 3.20$ and 4.67 Standard Iris Diameters [6]. The authors refer to [Donato 99] for filter design, however, [Donato 99] report spatial frequency

[6] A Standard Iris Diameter is defined as $1/7$ of the distance between the centre of the left and right eyes.

values of $\nu \in \{0, 1, 2, 3, 4\}$ and go on to describe a further test using high frequency values of $\nu \in \{0, 1, 2\}$ (scale is the inverse of frequency) and low frequency values of $\nu \in \{2, 3, 4\}$. They state that the performance of the high frequency subset $\nu \in \{0, 1, 2\}$ was almost the same as $\nu \in \{0, 1, 2, 3, 4\}$. It should be noted that the task at hand was the classification of FACS Action Units. Intuitively, one would expect high scale (low frequency) to generalise or provide a better representation of expressions than low scale (high frequency), since the former is more likely to ignore artifacts.

The MPEG-7 [MPEG-7] *Homogeneous Texture Descriptor* standard has made use of Gabor filter banks of 6 orientations and 5 scales. The number of orientations and scales were based on previous results from [Manjunath 96, Ro 01]. It's use is pitched towards automatic searching and browsing of images and the mean and standard deviation of each filtered image, plus the mean and std of the input image are typically used as features ($30 \times 2 + 2 = 62$ features).

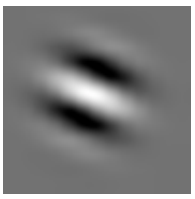
After much experimentation, values of $\sigma = \pi$ and $nu \in \{0.0, 0.06, 1.4\}$ were used as in [Bhuiyan 07]. In this experiment, the kernel or mask width is automatically decided by the spatial extent of the Gaussian envelope and was obtained from the formulation implemented in [Zhou 06]

$$6 \times \sigma / \frac{k_{max}}{f^{\nu}}) + 1 \tag{5.3.3}$$

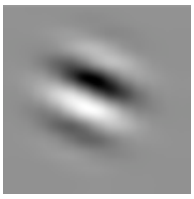
Truncating the Gabor filters to a width of $6\sigma + 1$ points (pixels) as in [Dunn 95] and using $\sigma = \pi$, $F = \sqrt{2}$, $K_{max} = \pi/2$ and $\nu = 0.06$, gives

$$6 \times (3.14159265 / \frac{3.14159265/2}{1.41421356^{0.06}}) + 1 \tag{5.3.4}$$

gives a filter size of 13×13 pixels. A sample filter with scale = 1.4, orientation= $\pi/8$ can be seen at Figure 5.6. Images of the regions R1, R2 and R3 filter response magnitude are shown at figures 5.7.



(a) Real part of wavelet

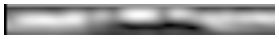


(b) Imaginary part of wavelet

Figure 5.6: Real and Imaginary part of a Gabor wavelet, scale = 1.4, orientation= $\pi/8$ ($5 \times$ actual size)

(a) Eyebrow region R1 magnitude response

(b) Eye region R2 magnitude response



(c) Mouth region R3 magnitude response

Figure 5.7: Magnitude responses of regions R1, R2 and R3 after convolution ($3 \times$ actual size)

Classification

In all of the experiments described in this chapter, a Radial Basis Function (RBF) SVM kernel type with a regularized support vector classification (standard algorithm C-SVC) type of SVM was used. As discussed previously, datasets of features were output from the NXS system as training and testing sets. Optimal parameters for the SVM models were found using the "grid.py" program. The training set data was reused for testing in a 5-fold cross validation setup, as depicted in Figure 5.8. In this Figure, "experiment " is used to describe the training run. Once found, the optimal parameters were transcribed to the RapidMiner product [RapidMiner] in order to produce a confusion matrix of CA of the type shown in Figure 5.9. There were slight differences between LIBSVM results and RapidMiner which was very likely to due to differences in the cross-fold validation sampling algorithms between the products. RapidMiner has several algorithms of its own available and the stratified sampling type was used to

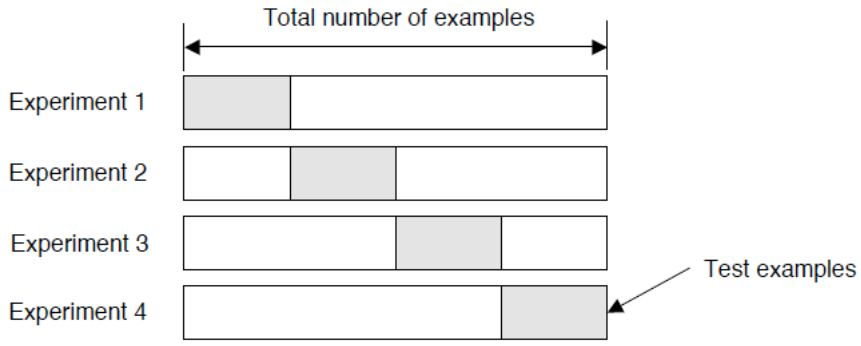


Figure 5.8: Leave-one-out cross validation [PRISM]

create random subsets while keeping class distributions constant.

5.4 Presentation and Analysis of Data

The sample size in all of the experiments was small. Therefore, even the misclassification of one or two expressions has a relatively high impact on the CA. Thus, some caution is needed in interpreting results with regards to CA, where differences of only a few percentage exist, and throughout this summary they are ignored. In the table column headings, "true" denotes the actual or real classification and "pred.", abbreviated from predicted, denotes the derived classification.

5.4.1 Experiment 1

The first experiment was to determine if the NXS system could be trained to differentiate between prototypical facial expressions labelled as 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and ' Neutral' from the Cohn-Kanade database. The recognition results using shape, shape and Gabor magnitudes concatenated, shape and AAM texture parameters concatenated, Gabor magnitudes and AAM texture parameters are given in Figures 5.9, and using R1, R2 and R3 shape in Figure 5.10.

Classification using the shape concatenated with the Gabor magnitudes yielded the best overall CA. Given their holistic nature, one would have thought that the shape concatenated with the AAM texture parameters would have performed better than the rest. Perhaps there was an advantage of having all of the image patches set to a fixed canonical size, as was the case with the Gabor filter pre-processing. There were no exceptionally performing feature sets that provided a CA much higher than the others.

Of the CA results obtained using R1, R2 and R3 (Figure 5.10), the eyebrow (R1) shape features achieved a 90.91% accuracy in the prediction of Anger. In nearly all of the individual results, where surprise was misclassified, it was most often misclassified as fear. Interestingly, the converse was not true.

Anger was most often the expression with the highest CA, regardless of the feature

sets that were used to build the classifer.

Accuracy: 76.62%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	27	3	1	5	0	0	75.00%
pred. Fear	2	10	0	0	3	0	66.67%
pred. Happy	0	3	25	1	1	0	83.33%
pred. Neutral	3	0	4	29	1	6	67.44%
pred. Surprise	0	0	0	0	16	0	100.00%
pred. Sad	1	0	1	1	0	11	78.57%
class recall	81.82%	62.50%	80.65%	80.56%	76.19%	64.71%	

(a) Confusion matrix of recognition using shape only

Accuracy: 85.74%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	28	0	1	2	0	1	87.50%
pred. Fear	2	12	0	0	2	0	75.00%
pred. Happy	0	3	27	1	0	0	87.10%
pred. Neutral	3	1	2	33	1	1	80.49%
pred. Surprise	0	0	0	0	17	0	100.00%
pred. Sad	0	0	1	0	1	15	88.24%
class recall	84.85%	75.00%	87.10%	91.67%	80.95%	88.24%	

(b) Confusion matrix of recognition using shape and Gabor magnitudes

Accuracy: 83.14%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	27	1	1	2	0	0	87.10%
pred. Fear	2	13	1	0	2	0	72.22%
pred. Happy	0	2	25	1	2	0	83.33%
pred. Neutral	4	0	4	32	0	3	74.42%
pred. Surprise	0	0	0	0	17	0	100.00%
pred. Sad	0	0	0	1	0	14	93.33%
class recall	81.82%	81.25%	80.65%	88.89%	80.95%	82.35%	

(c) Confusion matrix of recognition using shape and AAM texture parameters

Accuracy: 82.45%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	28	1	1	2	0	0	87.50%
pred. Fear	2	11	0	0	1	0	78.57%
pred. Happy	0	0	25	3	1	0	86.21%
pred. Neutral	3	2	5	30	2	1	69.77%
pred. Surprise	0	2	0	0	17	0	89.47%
pred. Sad	0	0	0	1	0	16	94.12%
class recall	84.85%	68.75%	80.65%	83.33%	80.95%	94.12%	

(d) Confusion matrix of recognition using Gabor magnitudes in eyebrow, eye and mouth regions

Accuracy: 78.62%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	27	3	1	3	0	3	72.97%
pred. Fear	2	10	0	0	2	0	71.43%
pred. Happy	0	1	26	1	2	0	86.67%
pred. Neutral	4	2	4	31	1	2	70.45%
pred. Surprise	0	0	0	0	15	0	100.00%
pred. Sad	0	0	0	1	1	12	85.71%
class recall	81.82%	62.50%	83.87%	86.11%	71.43%	70.59%	

(e) Confusion matrix of recognition using AAM texture parameters from entire face

Figure 5.9: Experiment 1 - Recognition results

Accuracy: 76.09%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	30	1	1	2	0	1	85.71%
pred. Fear	0	8	0	0	2	1	72.73%
pred. Happy	0	2	23	3	0	1	79.31%
pred. Neutral	3	2	7	28	2	2	63.64%
pred. Surprise	0	2	0	2	16	0	80.00%
pred. Sad	0	1	0	1	1	12	80.00%
class recall	90.91%	50.00%	74.19%	77.78%	76.19%	70.59%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape

Accuracy: 79.27%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	29	0	2	3	1	0	82.86%
pred. Fear	0	11	0	0	2	2	73.33%
pred. Happy	3	2	26	2	0	0	78.79%
pred. Neutral	1	1	3	28	2	3	73.68%
pred. Surprise	0	1	0	1	16	0	88.89%
pred. Sad	0	1	0	2	0	12	80.00%
class recall	87.88%	68.75%	83.87%	77.78%	76.19%	70.59%	

(b) Confusion matrix of recognition using eye region (R2) shape

Accuracy: 76.00%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	class precision
pred. Anger	24	2	1	5	1	2	68.57%
pred. Fear	1	10	0	0	3	0	71.43%
pred. Happy	3	1	25	1	1	0	80.65%
pred. Neutral	4	1	5	29	0	2	70.73%
pred. Surprise	1	2	0	0	16	0	84.21%
pred. Sad	0	0	0	1	0	13	92.86%
class recall	72.73%	62.50%	80.65%	80.56%	76.19%	76.47%	

(c) Confusion matrix of recognition using mouth region (R3) shape

Figure 5.10: Experiment 1 - Recognition results using shape from eyebrow (R1), eye (R2) and mouth (R3) regions

5.4.2 Experiment 2

The second experiment was to determine if the system could differentiate between facial expressions *nominated* as anxiety against those *nominated* as fear. The fitting process did not work well with 3 images (this was likely due to the relatively small number of images being used for AAM training) and, for expediency, the images were removed from the training set. The revised number of images used in the exercise is shown in Figure 5.7. In total, there were 31 images selected for the experiment - 9 male and 22 female.

Fear	Anxious	**Total**
16	15	**31**

Table 5.7: Experiment 2 - Numbers of each expression from Cohn-Kanade database

The recognition results using shape, shape and Gabor magnitudes concatenated, shape and AAM texture parameters concatenated, Gabor magnitudes and AAM texture parameters are given in Figures 5.11, and using R1, R2 and R3 shape in Figure 5.12.

Overall, the CA was low. Shape alone, and individual R1, R2 and R3 shapes, yielded the best CAs. Classifiers that were built using texture features did not perform nearly as well as those built with shape features. One surprising result was the recognition performance using the shape extracted from the eyebrow region presented at 5.12(a). One could theorise that this was because poll participants made more use of the eyebrow region than the eye and mouth in their assessment of the expression. And, based on the prior discussion of anxious features at Subsection 2.5.2, it would seem that the less exaggerated brow movements with an anxious expression *would* be a discriminating factor between fear and anxiety expressions. This, of course, is quite speculative, with such a small sample size and the fact that there were only 14 poll participants.

Another explanation considered was that it was simply due to the efficacy of the

Accuracy: 70.48%

	true Fear	true Anxious	class precision
pred. Fear	10	3	76.92%
pred. Anxious	6	12	66.67%
class recall	62.50%	80.00%	

(a) Confusion matrix of recognition using shape only

Accuracy: 67.62%

	true Fear	true Anxious	class precision
pred. Fear	10	4	71.43%
pred. Anxious	6	11	64.71%
class recall	62.50%	73.33%	

(b) Confusion matrix of recognition using shape and Gabor magnitudes

Accuracy: 57.62%

	true Fear	true Anxious	class precision
pred. Fear	8	5	61.54%
pred. Anxious	8	10	55.56%
class recall	50.00%	66.67%	

(c) Confusion matrix of recognition using shape and AAM texture parameters

Accuracy: 58.10%

	true Fear	true Anxious	class precision
pred. Fear	14	11	56.00%
pred. Anxious	2	4	66.67%
class recall	87.50%	26.67%	

(d) Confusion matrix of recognition using Gabor magnitudes in eyebrow, eye and mouth regions

Accuracy: 54.76%

	true Fear	true Anxious	class precision
pred. Fear	12	10	54.55%
pred. Anxious	4	5	55.56%
class recall	75.00%	33.33%	

(e) Confusion matrix of recognition using AAM texture parameters from entire face

Figure 5.11: Experiment 2 - Recognition results

SVM processing. To examine the phenomenon further, two post-hoc experiments were devised reusing the eyebrow region (R1) data. The first was to randomly change the class labels in the samples that had been labelled as anxious and fear and to re-run the experiment. The results are shown in Figure 5.13(a). This resulted in a much lower CA - 67.62%.

The second post-hoc experiment was to test how well regression analysis would separate the classes. Epsilon Support Vector Regression (SVR) was used and the optimal parameters determined using an alternative grid search program.[7] The result is shown in Figure 5.13(b). This time the CA was lower - 64.76%.

[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#grid_parameter_search_for_regression`, last accessed 1 March 2010

Accuracy: 90.48%

	true Fear	true Anxious	class precision
pred. Fear	15	2	88.24%
pred. Anxious	1	13	92.86%
class recall	93.75%	86.67%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape

Accuracy: 77.62%

	true Fear	true Anxious	class precision
pred. Fear	13	4	76.47%
pred. Anxious	3	11	78.57%
class recall	81.25%	73.33%	

(b) Confusion matrix of recognition using eye region (R2) shape

Accuracy: 73.81%

	true Fear	true Anxious	class precision
pred. Fear	11	3	78.57%
pred. Anxious	5	12	70.59%
class recall	68.75%	80.00%	

(c) Confusion matrix of recognition using mouth region (R3) shape

Figure 5.12: Experiment 2 - Recognition results using shape from eyebrow (R1), eye (R2) and mouth (R3) regions

Accuracy: 67.62%

	true Fear	true Anxious	class precision
pred. Anxious	10	5	66.67%
pred. Fear	5	11	68.75%
class recall	66.67%	68.75%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape and randomised class labels

Accuracy: 64.76%

	true Fear	true Anxious	class precision
pred. Anxious	9	5	64.29%
pred. Fear	6	11	64.71%
class recall	60.00%	68.75%	

(b) Confusion matrix of regression analysis using eyebrow region (R1) shape

Figure 5.13: Experiment 2 - Post-hoc

5.4.3 Experiment 3

The third experiment was to establish how well the facial expressions of anxiety could be classified when infused into a larger set of emotional expressions including 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and ' Neutral'. Thus, the focus of the experiment was not the overall classification but to establish if an anxious expression could be distinguished from 6 prototypical expressions, which included the fear expressions used in Experiment 2.

The recognition results using shape, shape and Gabor magnitudes concatenated, shape and AAM texture parameters concatenated, Gabor magnitudes and AAM texture parameters are given in Figures 5.14, and using R1, R2 and R3 shape in Figure 5.15. It was anticipated that fear and anxiety would have the lowest CA and this was confirmed in the CA of every classifier. One might have expected that fear would have been misclassified most often as anxious, rather than any other expression, and vice-versa but this was not the case. Whilst there was a slight tendency towards anxious expressions being misclassified as fear, they were also misclassified as every other expression, other than those labelled as 'Sad'.

Use of shape concatenated with Gabor magnitudes produced the best overall results, despite the CA performance of AAM texture parameters being slightly less than that of Gabor magnitude. However, as stated previously, with such small sample CA differences of just a few percentage not significant.

Accuracy: 80.45%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	29	1	1	1	0	0	0	90.62%
pred. Fear	1	10	0	0	3	0	3	58.82%
pred. Happy	2	1	28	0	0	0	2	84.85%
pred. Neutral	1	0	1	33	1	7	1	75.00%
pred. Surprise	0	2	0	0	17	0	0	89.47%
pred. Sad	0	0	1	2	0	10	0	76.92%
pred. Anxious	0	2	0	0	0	0	9	81.82%
class recall	87.88%	62.50%	90.32%	91.67%	80.95%	58.82%	60.00%	

(a) Confusion matrix of recognition using shape only

Accuracy: 84.01%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	31	0	1	3	0	1	0	86.11%
pred. Fear	0	12	0	0	3	0	4	63.16%
pred. Happy	0	1	28	0	0	0	2	90.32%
pred. Neutral	2	1	1	32	0	3	1	80.00%
pred. Surprise	0	1	0	0	18	0	0	94.74%
pred. Sad	0	0	1	1	0	13	0	86.67%
pred. Anxious	0	1	0	0	0	0	8	88.89%
class recall	93.94%	75.00%	90.32%	88.89%	85.71%	76.47%	53.33%	

(b) Confusion matrix of recognition using shape and Gabor magnitudes

Accuracy: 82.26%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	29	2	1	1	0	0	0	87.88%
pred. Fear	0	10	0	0	2	0	3	66.67%
pred. Happy	0	0	28	1	1	0	3	84.85%
pred. Neutral	4	0	2	33	0	3	2	75.00%
pred. Surprise	0	2	0	0	18	0	0	90.00%
pred. Sad	0	0	0	1	0	14	0	93.33%
pred. Anxious	0	2	0	0	0	0	7	77.78%
class recall	87.88%	62.50%	90.32%	91.67%	85.71%	82.35%	46.67%	

(c) Confusion matrix of recognition using shape and AAM texture parameters

Accuracy: 74.56%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	28	1	1	4	0	1	0	80.00%
pred. Fear	0	6	0	0	2	0	4	50.00%
pred. Happy	0	1	25	3	0	0	2	80.65%
pred. Neutral	5	2	3	28	2	1	1	66.67%
pred. Surprise	0	1	0	0	17	0	1	89.47%
pred. Sad	0	0	0	1	0	15	0	93.75%
pred. Anxious	0	5	2	0	0	0	7	50.00%
class recall	84.85%	37.50%	80.65%	77.78%	80.95%	88.24%	46.67%	

(d) Confusion matrix of recognition using Gabor magnitudes in eyebrow, eye and mouth regions

Accuracy: 76.33%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	29	1	1	1	1	3	1	78.38%
pred. Fear	1	9	0	1	3	0	5	47.37%
pred. Happy	0	2	27	2	0	0	0	87.10%
pred. Neutral	3	1	3	31	2	3	1	70.45%
pred. Surprise	0	1	0	0	15	0	1	88.24%
pred. Sad	0	0	0	1	0	11	0	91.67%
pred. Anxious	0	2	0	0	0	0	7	77.78%
class recall	87.88%	56.25%	87.10%	86.11%	71.43%	64.71%	46.67%	

(e) Confusion matrix of recognition using AAM texture parameters from entire face

Figure 5.14: Experiment 3 - Recognition results

Accuracy: 76.31%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	30	2	1	2	0	1	0	83.33%
pred. Fear	0	8	0	0	4	3	2	47.06%
pred. Happy	2	1	25	1	0	0	3	78.12%
pred. Neutral	1	0	3	32	0	2	2	80.00%
pred. Surprise	0	1	0	0	15	0	0	93.75%
pred. Sad	0	2	0	1	2	11	0	68.75%
pred. Anxious	0	2	2	0	0	0	8	66.67%
class recall	90.91%	50.00%	80.65%	88.89%	71.43%	64.71%	53.33%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape

Accuracy: 73.96%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	27	0	2	3	0	0	0	84.38%
pred. Fear	0	9	0	0	2	1	2	64.29%
pred. Happy	1	1	26	1	0	0	1	86.67%
pred. Neutral	4	1	3	30	5	5	1	61.22%
pred. Surprise	0	1	0	1	14	0	3	73.68%
pred. Sad	1	1	0	1	0	11	0	78.57%
pred. Anxious	0	3	0	0	0	0	8	72.73%
class recall	81.82%	56.25%	83.87%	83.33%	66.67%	64.71%	53.33%	

(b) Confusion matrix of recognition using eye region (R2) shape

Accuracy: 75.72%

	true Anger	true Fear	true Happy	true Neutral	true Surprise	true Sad	true Anxious	class precision
pred. Anger	27	1	2	3	1	2	1	72.97%
pred. Fear	0	10	0	0	2	0	3	66.67%
pred. Happy	0	0	24	1	0	0	2	88.89%
pred. Neutral	6	0	4	31	0	2	1	70.45%
pred. Surprise	0	1	0	0	16	0	1	88.89%
pred. Sad	0	1	0	1	1	13	0	81.25%
pred. Anxious	0	3	1	0	1	0	7	58.33%
class recall	81.82%	62.50%	77.42%	86.11%	76.19%	76.47%	46.67%	

(c) Confusion matrix of recognition using mouth region (R3) shape

Figure 5.15: Experiment 3 - Recognition results using shape from eyebrow (R1), eye (R2) and mouth (R3) regions

5.4.4 Experiment 4 - Baseline

The ultimate objective of the fourth experiment was to test whether the classifier built from the Cohn-Kanade database of images, in Experiment 1, could be used to predict the facial expressions in the Feedtum database of images, which were recorded using different subjects and under different lighting conditions. Classifiers trained using shape vectors, shape and Gabor texture, Gabor texture, eye (R1), eye (R2) and mouth (R3) regions were used.

As a preliminary step, a CA test was conducted using an SVM built from prototypical facial expressions of 'Fear', 'Anger', 'Happy', 'Sad', 'Surprise', and 'Neutral', sourced entirely from images within the Feedtum database. This was so that the CA could be compared to that attained in Experiment 1, which used the Cohn-Kanade database. This is referred to here as the "baseline" experiment. The "baseline" was acquired by SVM classification, similar to Experiment 1, using 5-fold cross validation.

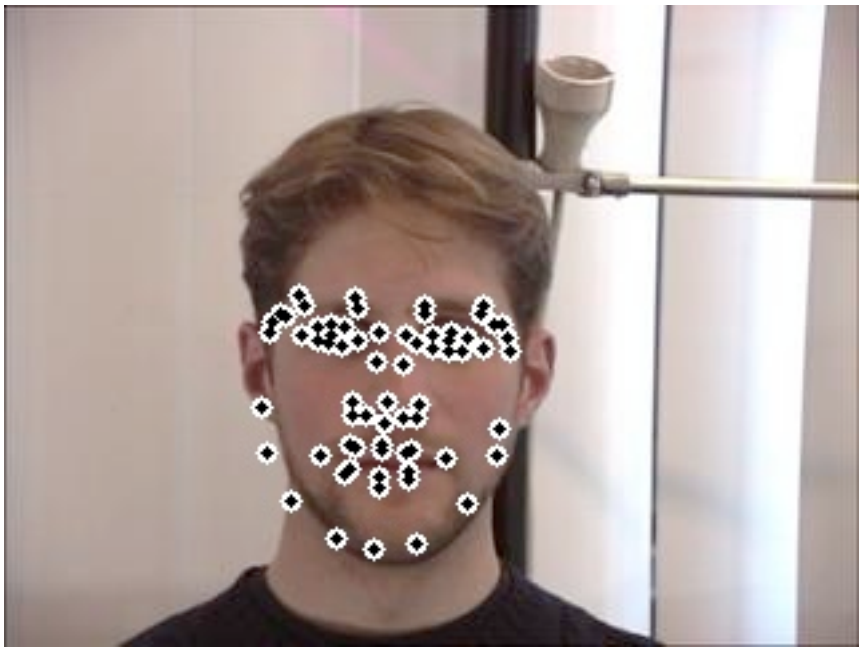
At first, the AAM that had been built for Experiment 1, using images from the Cohn-Kanade database, was used to fit the landmark points to images from the Feedtum database. As can be seen in 5.16(a), the landmark points were not placed perfectly. This was likely due to the relatively small number of samples used from the Cohn-Kanade database (≈ 200) to build the initial AAM ($500 - 1,000$ would be much better). Since the object of the experiments was not to test the AAM per se, to overcome this, a specific Feedtum AAM was built. The accuracy of fitting with the new model was much better and an example is shown at 5.16(b).

The baseline recognition results using shape, shape and Gabor magnitudes concatenated, shape and AAM texture parameters concatenated, Gabor magnitudes and AAM texture parameters are given in Figures 5.17, and using R1, R2 and R3 shape in Figure 5.18.

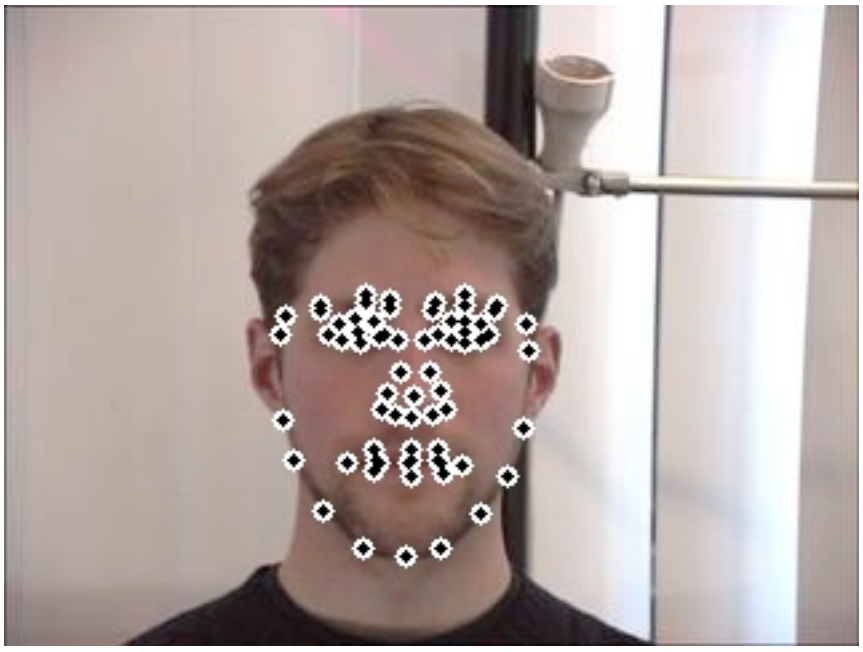
Figures 5.17 and 5.18 show that the CA was lower than in Experiment 1. However,

the number of samples is smaller and a relatively larger variation in the CA will result from each misclassification. Notwithstanding that, a plausible reason for the lower CA is that the expressions portrayed in the Feedtum database are much less pronounced. Figures 5.19(a) and 5.19(b) are reported within the Feedtum database transcriptions,[8] as the apex of expressions of anger and fear respectively. One would think that human judges might have difficulty in correctly classifying these expressions. In addition, the intensity of the expressions are clearly in contrast to that shown in Figure 5.4.

[8]Feedtum metadata transcriptions at `http://www.mmk.ei.tum.de/~waf/fgnet/metadata-feedtum.csv`, image files are anger/0003_3/p_086.jpg and fear/0007_2/p_110.jpg, last access 1 March 2010



(a) Feedtum image fitted using general AAM trained on Cohn-Kanade database



(b) Feedtum image fitted using specific AAM trained on Feedtum database

Figure 5.16: Experiment 4 - Images fitted using generalised and specific AAMs

Accuracy: 55.71%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	7	0	0	0	1	0	87.50%
pred. Fear	0	6	0	1	1	1	66.67%
pred. Happy	1	2	10	2	1	1	58.82%
pred. Sad	0	1	0	2	2	4	22.22%
pred. Surprise	0	0	2	1	5	0	62.50%
pred. Neutral	1	2	2	4	1	9	47.37%
class recall	77.78%	54.55%	71.43%	20.00%	45.45%	60.00%	

(a) Confusion matrix of recognition using shape only

Accuracy: 74.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	8	0	0	0	0	1	88.89%
pred. Fear	0	8	0	0	2	1	72.73%
pred. Happy	0	1	11	0	1	1	78.57%
pred. Sad	0	1	0	6	0	1	75.00%
pred. Surprise	0	0	2	1	8	0	72.73%
pred. Neutral	1	1	1	3	0	11	64.71%
class recall	88.89%	72.73%	78.57%	60.00%	72.73%	73.33%	

(b) Confusion matrix of recognition using shape and Gabor magnitudes

Accuracy: 54.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	7	2	0	0	2	0	63.64%
pred. Fear	1	4	2	1	1	1	40.00%
pred. Happy	0	1	9	0	2	0	75.00%
pred. Sad	0	0	1	4	1	4	40.00%
pred. Surprise	0	1	2	1	4	0	50.00%
pred. Neutral	1	3	0	4	1	10	52.63%
class recall	77.78%	36.36%	64.29%	40.00%	36.36%	66.67%	

(c) Confusion matrix of recognition using shape and AAM texture parameters

Accuracy: 64.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	7	0	0	0	0	3	70.00%
pred. Fear	1	7	0	1	3	1	53.85%
pred. Happy	0	0	11	0	1	1	84.62%
pred. Sad	0	2	0	5	0	2	55.56%
pred. Surprise	0	1	1	2	7	0	63.64%
pred. Neutral	1	1	2	2	0	8	57.14%
class recall	77.78%	63.64%	78.57%	50.00%	63.64%	53.33%	

(d) Confusion matrix of recognition using Gabor magnitudes in eyebrow, eye and mouth regions

Accuracy: 64.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	7	1	3	2	0	1	50.00%
pred. Fear	0	6	0	3	0	1	60.00%
pred. Happy	1	1	10	0	0	0	83.33%
pred. Sad	0	0	0	2	2	1	40.00%
pred. Surprise	0	0	0	0	8	0	100.00%
pred. Neutral	1	3	1	3	1	12	57.14%
class recall	77.78%	54.55%	71.43%	20.00%	72.73%	80.00%	

(e) Confusion matrix of recognition using AAM texture parameters from entire face

Figure 5.17: Experiment 4 - Baseline recognition results using Feedtum database

Accuracy: 45.71%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	6	1	0	0	1	0	75.00%
pred. Fear	0	6	2	1	3	1	46.15%
pred. Happy	0	2	8	1	2	2	53.33%
pred. Sad	1	2	0	2	0	6	18.18%
pred. Surprise	1	0	2	1	4	0	50.00%
pred. Neutral	1	0	2	5	1	6	40.00%
class recall	66.67%	54.55%	57.14%	20.00%	36.36%	40.00%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape

Accuracy: 57.14%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	8	1	0	0	1	0	80.00%
pred. Fear	0	4	1	0	5	2	33.33%
pred. Happy	0	1	10	1	1	0	76.92%
pred. Sad	0	2	1	5	0	4	41.67%
pred. Surprise	0	2	2	2	4	0	40.00%
pred. Neutral	1	1	0	2	0	9	69.23%
class recall	88.89%	36.36%	71.43%	50.00%	36.36%	60.00%	

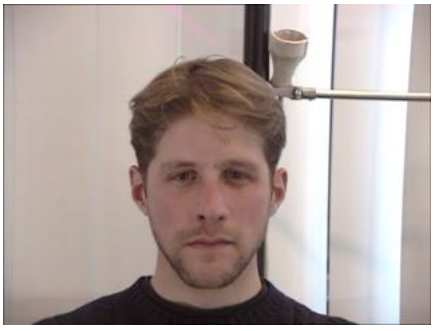
(b) Confusion matrix of recognition using eye region (R2) shape

Accuracy: 58.57%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	7	0	0	0	2	0	77.78%
pred. Fear	1	5	1	0	0	3	50.00%
pred. Happy	0	1	11	0	2	0	78.57%
pred. Sad	0	2	0	3	1	3	33.33%
pred. Surprise	1	0	2	1	6	0	60.00%
pred. Neutral	0	3	0	6	0	9	50.00%
class recall	77.78%	45.45%	78.57%	30.00%	54.55%	60.00%	

(c) Confusion matrix of recognition using mouth region (R3) shape

Figure 5.18: Experiment 4 Baseline Feedtum database - Recognition results using shape from eyebrow (R1), eye (R2) and mouth (R3) regions



(a) Feedtum image of anger expression

(b) Feedtum image of fear expression

Figure 5.19: Experiment 4 - Feedtum images of anger and fear.

5.4.5 Experiment 4 - Classification against Cohn-Kanade SVM

Next, an attempt was made to automatically classify expressions in images from the Feedtum database against the SVM models built in Experiment 1 using images from the Cohn-Kanade database. As in the first part of the experiment, the images were fitted with the Feedtum-specific AAM.

The PCA coefficients, after performing PCA on the Gabor filter magnitudes, were obtained by projecting into the eigenspace that was created in Experiment 1. Similarly, the scaling parameters obtained in Experiment 1 were applied when scaling the features obtained in Experiment 4.

The recognition results for the second part of Experiment 4 using shape, shape and Gabor magnitudes concatenated, shape and AAM texture parameters concatenated, Gabor magnitudes and AAM texture parameters is given in Figure 5.20, and using R1, R2 and R3 shape in Figure 5.21.

The CA results from this experiment were all very low. One reason that was considered is the practice of scaling the data prior to building the classifier. *Scaling* is used not only in SVM, but also in Neural Network classification. The case for scaling is presented in [Hsu 03]:

> "The main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Because kernel values usually depend on the inner products of feature vectors, e.g. the linear kernel and the polynomial kernel, large attribute values might cause numerical problems. We recommend linearly scaling each attribute to the range $[-1; +1]$ or $[0; 1]$."

A key requirement in the use of scaling is that the same method used to scale the training data is applied to the test data. Operationally, the scaling parameter that is

found for each feature during training needs to be saved and then applied to the test data. One of the criticisms of this approach is that it "overtrains" the data. It has the potential to increase the CA in exercises where the training set is used for testing after the entire training set has been scaled, i.e. no new previously unseen data is introduced in testing. This is sometimes referred to as, "peeping the data". When applied to new and previously unseen data, there is no guarantee that the scaling parameters will have a valid relationship, as might have been the case in the second part of Experiment 4.

Experiments 1, 2, 3, and 4-baseline were all conducted using scaled and normalised data. To get some idea of the effect that scaling might have had on the second part of Experiment 4, an informal, post-hoc classification exercise was conducted, whereby the data was normalised but not scaled. The results, which are not included here, did not vary significantly and scaling was ruled out as a major contributor to the poor CA performance.

Accuracy: 32.86%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	6	6	1	8	4	9	17.65%
pred. Fear	2	2	1	0	2	2	22.22%
pred. Happy	0	1	11	0	2	3	64.71%
pred. Sad	0	0	0	0	0	0	0.00%
pred. Surprise	1	2	1	0	3	0	42.86%
pred. Neutral	0	0	0	2	0	1	33.33%
class recall	66.67%	18.18%	78.57%	0.00%	27.27%	6.67%	

(a) Confusion matrix of recognition using shape only

Accuracy: 41.43%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	9	2	1	4	0	8	37.50%
pred. Fear	0	5	3	2	4	4	27.78%
pred. Happy	0	1	10	1	3	2	58.82%
pred. Sad	0	0	0	0	0	0	0.00%
pred. Surprise	0	3	0	0	4	0	57.14%
pred. Neutral	0	0	0	3	0	1	25.00%
class recall	100.00%	45.45%	71.43%	0.00%	36.36%	6.67%	

(b) Confusion matrix of recognition using shape and Gabor magnitudes

Accuracy: 17.14%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	6	1	5	0	0	0	50.00%
pred. Fear	3	5	1	6	5	9	17.24%
pred. Happy	0	4	1	4	6	6	4.76%
pred. Sad	0	0	0	0	0	0	0.00%
pred. Surprise	0	1	0	0	0	0	0.00%
pred. Neutral	0	0	7	0	0	0	0.00%
class recall	66.67%	45.45%	7.14%	0.00%	0.00%	0.00%	

(c) Confusion matrix of recognition using Gabor magnitudes in eyebrow, eye and mouth regions

Figure 5.20: Experiment 4 - Recognition results using SVMs built in experiment 1

Accuracy: 25.71%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	0	0	5	5	0	9	0.00%
pred. Fear	0	4	0	2	0	1	57.14%
pred. Happy	0	1	2	1	0	1	40.00%
pred. Sad	1	0	1	0	1	0	0.00%
pred. Surprise	8	5	5	1	10	2	32.26%
pred. Neutral	0	1	1	1	0	2	40.00%
class recall	0.00%	36.36%	14.29%	0.00%	90.91%	13.33%	

(a) Confusion matrix of recognition using eyebrow region (R1) shape

Accuracy: 14.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	3	4	11	9	2	10	7.69%
pred. Fear	6	4	1	1	5	5	18.18%
pred. Happy	0	1	0	0	0	0	0.00%
pred. Sad	0	0	0	0	1	0	0.00%
pred. Surprise	0	0	0	0	3	0	100.00%
pred. Neutral	0	2	2	0	0	0	0.00%
class recall	33.33%	36.36%	0.00%	0.00%	27.27%	0.00%	

(b) Confusion matrix of recognition using eye region (R2) shape

Accuracy: 24.29%

	true Anger	true Fear	true Happy	true Sad	true Surprise	true Neutral	class precision
pred. Anger	6	5	10	2	2	6	19.35%
pred. Fear	0	0	2	0	3	0	0.00%
pred. Happy	0	0	1	0	0	0	100.00%
pred. Sad	1	2	0	6	0	6	40.00%
pred. Surprise	0	0	0	0	1	0	100.00%
pred. Neutral	2	4	1	2	5	3	17.65%
class recall	66.67%	0.00%	7.14%	60.00%	9.09%	20.00%	

(c) Confusion matrix of recognition using mouth region (R3) shape

Figure 5.21: Experiment 4 - Recognition results using shape from eyebrow (R1), eye (R2) and mouth (R3) regions against SVMs built in experiment 1

5.5 Conclusions and Evaluation

5.5.1 Hypothesis 1

Using computerised facial expression recognition techniques, anxious expressions can be differentiated from fearful expressions.

Fear and anxious expressions were sourced from a set of expressions having action units corresponding to both fear and anxiety. In the absence of a verified database of anxious expressions, the Cohn-Kanade database was a useful starting point, but it is not clear how the slightly pronounced nature of the fear expressions might have affected the results. A larger, more ecologically valid set of training data, both for the fearful and the anxious expressions, would provide a more reliable set of results to support this hypothesis. Obviously, procuring such a database is a major undertaking. Nevertheless, the results from Experiment 2 suggest that anxiety *can* be automatically distinguished from fear.

The results from Experiment 2 suggest that shape is much better discriminator than texture in the differentiation of fear and anxious expressions. Although much more effort is need to draw a conclusion, human judges may rely, to a large extent, on the shape of eyebrow region when trying to differentiate between fear and anxious expression.

5.5.2 Hypothesis 2

Using computerised facial expression recognition techniques, anxious expressions can be differentiated from a larger set of prototypical expressions.

Anxious expressions were not classified with a high degree of accuracy. In Experiment 3, although they were quite often misclassified as fear, they were also misclassified as every other expression except those labelled as ‘Sad’. Again, a much

larger sample set is needed to draw a conclusion. One possibility to improve accuracy would be to use an ensemble of classifiers, recognising, first, a more broadly labelled set of expressions which combined both fear and anxiety under one label, and, second, performing a binary classification between fear and anxiety with the use of just shape.

5.5.3 Question Set 1

How does facial expression recognition performance, i.e. CA, vary when using:

- *the location of facial landmark points only;*
- *the location of facial landmark points concatenated with Gabor magnitude ;*
- *the location of facial landmark points concatenated with AAM texture parameters;*
- *AAM texture parameters only; and*
- *Gabor magnitude only?*

In Experiment 1, shape when concatenated with Gabor magnitudes produced the highest CA but, given the number of samples, there was no significant percentage advantage over the use of shape concatenated with AAM texture parameters, or use of only the Gabor magnitudes. Again in Experiment 3, shape concatenated with Gabor magnitudes produced the highest CA, but in this case, it outperformed the use of only the Gabor magnitudes.

Shape concatenated with Gabor magnitudes also produced the best CA in the "baseline" Experiment 4. This was a slightly odd result where, even though the overall CA of Gabor magnitudes and AAM texture parameters was the same, concatenating Gabor magnitudes to shape improved the CA (achieved by shape) by 20%, yet concatenating AAM texture parameters to shape resulted in a slight decrease in CA. One could speculate about the manner in which the feature data from the Gabor magnitudes better

compliments the shape data. However, the topic of how best to fuse heterogeneous feature sets, i.e. shape and texture data, in order to achieve the best CA needs far more investigation.

5.5.4 Question Set 2

In the NXS system, described in Chapter 4, the face is subdivided into three regions:

- *R1 - The eyebrow region;*
- *R2 - The eye region; and*
- *R3 - the mouth region*

There are two parts to this question:

1. *is one facial region* generally *more reliable for recognition of expressions?*
   It could not concluded that there was *generally* one region of the face that was consistently associated with a CA rate, higher than the other regions.

2. *is one facial region more reliable for recognition of a* specific *expression?*
   One surprising result was the CA based on just the eyebrow (R1) shape features, shown at Figure 5.12(a). Although purely speculative, it suggests that, when faced with a binary choice to differentiate fear and anxious expressions, a viewer will place more emphasis on the eyebrow features.

   Knowing which facial regions are used by humans to differentiate facial expressions would be useful for the field of facial expression recognition.

5.5.5 Question Set 3

Would performance be sufficient to achieve on-line recognition of facial expressions, in a video running at 30 frames per second? It was clear when running the experiments that the execution time to classify the sets of images increased noticeably as the number of Gabor filters increased. However, the results of this step are only anecdotal and have not been formally recorded.

5.6 Overall Evaluation

In summary, despite the lack of samples and natural data, the results suggest that the recognition of anxious expressions is possible but becomes more difficult when fearful expressions are also present. The difficulty increases when more primary expressions are added to the classification problem. The exercise demonstrates that facial expression classification is, in general, a difficult task and, in some situations, in the absence of contextual information and/or temporal data revealing facial dynamics, may not even be possible. Moreover, even with contextual and temporal evidence present, the fact that a prototypical expression can take many forms, e.g. a 'happy' expression can be portrayed with or without opening the mouth, compounds the degree of difficulty. Any attempt at recognition may not be reliable without the presence of semantic information.

The second part of Experiment 4 demonstrated that, even using two popular and creditable databases, a classifier built from images from one database, did not achieve a high CA in predicting expressions from the other, despite Gabor filtering being reasonably invariant to lighting conditions. This echoes the preliminary results reported in [Whitehill 09] (albeit, much worse).

In this instance, it seems that the notion of using a database recorded under one set of conditions to recognise expression in a database recorded under a different set

of lighting conditions was overly simplistic. Part of the solution, as demonstrated by [Whitehill 09] would be to 1) vastly increase the size of the training set; and 2) include image samples recorded under a wide range of lighting conditions.

Gabor filter processing proved cumbersome. Several studies have attempted to improve the efficiency and reduce the computational overhead of it, either by:

- automatically selecting the best features to convolve with the Gabor filters.
  There are many proposed schemes for doing this [Littlewort 06, Shen 05, Zhou 06, Zhou 09] (although in [Zhou 06, Zhou 09] the processing is applied to reduce the Gabor feature set *after* convolution); or

- by optimising the Gabor wavelet basis for convolution, e.g. use of genetic algorithms (GA)
  The GA approach has shown some promise in addressing both problems but, ironically, it too imposes a computational burden [Li 07, Tsai 01, Wang 02].

In this set of experiments, all three facial regions were convolved using the same basis for convolution and no attempt was made to optimise for individual facial regions. [Ro 01] suggests a way to improve Gabor filter performance is to reduce the computation load by selecting the Gabor filter basis in a pattern-dependent way. Given the similarity of the images patches used in the this set of experiments, it is difficult to envisage that using specific Gabor filter parameters would yield a significantly higher CA or faster convolution times.

During the setup stage of the experiments, a great deal of effort was spent trying to optimise the system in areas such as Gabor filter settings, facial region image sizes, data scaling and PCA process. The impact on percentage CA from any of these measures was not as significant as changes to the SVM parameters, and not comparable to the variation in Experiment 2 between shape-based and texture-based classification. Re-architecting

the recognition process to make use of a hierarchy of specialised classifiers, depending on the expression, would be much likelier improve the CA.

You largely constructed your depression. It wasn't given to you. Therefore, you can deconstruct it.

Albert Ellis

6

Depression Analysis Using Computer Vision

6.1 Introduction and Motivation for Experiments

Facial expressions convey information about a person's internal state. The ability to correctly interpret another's facial expressions is important to the quality of social interactions. In phatic speech, in particular, the affective facial processing loop, i.e. interpreting another's facial expressions then responding with one's own facial expression, plays a critical role in the ability to form and maintain relationships.

The affective facial processing loop was discussed in Subsection 2.6.2 of Chapter 3. A predisposition to misinterpret expressions often underlies dysphoria, e.g. anxiety and depression. More specifically, the impaired inhibition of negative affective material could be an important cognitive factor in depressive disorders. Recent studies using *fMRI*, have provided some evidence that the severity of depression in MDD groups correlates with increased neural responses to pictures with negative content [Fu 08, Lee 07]. In turn, this bias to favour negative material by depressed patients has been shown to be signaled in their resulting facial expressions.

This chapter explores the feasibility of using state-of-the-art, low-cost, unobtrusive techniques to measure facial activity and expressions, and then applies them to a real-life application. The motivation for the experiments presented in this chapter is to:

- prove the concepts described in previous chapters can be applied in a practical experiment involving the analysis of video;
- taking depression as an example, try to establish if automatic expression analysis can be used to track facial activity and expression in video; and
- test if automatic facial activity and expression analysis could be used in a real-life application.

Section 6.2 states the hypotheses that are to be tested. Section 6.3 describes the methodology used in the experiments. The results are presented in Section 6.4 and Section 6.5 concludes the chapter.

6.2 Hypotheses

To sharpen the focus of the experiments, the following hypotheses and questions were generated on the basis of the motivation for the experiments and the literature review in Chapter 3:

1. When viewing the stimuli, patients with a clinical diagnosis of unipolar melancholic depression will show less facial activity than control subjects and patients with other types of depression; and

2. When viewing the stimuli, patients with a clinical diagnosis of unipolar melancholic depression will show *less repertoire* of facial expressions than control subjects and patients with other types of depression.

6.3 Methodology

6.3.1 Experimental setup

The experiment described in this chapter is currently incorporated in a collaborative project at the Black Dog Institute, Sydney, Australia. 27 participants were recorded using a high-quality video camera [AVT] while viewing affective content and answering emotive questions. Table 6.1 summarises the ratios of female to male and controls to patients in the interview,[1] while Table 6.2 summarises the subjects in the exercise.

	Control	Patient	Total
Male	7	7	14
Female	9	4	13
Total	16	11	27

Table 6.1: Participant details

The experimental mood/affect induction paradigm is presented to participants of the trial by way of an interactive computer package [NBS], in a setup conceptually similar to that in Figure 6.1.

In the paradigm, the participant's facial and vocal expressions are recorded as they face the computer display. In its entirety, the interview or session takes around 30 minutes. The experimental paradigm includes:

[1]The terms "interview", "session" and "recording" are used synonymously unless otherwise stated.

Patient Id	Gender	Diagnosis Clinical	Diagnosis MINI	Control	Patient
Co_m_01	m			1	
Co_m_02	m			1	
Co_f_03	f			1	
Co_m_04	m			1	
Co_f_05	f			1	
Co_f_06	f			1	
Co_f_07	f			1	
Co_f_08	f			1	
Co_f_09	f			1	
Co_f_10	f			1	
Pa_m_UP-Mel_01	m	Unclear - Grey	UP-Mel		1
Pa_m_UP-Mel_02	m	UP-Mel	UP-Mel		1
Pa_m_UP-Mel_03	m	UP-Mel	UP-Mel		1
Pa_m_UP-Mel_04	m	UP-Mel	UP-Mel		1
Co_m_11	m			1	
Co_f_12	f			1	
Co_m_13	m			1	
Co_m_14	m			1	
Co_f_15	f			1	
Co_m_16	m			1	
Pa_m_UP-Mel_05	m	UP-MEL	UP-MEL		1
Pa_f_UP-NonMel_06	f	BP2, meancholic	UP NON-MEL		1
Pa_m_Unkown_07	m	Unkown	Unkown		1
Pa_f_UP_BP2_08	f	BP2	BP2		1
Pa_f_UP-NonMel_09	f	no clinical diagnosis	UP-NON MEL		1
Pa_f_UP-Mel_10	f	UP-MEL	UP-MEL		1
Pa_m_PD_11	m	PD	UP-MEL		1
Total Participants				**16**	**11**

Participant Ids in the table take the form
XX_G_CD_ID

XX - "Co" for Control or "Pa" for Patient, G - Gender, CD - Diagnosis (Patients only), ID - Sequential Id number (control and patients numbered separately)

Table 6.2: Participant details and diagnosis

Watching movie clips Short movie segments of around two minutes each, some positive and some negative are presented. With the exception of one clip, each movie has previously been rated for its affective content [Gross 95].

Watching and rating International Affective Picture System (IAPS) pictures Pictures from the IAPS [Lang 05] compilation are presented and participants rate each image as either positive or negative. Reporting logs enable correlation of the image presentations, the participant's ratings, and their facial activity.

Reading sentences containing affective content Two sets of sentences used in the study by [Brierley 07, N. Medforda et al 05] are read aloud. The first set contains emotionally arousing "target" words. The second set repeats the first, with the

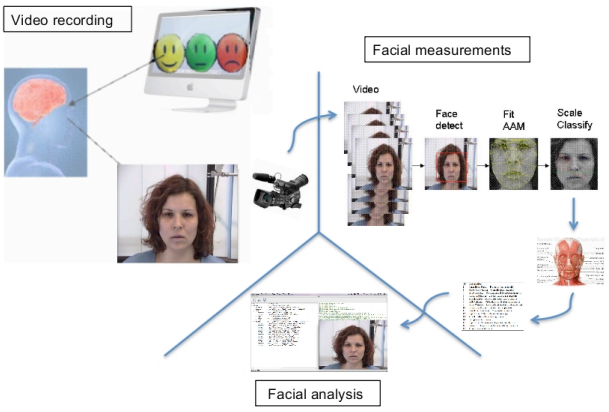


Figure 6.1: Experimental setup

"target" words replaced by well-matched neutral words.

Answering specific questions Finally, participants are asked to describe events that had aroused significant emotions. For instance, ideographic questions such as, "Describe an incident that made you feel really sad."

It is important to note that this chapter and thesis reports only on the first section of the experimental paradigm, i.e. watching movie clips. The initial experimental setup consisted of the movies listed in Table 6.3, with intended induced emotion shown in parenthesis. This is referred to as the "Old Paradigm". After some evaluation, the movie sequence was changed and this is referred to as the "New Paradigm", as shown in Table 6.4.

Movie (emotion)
Bill Crosby (Happy)
The Champ (Sad)
Weather (Happy)
Sea of Love (Surprise)
Cry Freedom (Anger)

Table 6.3: Old paradigm - movie list

A summary of the numbers of participants in each paradigm is shown at Table 6.5.

Figure 6.2 shows a control subject being recorded as he watches the movie Cry Freedom. The interview, from a participant's view, is shown in Figure 6.3. When the

Movie (emotion)
Bill Crosby (Happy)
The Champ (Sad)
Weather (Happy)
Silence of the Lambs (Fear)
Cry Freedom (Anger)
The Shining (Fear)
Capricorn One (Surprise)

Table 6.4: New paradigm - movie list

	Control	**Patient**	**Total**
Old Paradigm	10	4	14
New Paradigm	6	7	13
Total	16	11	27

Table 6.5: Participant summary

interview is in progress, the Research Assistant can monitor the session on the laptops shown in Figure 6.4. The laptop on the left in Figure 6.4 is displaying a frame from the movie clip Bill Cosby, while the one on the show the recording of the participant.

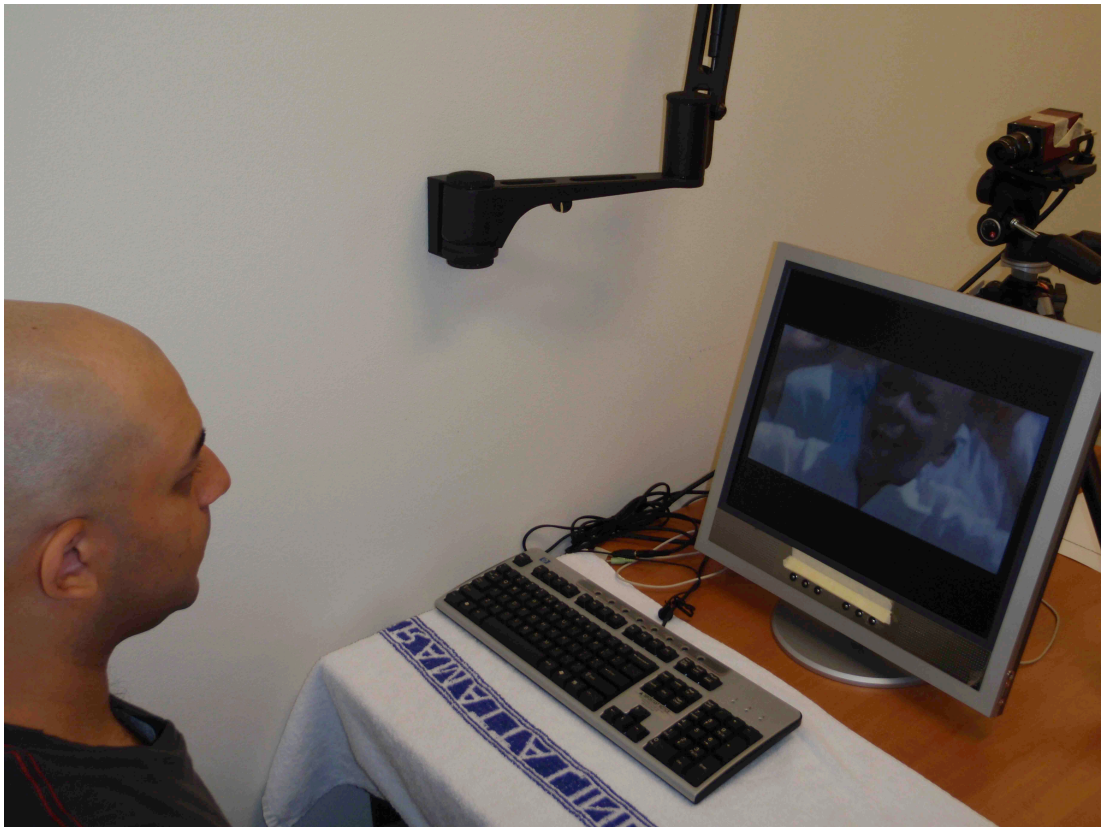


Figure 6.2: Control subject watching video clip - Cry Freedom

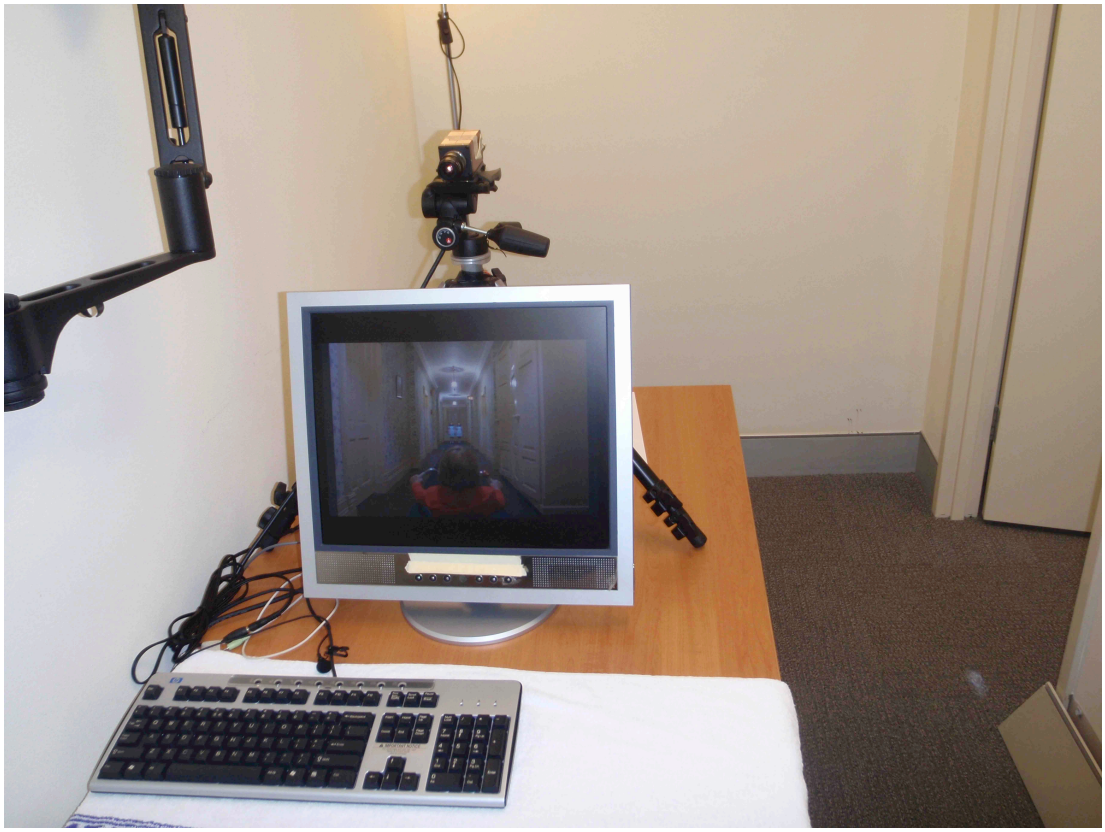


Figure 6.3: Participant's view of the interview (video clip - Silence of the Lambs)

6.3.2 System Setup and Processing

The experimental mood/affect induction paradigm is presented to participants by way of an interactive computer package [NBS]. Videos of the subjects viewing the paradigm are recorded with an Allied Vision Technologies (AVT) Pike 100C camera [AVT], capturing 800×600 pixels images at 24.94 frames per second. The camera is connected through a Firewire IEEE 1394b connection to an Apple MacBook Pro. The videos are recorded in very high quality in order to support future research. At this point in time, however, prior to analysis, the videos are exported to Microsoft's AVI format. This is because AVI is the only container format supported by OpenCV [OpenCV], which is used in the analysis software (the *NXS* system).

Once the video has been recorded, analysis begins by capturing sample frames

Figure 6.4: Laptops displaying stimuli and recording of subject

from each video, which are used to 1) build a person-specific AAM for each person (person-specific AAMs give better fitting quality [Gross 05] and there is no need at this juncture to have generic AAMs); and 2) construct an SVM classifier for each person's emotional expressions (which are rated subjectively at this point in time). For the reasons explained in Section 5.3.2 of Chapter 5, the IEBM method [Saragih 06] of building the AAM and fitting the model to the image has been chosen, as was the LIBSVM [Chang 01] implementation of SVM.

Once the AAM and SVM have been built, frames are then captured from the video at 200 ms intervals (this seemed a reasonable choice of interval based, anecdotally, on the speed of movement of facial features, however, there is no restriction on the rate). As each frame is captured, frontal facial images are detected using the Viola and Jones [Viola 01] technique to determine the global location of the face in the image. Next,

the AAM is used to track and measure the local shape and texture features. As described in Chapter 5, "shape" refers to the collective landmark points, which are captured as a set of normalised, x, y Cartesian coordinates. The features are then used to classify the expressions using an SVM classifier [Chang 01]. All outputs are stored within the system to allow for post-processing, which is described in Subsections 6.3.2 and 6.3.2.

Measuring Facial Activity

With the raw feature data captured, Algorithm 4 is used to measure the collective movement of the landmark points between each frame. Although not shown in the algorithm, extreme movements are ignored if the movement falls outside of predefined thresholds. This is to cater for situations where the face detection in a frame has failed and the AAM "fitting" has not converged, which typically leaves the landmark points spread around the image.

Algorithm 4: Measuring and facial activity

input: set of facial landmark points for every image for each video

int i = 0

int j = 0

for *each video* **do**

  for *each set of facial landmark points* **do**

    tempx ← distance between x coordinates of this and previous frame

    tempy ← distance between y coordinates of this and previous frame

    `// one.norm is the sum of absolute values`

    allSubjectsMovements[i][j] ← one.norm(tempx) + one.norm(tempy)

    j++;

  i++

output: set of scalar values representing distance between each set of landmark points for each video

The *NXS* system outputs the facial activity measurements for each subject into a file of comma-separated values, which can then be imported to a third-party product for further analysis, e.g. Excel.

Tracking Prototype Expressions

The classified expression is stored within the system for each captured image. The images, captured at a rate of one every 200 ms and marked up with the automatically fitted landmark points, can be assembled as an image sequence and played as a short video. This assists in verifying that the AAM has fitted properly, and consistently between frames. The coloured slider below the images, shown in Figure 6.5, provides a visual representation of the facial expressions over the course of the interview. The colours represent the classification, pink - happy, blue - sad and white - neutral. In Figure 6.5, the slider has been positioned to a period of happy expressions, which coincides with the Bill Cosby film clip. This allows a visual confirmation that the expression recognition has worked successfully. Each reconstructed participant "movie" can be played individually or along with several other participant "movies", thus allowing a comparison of participants' facial responses at a specific time in the interview.

The *NXS* system outputs the list of classifications for each subject into a file of comma-separated values, which can then be imported to a third-party product for further analysis, e.g. Excel.

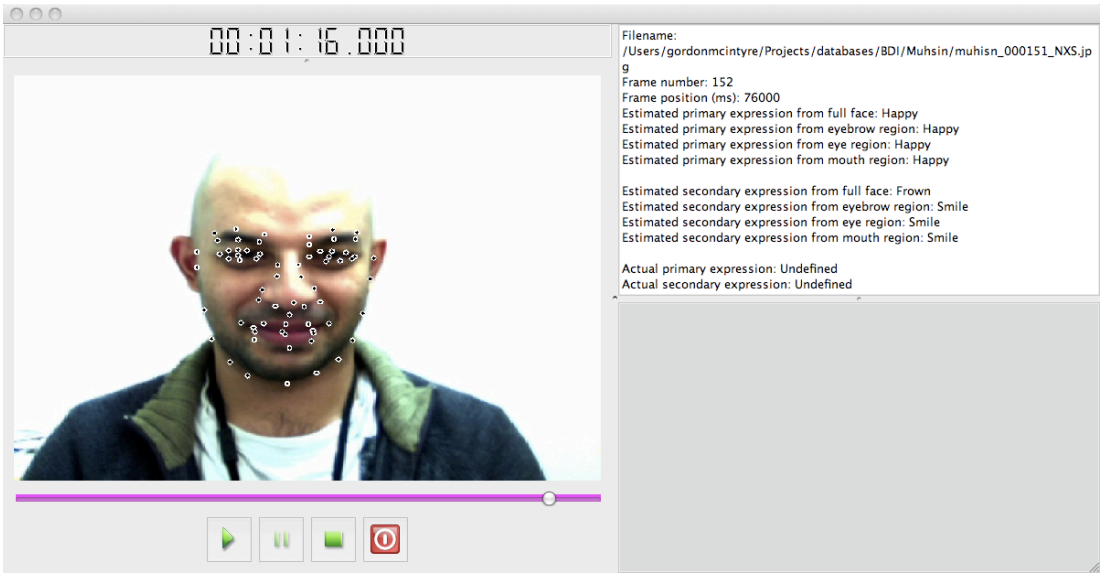


Figure 6.5: The *NXS* System - Replaying captured images

6.4 Presentation and Analysis of Data

6.4.1 Introduction

Two sets of results are presented, one for each paradigm, in the form of charts at the end of the chapter. The data used to construct the charts is available in Appendix B.

6.4.2 Old Paradigm

Figure 6.6 displays a stacked column chart of facial activity for every Old Paradigm participant over the entire video clip session. Overall, control subjects have tended to have a higher facial activity score than patients. Figure 6.7 displays a clustered column chart of the same data as Figure 6.6 facial activity. This is simply another view of the data in Figure 6.6. Figure 6.8 is a comparison of the accumulated facial activity over time, across the entire series of movie clips. Each sub-figure in Figure 6.9 shows the facial activity specific to the relevant movie clip.

Overall, control subject, Co_m_04 had a low facial activity score, but his score during the "sad" stimuli was in keeping with the other controls. This was an interesting result, since, anecdotally, the "sad" movie clip (The Champ), seemed to evoke strong feelings in all of the other control subjects. On viewing Co_m_04's recording, he seems of Asian appearance and it is not known if there was a cultural factors influencing the results.

Figure 6.10 shows the number of happy expressions displayed by each subject, for each film clip over time over the entire series of clips. Figure 6.11 shows the number of sad expressions displayed by each subject, for each film clip over time over the entire series of clips. Figure 6.12 shows the number of neutral expressions displayed by each subject, for each film clip over time over the entire series of clips.

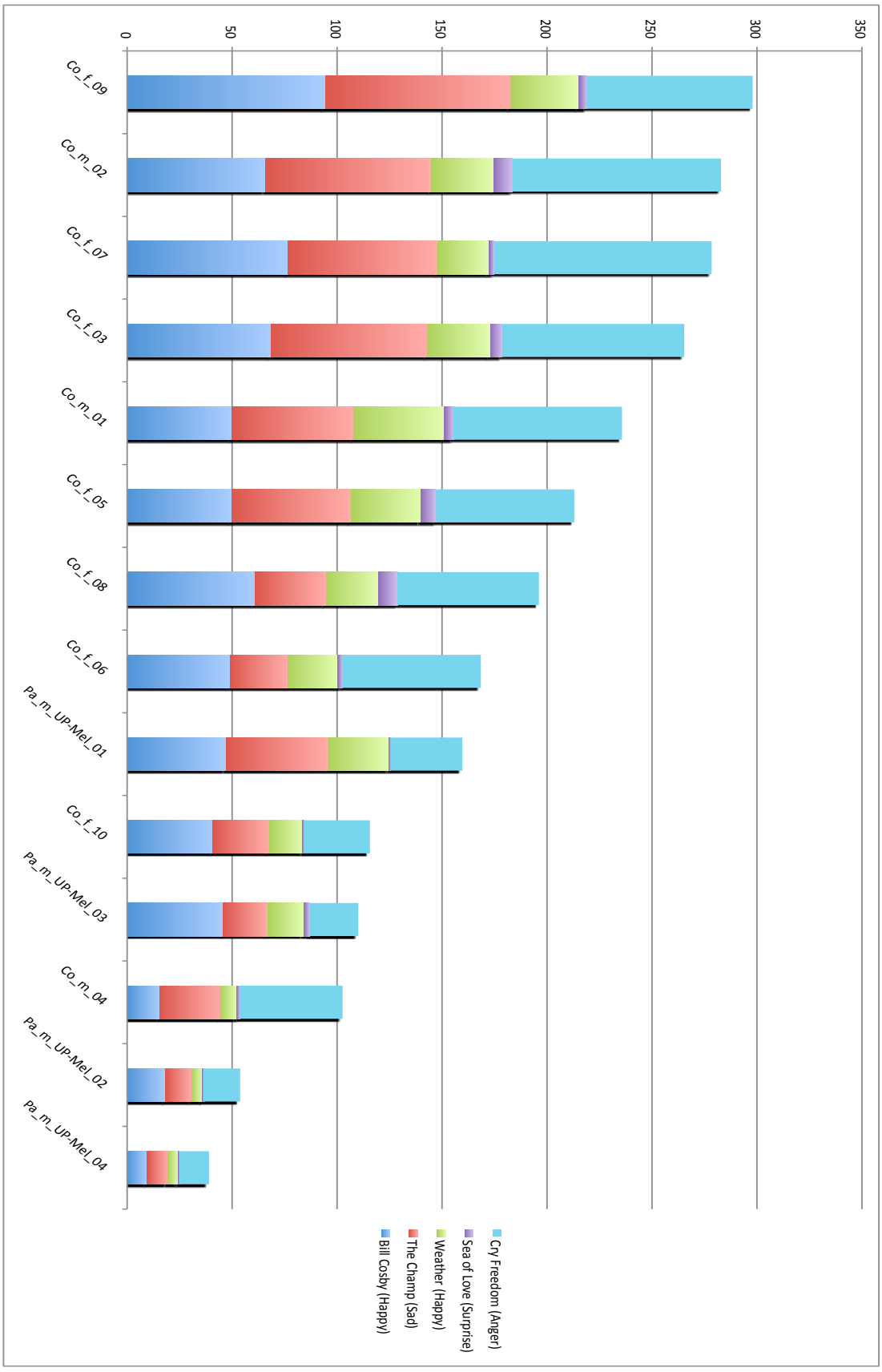


Figure 6.6: Old Paradigm - Stacked column chart comparing facial activity (Co - Control, Pa - Patient)

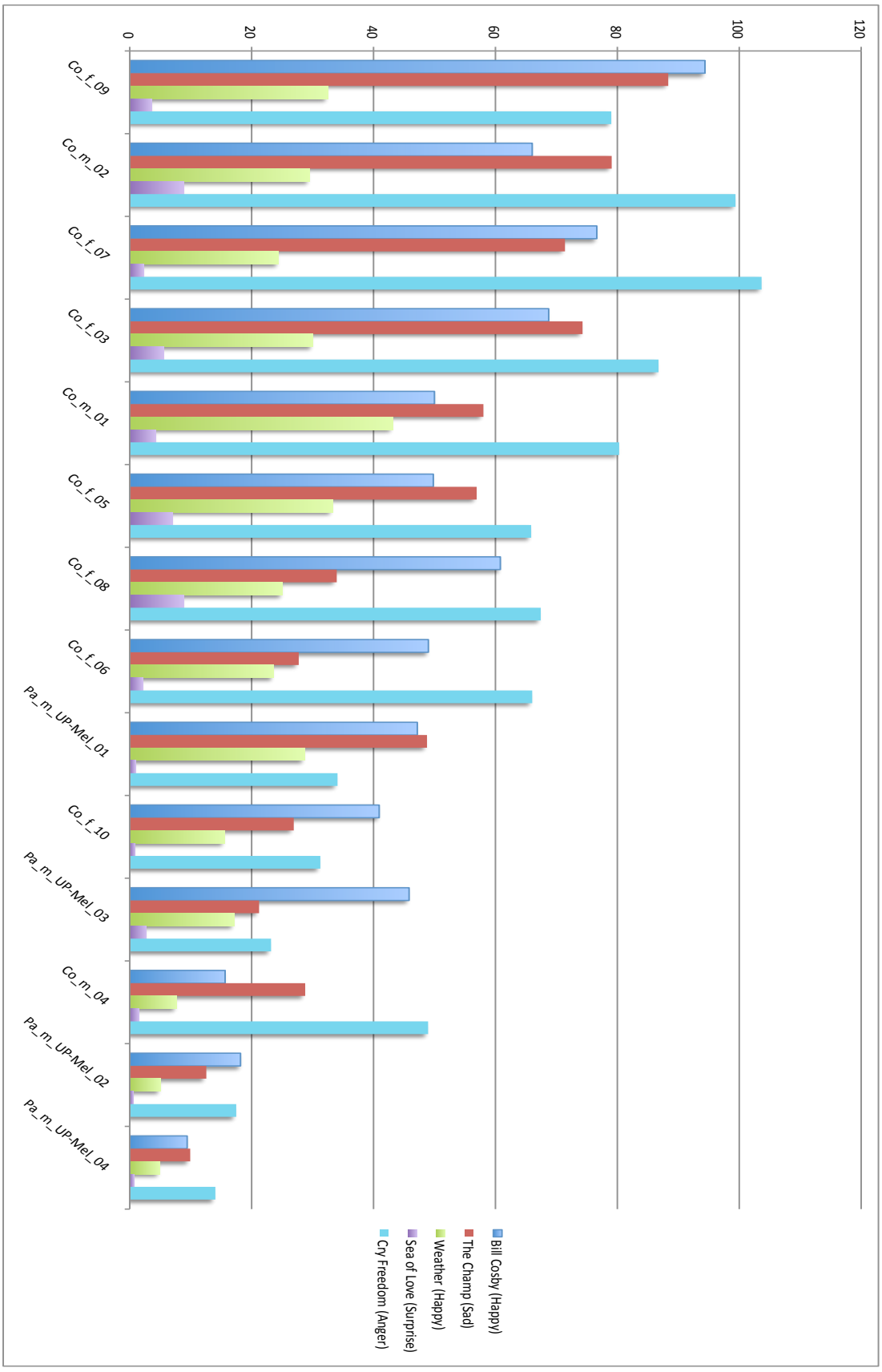


Figure 6.7: Old Paradigm - Clustered column chart comparing facial activity (Co - Control, Pa - Patient)

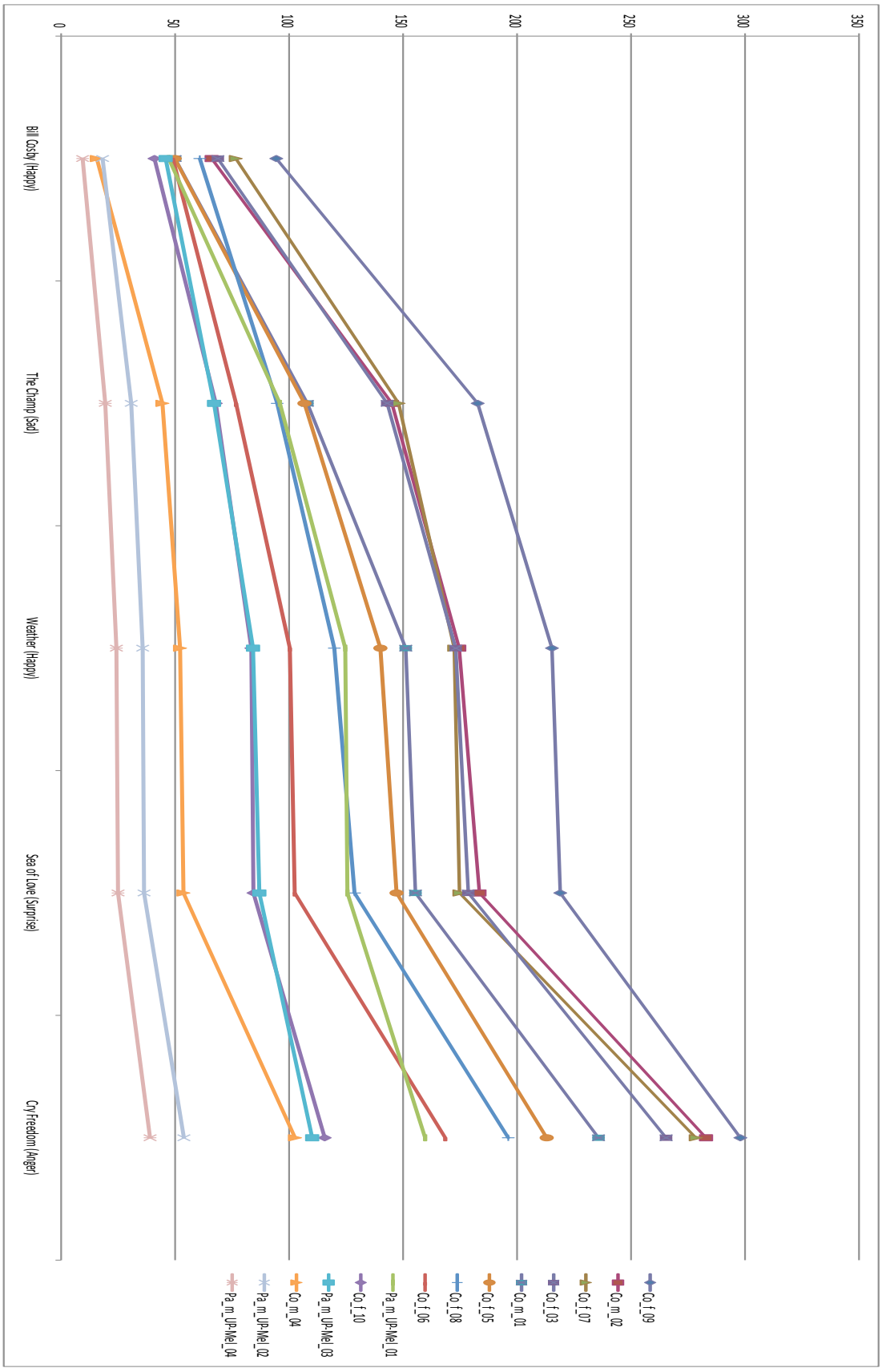
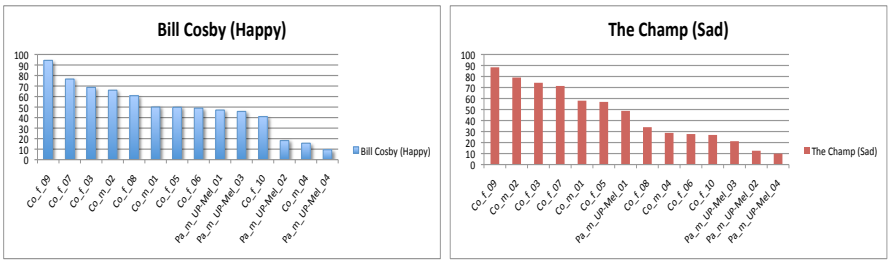
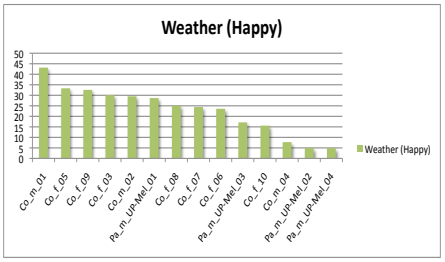


Figure 6.8: Old Paradigm - Line chart comparing accumulated facial activity (Co - Control, Pa - Patient)

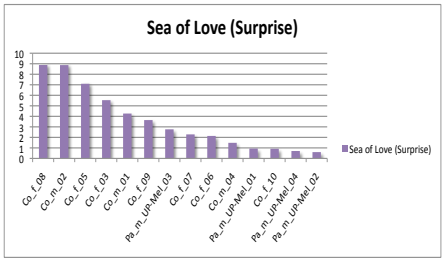


(a) Bill Cosby

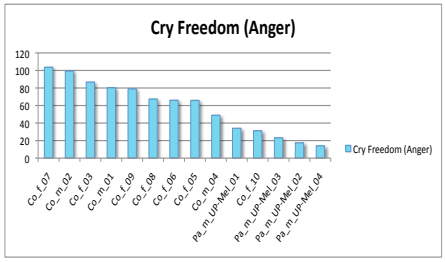
(b) The Champ



(c) Weather



(d) Sea of Love



(e) Cry Freedom

Figure 6.9: Old Paradigm - Facial activity for each video

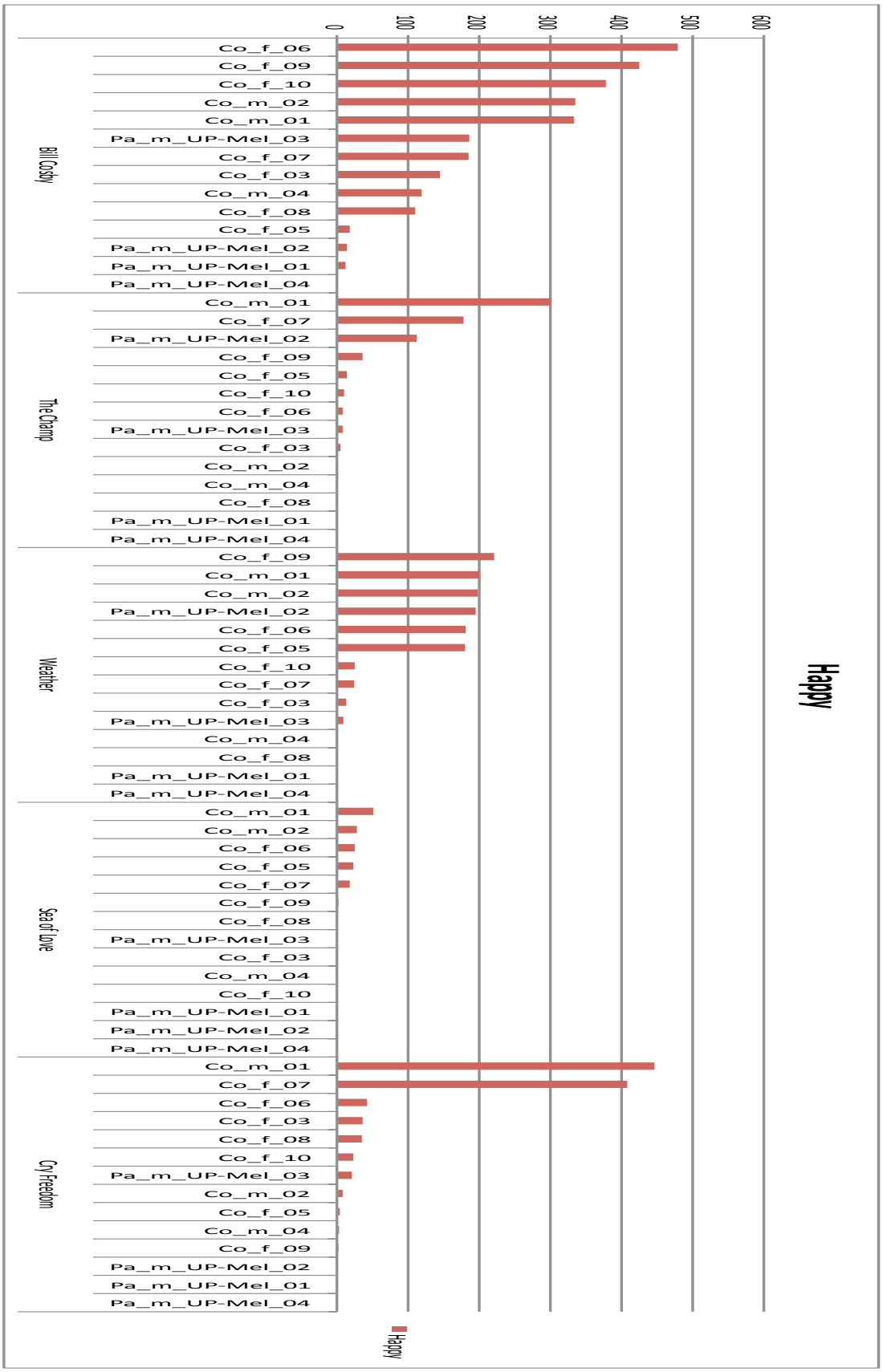


Figure 6.10: Old Paradigm - Number of happy expressions

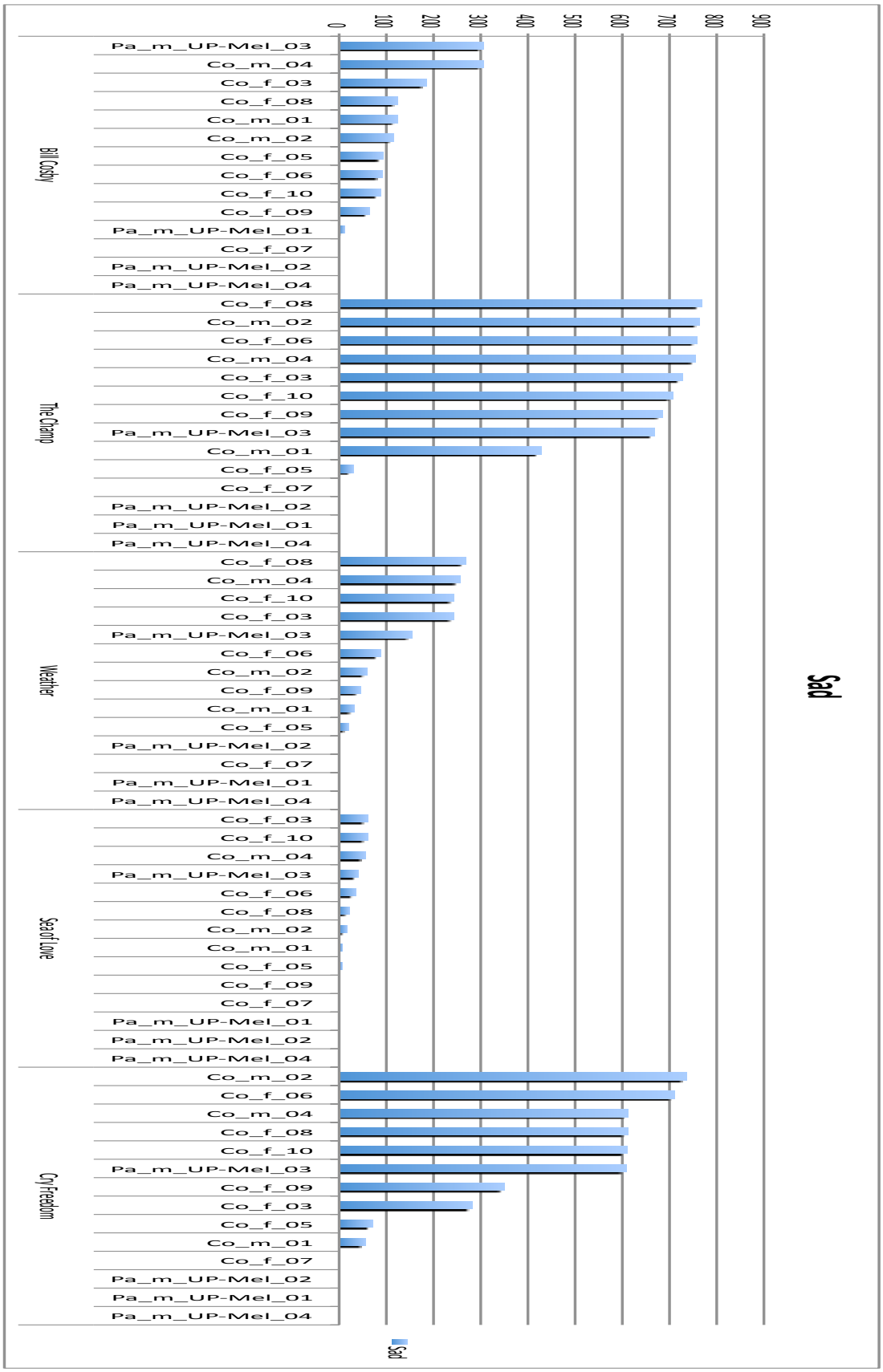


Figure 6.11: Old Paradigm - Number of sad expressions

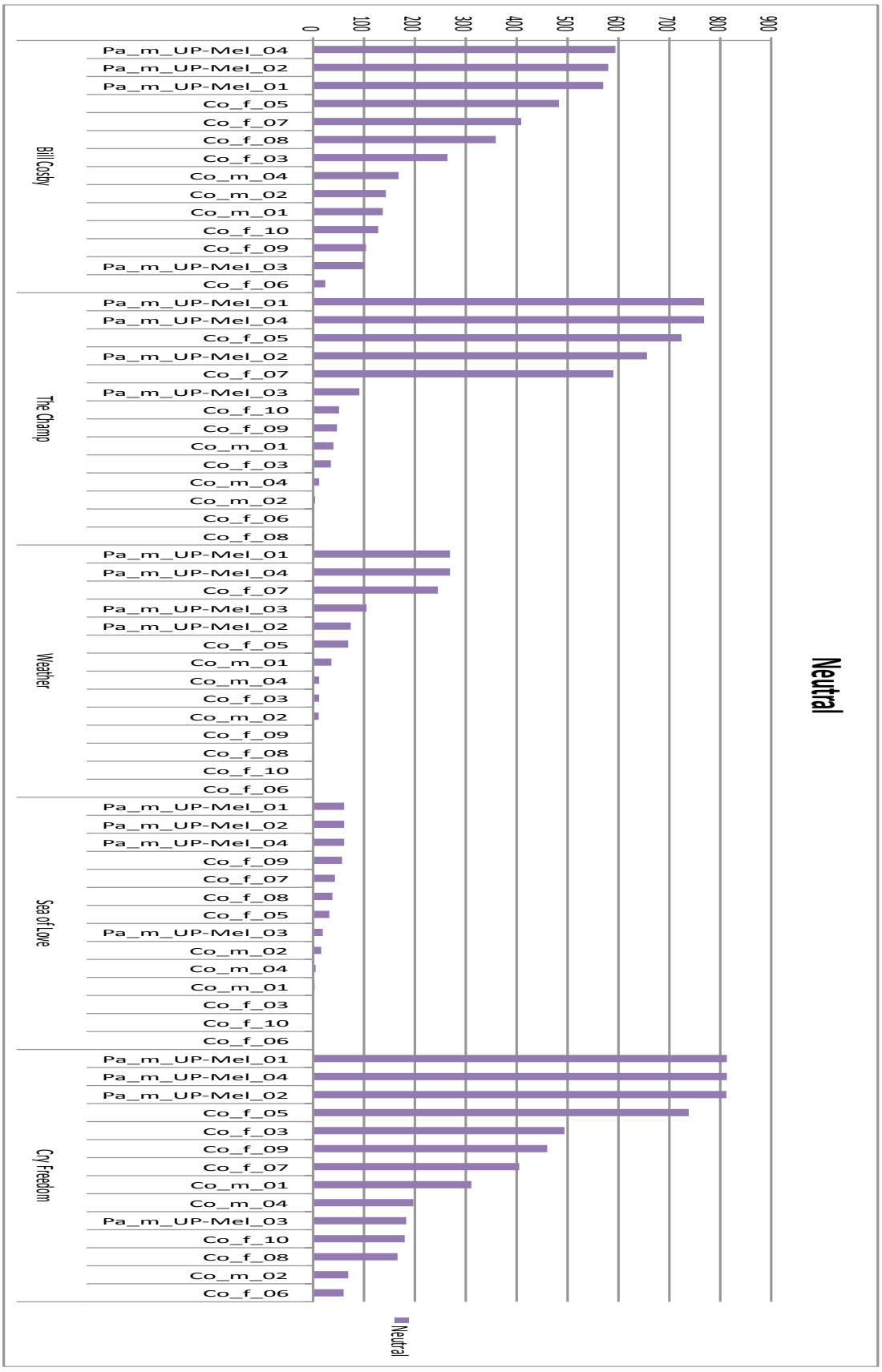


Figure 6.12: Old Paradigm - Number of neutral expressions

6.4.3 New Paradigm

Figure 6.13 displays a stacked column chart of facial activity. Although patient Pa_f_UP_BP2_08 has a very high score, examination of the video revealed that she displayed non-purposeful or habitual mouth movement throughout the recording. Two patients, Pa_f_UP-NonMel_09 and Pa_f_UP-NonMel_06, with a clinical diagnosis of unipolar **non**-melancholic depression, also had a high facial activity score. Patients Pa_f_UP-Mel_10 and Pa_m_UP-Mel_05, both diagnosed with unipolar melancholic depression score lowest in the facial activity scale. Patient Pa_m_PD_11, who had a Mini-diagnosis of unipolar melancholic depression and a clinical diagnosis of panic disorder, had a low facial activity score. Figure 6.14 displays a clustered column chart of the same data as Figure 6.6 facial activity. Each of the sub-figures in Figure 6.15 shows the facial activity specific to a movie clip.

Figure 6.16 shows the number of happy expressions displayed by each subject, for each film clip over time over the entire series of clips. Figure 6.17 shows the number of sad expressions displayed by each subject, for each film clip over time over the entire series of clips. Figure 6.18 shows the number of neutral expressions displayed by each subject, for each film clip over time over the entire series of clips.

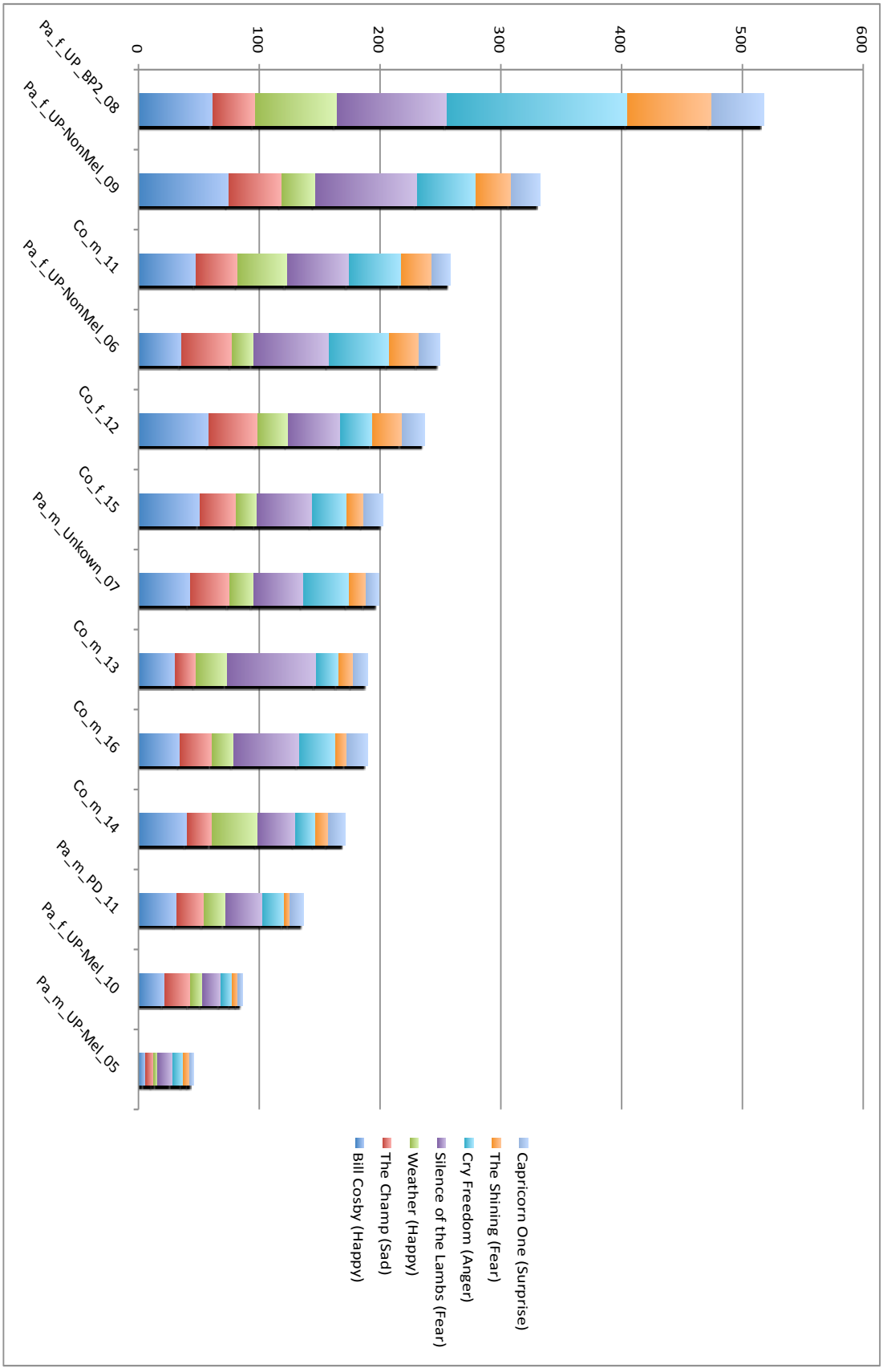


Figure 6.13: New Paradigm - Stacked column chart comparing facial Activity (Co - Control, Pa - Patient)

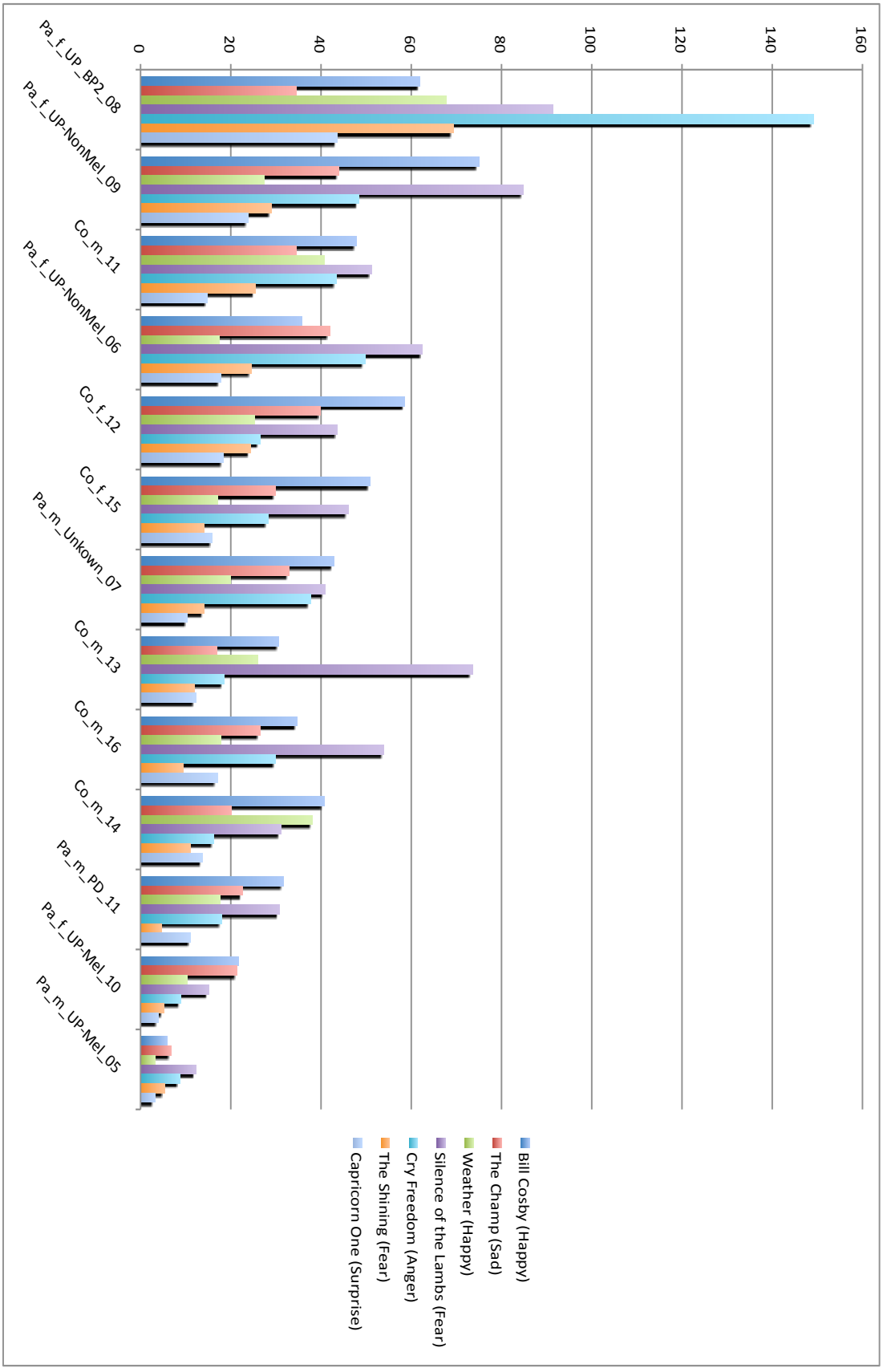
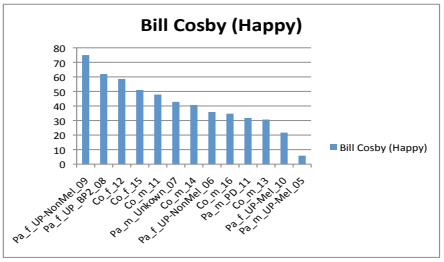
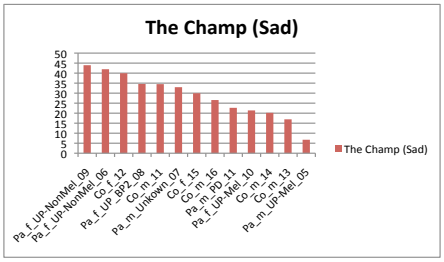


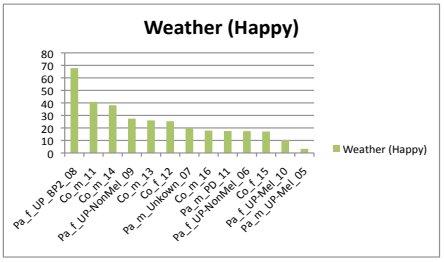
Figure 6.14: New Paradigm - - Clustered column chart comparing facial activity (Co - Control, Pa - Patient)



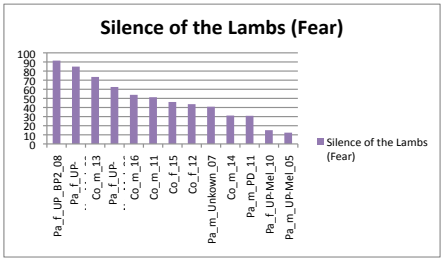
(a) Bill Cosby



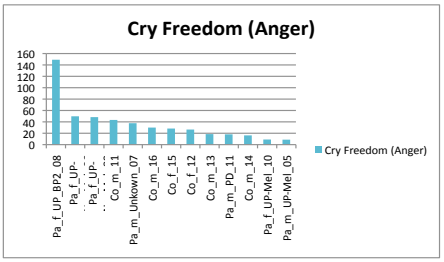
(b) The Champ



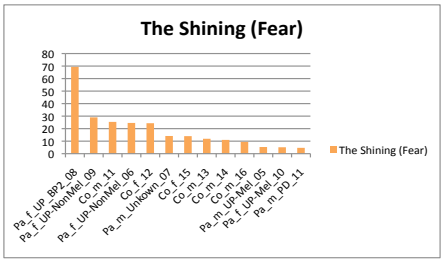
(c) Weather



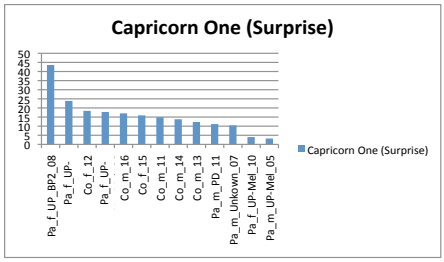
(d) Silence of the Lambs



(e) Cry Freedom



(f) The Shining



(g) Capricorn One

Figure 6.15: New Paradigm - Facial Activity for each video

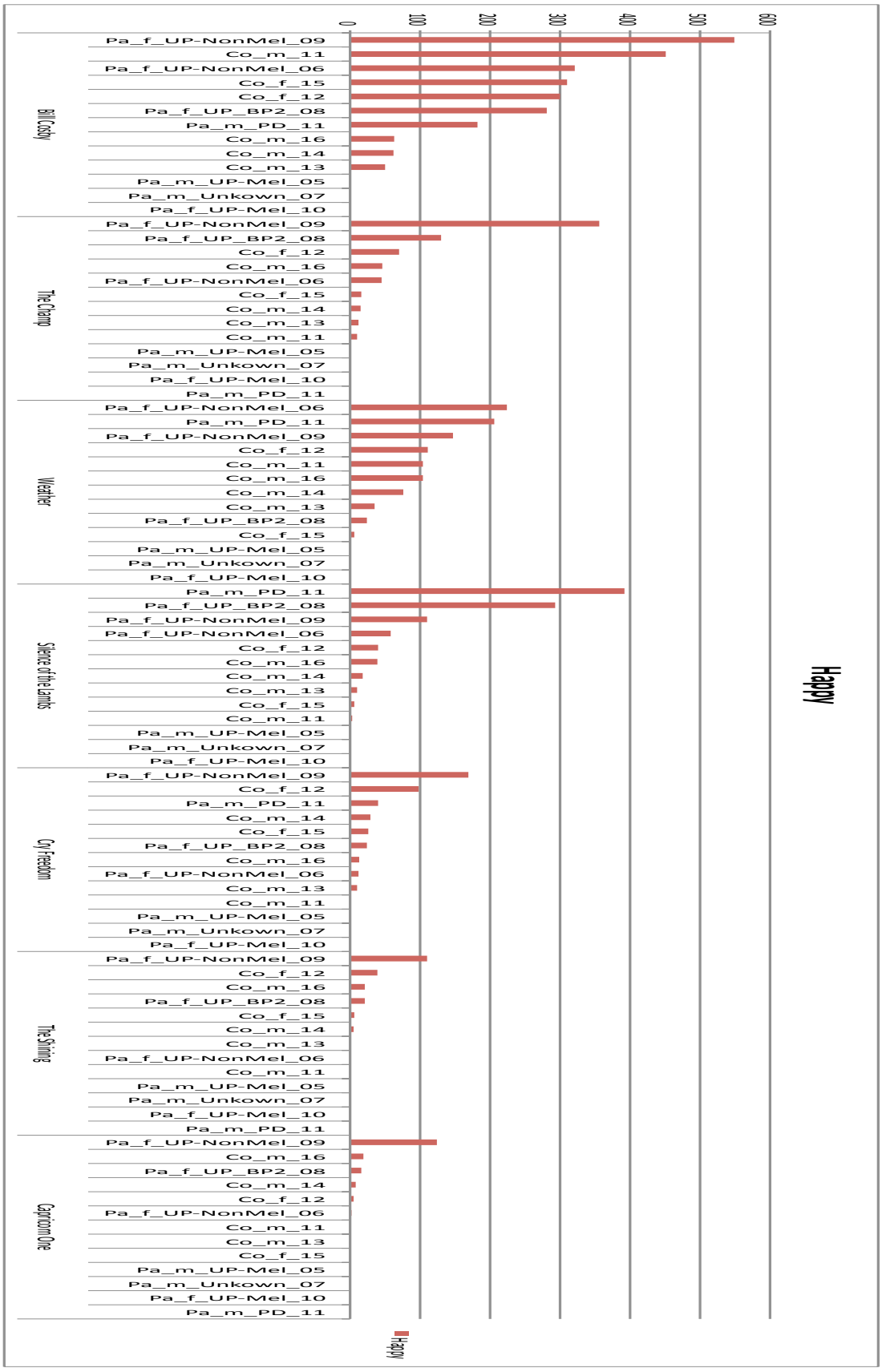


Figure 6.16: New Paradigm - Number of happy expressions

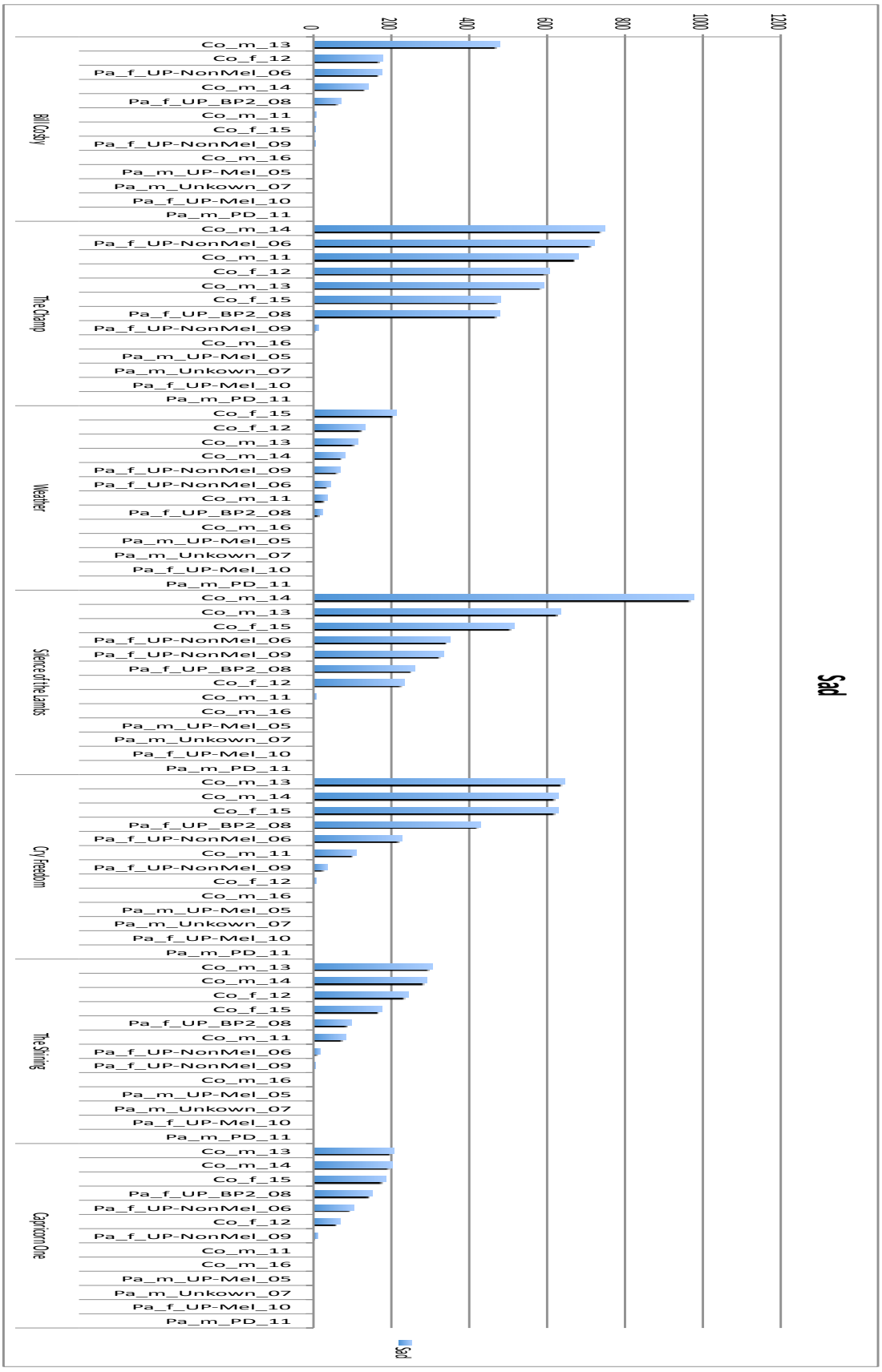


Figure 6.17: New Paradigm - Number of sad expressions

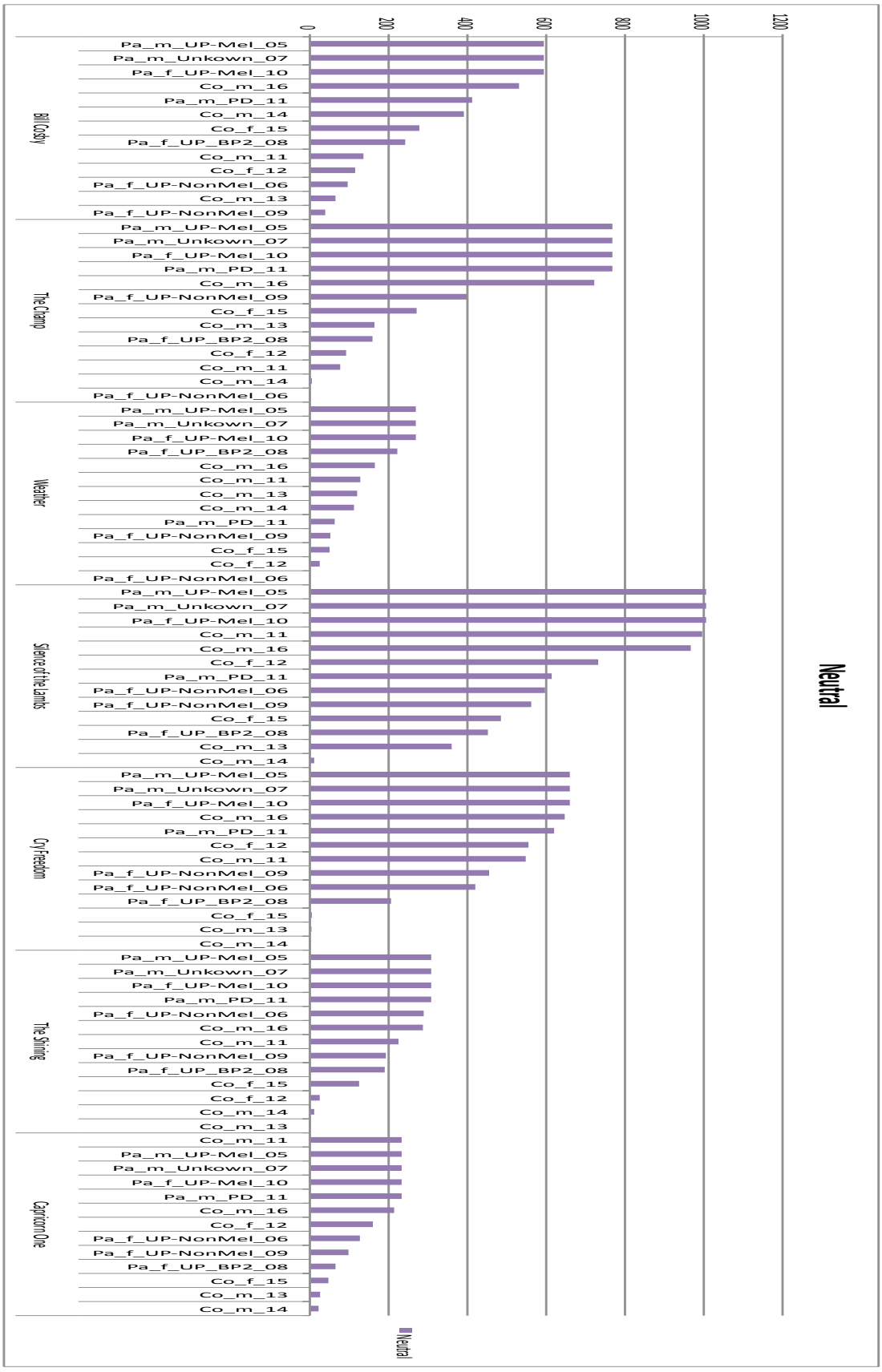


Figure 6.18: New Paradigm - Number of neutral expressions

6.5 Evaluation and Conclusions

There was some support for both hypotheses.

6.5.1 Hypothesis 1

When viewing the stimuli, patients with a clinical diagnosis of unipolar melancholic depression will show less facial activity than control subjects and patients with other types of depression.

In both the Old Paradigm and New the Paradigm results, there was a tendency for patients with unipolar melancholic depression to have reduced facial activity.

6.5.2 Hypothesis 2

When viewing the stimuli, patients with a clinical diagnosis of unipolar melancholic depression will show less repertoire *of facial expressions than control subjects and patients with other types of depression.*

In both the Old Paradigm and New Paradigm results, there was a tendency for patients with unipolar melancholic depression to show less positive facial expressions. However, analysis of the results reveals that the same set of patients also display less negative (sad) expressions. This is in keeping with the lower facial scores and the tendency towards a high number of neutral expressions.

6.6 Overall Evaluation

Although the results would tend to confirm the hypotheses, realistically, the sample size is far too small. One possibility raised, after the analysis had been undertaken, is that some of the patients show psychomotor agitation, whereas others show retardation - this would suggest 2 clusters of patients. Many more recordings are needed

before any conclusions could be drawn. Other factors such as age, gender and cultural background need to be considered, i.e. to test if there are other attributes that closely correlate with facial activity and expressiveness.

The motivation for the exercise was to test the feasibility of the approach and, on that basis, the results confirm that this method of facial activity and expression analysis could be used successfully in this and similar studies. Anecdotally, even with the small sample size, there were some interesting patterns. For instance, other types of MDD participants seemed to have slightly higher levels of facial activity and expressions than control subjects and patients with unipolar melancholic depression. Another interesting event was that, in one case ethnic background seemed to influence the facial activity and expression responses to the emotion in the video clip. The control subject's responses during the sad clip was similar to other control subjects, whereas his responses during other clips was much lower. Obviously, much more samples would be required to support the notion, but it would make an interesting follow up study.

The map is not the territory.

Alfred Korzybski

7

Semantics and Metadata

7.1 Introduction

The chapter reflects on some of the lessons learned in the earlier experimental tasks, and in keeping with objective 4 of the dissertation

> *"Identify avenues for improvement in the emotional expression recognition process."*

considers the limitations in emotional expression recognition, and ways in which they could be overcome. The problem domain is expanded well beyond the experiments

described in this thesis, to the field of *affective computing*. *Affective computing* is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena [Picard 97].

This chapter is organised as follows:

Section 7.2 discusses some of the strengths and weaknesses of the FER approach, used in the earlier experiments. Approaches for improvement are suggested and, once the background has been explained, an example framework for affective computing is described in Section 7.3. This is further explained by way of examples in Sections 7.4 and 7.5.

7.2 Discussion

With the empirical footing in place, the requirements of an expression recognition system can be revisited, and several observations made regarding the earlier experiments:

1. Culture-specific rules influence the display of affect [Cowie 05a], and in the experiment in Chapter 6, the control subject Co_m_04's low facial activity score gave cause for speculation that there might be ethnic or cultural factors influencing the result. This raises the question of how far a purely *statistical* approach to emotional expression recognition can extend. One would think that accounting for ethnic background or culture, as well as other factors such as context and personality type are beyond the limitations of such an approach.

2. The results of the MDD experiment in Chapter 6 compared participants' expressions during specific movie clips (see Figure 6.10). The actual *stimuli* information was not recorded in NXS, and the expressions had to be matched to the temporal sequences manually and *a posteriori*. This was quite time consum-

ing, and although a technical solution could be found to synchronise the *start* of the video recording with the stimuli presentation, a solution that incorporated *detailed* temporal information about the *stimuli* would be more useful. For instance, knowing which frames the "punch-line" occurs in a movie clip would enable the latency between subjects' reactions and the stimuli to be easily measured and compared.

3. The experiment in Chapter 6 was confined to the first part of the paradigm, where participants view video clips, i.e. FER only. Subsequent steps in the experimental paradigm include viewing of IAPS [Lang 05] images and an open interview, i.e. a question & answer stage at the end. The processing of the audio-visual sections of the interview is much more difficult. During this stage, some method of detecting when each interlocutor speaks is necessary and some means of representing the dialogue is required. A participant's facial display of a particular emotion will obviously be different during speech to that when viewing a video clip. The AAM will need to be able to track the face during speech and, possibly, additional classifiers will be needed to be able to match expressions in speech.

4. With a larger sample size in the MDD experiment, it might be useful to incorporate other variables, e.g. the participant's age and gender, or, even to consider *soft variables* such as temperament [Parker 02] and personality type [Parker 06]. Of course, this *subject* information can be recorded on a spreadsheet or word processing document, as was the case in this experiment. However, as the amount of data increases, so too does the degree of difficulty in maintaining spreadsheets. Having the information stored within the system performing the analysis, NXS in this case, would potentially create much more comprehensive outputs for analysis.

5. In the course of both the experiments, although the aggregated or summarised data was exported to files of comma-separated values, the raw data was actually stored in a system-specific format. Details pertaining to the subjects in the images, the frame sequence numbers, the location of the facial landmark points and the expression classified for each frame, are all in a format, known only to NXS. Thus, without additional effort, the data is system specific and unlikely to be useable, in its raw form, by other systems or studies.

In recent research, there have been attempts to add rules and record descriptions to the emotional expression recognition process (audio and video), and each project has devised its own approach. Some have used rulebases and case-based reasoning type information [Pantic 04a, Valstar 06a], whereas others have attempted a more complex level of integration [Vinciarelli 08, Vinciarelli 09]. [Athanaselisa 05, Devillers 05, Liscombe 05] describe efforts to represent non-basic emotional patterns and context features in real-life emotions in audio-video data. [Athanaselisa 05] demonstrate that recognition of speech can be improved by incorporating a dictionary of affect with the standard ASR dictionary. [Cowie 05b] reports on a "fuzzy" rule based system for interpreting facial expressions.

Each of the studies mentioned so far has used its own technique to incorporate rules, and it implies 1) a need to devise some common method of defining rules; and, 2) a requirement to record rules and recognition results in a standard, reusable way. In the remainder of this section, two complementary and overlapping concepts are introduced. Subsection 7.2.1 introduces ontologies and Subsection 7.2.2 examines two description, or markup, schemes. One known as EARL is specifically for emotional content, and the other, MPEG-7, is broader and intended for audio-visual content (including a modest amount of affective content).

7.2.1 Use of Ontologies to Describe Content

An ontology is a statement of concepts which facilitates the specification of an agreed vocabulary within a domain of interest. Ontologies have been used for some time in the annotation of web pages. More recently, they have been used to semantically index commercial video productions [Benitez 03, Bertini 05, Grana 05, Hunter 02, Jaimes 03, Lagoze 01, Luo 06, Navigli 03, Obrenovic 05, Rahman 05, Tsinaraki 03, Tsinaraki 07, Song 05, Polydoros 06].

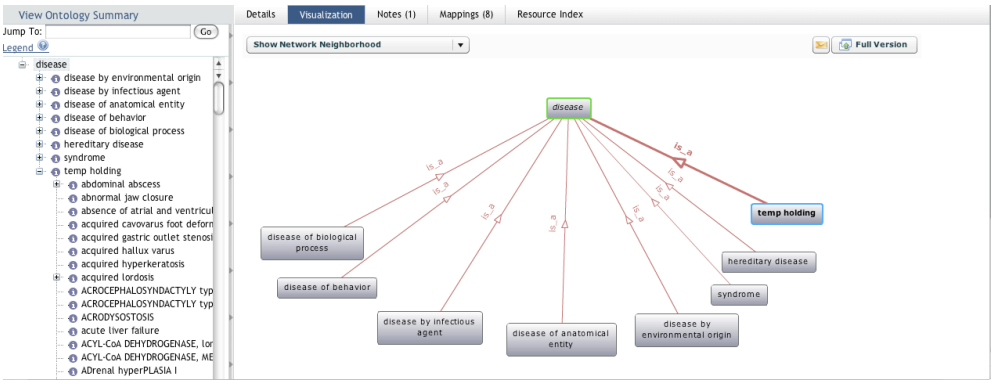


Figure 7.1: Human disease ontology

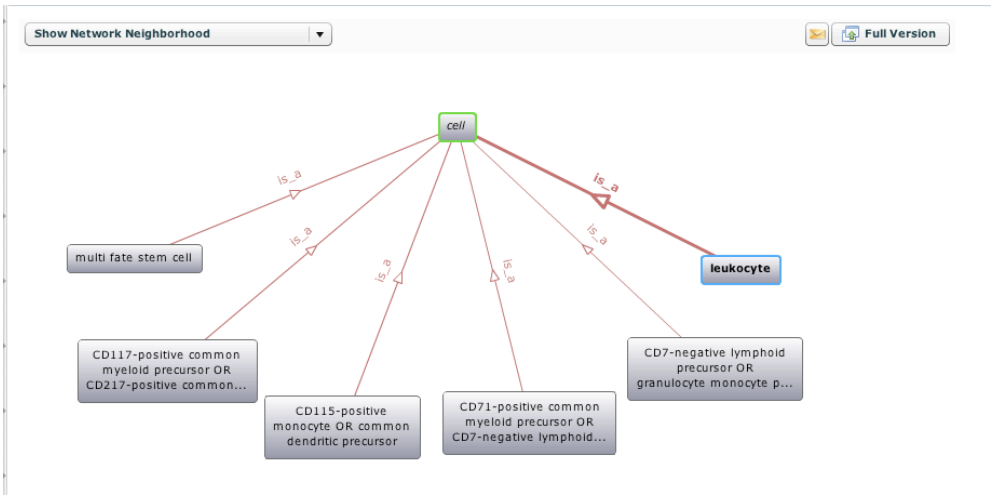


Figure 7.2: Cell ontology

In its crudest form, an ontology can be thought of as an hierarchical database of

concepts. Figures 7.1 and 7.2 illustrate 2 ontologies, one of human disease, and another of cells. Once the concepts have been populated with values, it becomes akin to a knowledge-base. One popular and well supported software product for building ontologies, *Protégé* [Protégé], is an ontology editor and knowledge-base framework, and is supported by Stanford University. One very powerful feature of an ontology, is its ability to link to other ontologies, e.g. medical and gene ontologies,[1].

7.2.2 Semantic Markup

Emotion Annotation and Representation Language (EARL)

One issue in emotion recognition, is reuse and verification of results, and, until recently, there has been no universally accepted system for describing emotional content. The Human-Machine Interaction Network on Emotion (HUMAINE) project has attempted to remedy this through the definition of the EARL [Baggia 08, Schröder 07]. EARL is an XML-based language for representing and annotating emotions in technological contexts. Using EARL, emotional expression can be described using either a set of forty-eight categories, dimensions or even appraisal theory. Examples of XML elements for annotation include "Emotion descriptor", which could be a category or a dimension,; "Intensity", expressed in terms of numeric values or discrete labels; and, "Start" and "End".

MPEG-7

Another initiative is that of the Moving Picture Experts Group (MPEG) which has developed the MPEG-7 standard for audio, audio-video and multimedia description [MPEG-7]. MPEG-7 uses metadata structures or Multimedia Description Schemes (MDS) for describing and annotating audio-video content. These are provided as a

[1]Examples can be located at `http://bioportal.bioontology.org/ontologies`, last accessed 22 April 2010.

standardised way of describing the important concepts in content description and content management in order to facilitate searching, indexing, filtering, and access. They are defined using the MPEG-7 Description Definition Language (DDL), which is XML Schema-based. The output is a description expressed in XML, which can be used for editing, searching, filtering. The standard also provides a description scheme for compressed binary form for storage or transmission [Chiariglione 01, Rege 05, Salembier 01]. Examples in the use of MPEG-7 exist in the video surveillance industry where streams of video are matched against descriptions of training data [Annesley 05]. The standard also caters for the description of affective content.

7.3 An Affective Communication Framework

To illustrate how the concepts previously described might be applied, an exemplary approach is presented which consists of 1) a generic model of affective communication; and 2) a set of *ontologies*. The model and ontologies, intended to be used in conjunction with one another, describe:

1. affective communication concepts;
2. affective computing research; and
3. affective computing resources.

Figure 7.3 presents an example of a base-level model of emotions in spoken language. It includes speaker and listener, in keeping with the Brunswikian lens model, as proposed by [Scherer 03]. Modelling attributes of both speaker and listener caters for the fact that the listener's cultural and social presentation vis-à-vis the speaker may also influence judgement of emotional content. It also includes a number of factors that influence the expression of affect in spoken language. Each of these factors is briefly

discussed, with more attention given to context, as this is seen as a much neglected factor in the study of automatic affective state recognition.

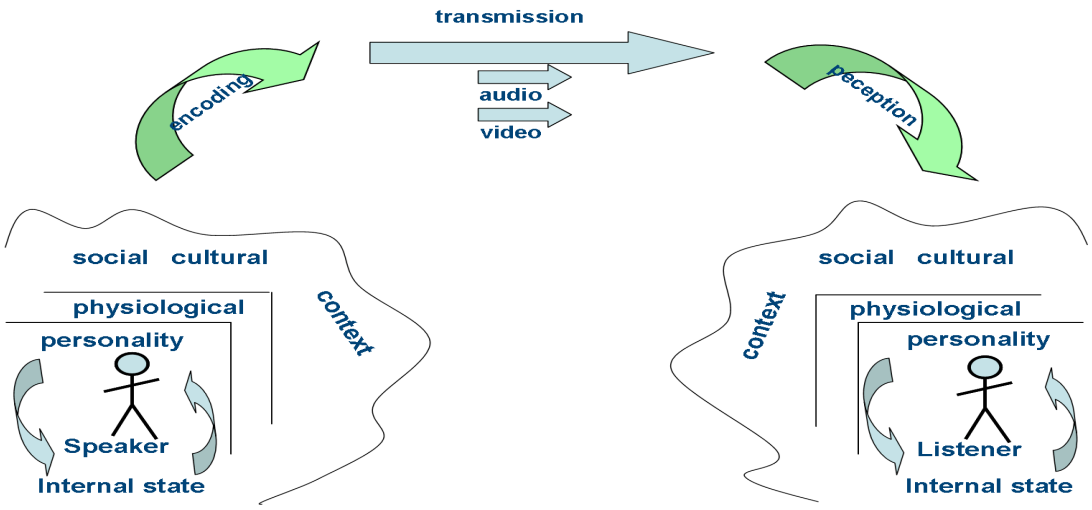


Figure 7.3: A generic model of affective communication

7.3.1 Factors in the Proposed Framework

Context

Context is linked to modality and emotion is strongly multi-modal in the way that certain emotions manifest themselves favouring one modality over the other [Cowie 05a]. Physiological measurements change depending on whether a subject is sedentary or mobile. A stressful context, such as an emergency hot-line, air-traffic control, or a war zone, is likely to yield more examples of affect than everyday conversation.

It is likely to produce quite different responses to a recording studio, used to record posed facial expressions for scientific experiments. [Stibbard 01] recommended

> *"...the expansion of the data collected to include relevant non-phonetic factors including contextual and inter-personal information."*

His findings underline the fact that most studies so far took place in an artificial environment, ignoring social, cultural, contextual and personality aspects which, in natural situations, are major factors modulating speech and affect presentation. The model depicted in Figure 7.3 takes into account the importance of context in the analysis of affect in speech.

There have been some attempts to include contextual data in emotion recognition research. [Devillers 05] includes context annotation as metadata in a corpus of medical emergency call centre dialogues. Context information was treated as either task-specific or global in nature. Unlike [Devillers 05], the model described in this dissertation does not differentiate between task-specific and global context as the difference is seen merely as temporal, i.e. pre-determined or established at "run-time".

The HUMAINE project [HUMAINE 06] has proposed that at least the following issues be specified:

- Agent characteristics (age, gender, race);
- Recording context (intrusiveness, formality, etc.);
- Intended audience (kin, colleagues, public);
- Overall communicative goal (to claim, to sway, to share a feeling, etc.);
- Social setting (none, passive other, interactant, group);
- Spatial focus (physical focus, imagined focus, none);
- Physical constraint (unrestricted, posture constrained, hands constrained); and
- Social constraint (pressure to expressiveness, neutral, pressure to formality).

"It is proposed to refine this scheme through work with the HUMAINE databases as they develop."

[Millar 04] developed a methodology for the design of audio-visual data corpora of the speaking face in which the need to make corpora re-usable is discussed. The methodology, aimed at corpus design, takes into account the need for *speaker* and *speaking environment* factors.

The model presented in this dissertation treats agent characteristics and social constraints separate to context information. This is because their effects on discourse are seen as separate topics for research.

Agent characteristics

As [Scherer 03] points out, most studies are either speaker oriented or listener oriented, with most being the former. This is significant when you consider that the emotional state of someone labelling affective content in a corpus could impact the label that is ascribed to a speaker's message, or facial expression.

The literature has not given much attention to the role that agent characteristics, such as personality type, play in affective presentation. This is surprising when one considers the obvious difference in expression between extroverted and introverted types. Intuitively, one would expect a marked difference in signals between these types of speakers. One would also think that knowing a person's personality type would be of great benefit in applications monitoring an individual's emotions [Parker 06].

At a more physical level, agent characteristics such as facial hair, whether they wear spectacles, and their head and eye movements all affect the ability to visually detect and interpret emotions.

Cultural

Culture-specific rules influence the display of affect [Cowie 05a], and gender and age are established as important factors in shaping conversation style and content in many societies. Studies by [Koike 98] and [Shigeno 98] have shown that it is difficult to

identify the emotion of a speaker from a different culture and that people will predominantly use visual information to identify emotion. Putting it in the perspective of the proposed model, cognisance of the speaker and listener's cultural backgrounds, the context, and whether visual cues are available, obviously influence the effectiveness of affect recognition.

Physiological

It might be stating the obvious but there are marked differences in speech signals and facial expressions between people of different age, gender and health. The habitual settings of facial features and vocal organs determine the speaker's range of possible visual appearances and sounds produced. The configuration of facial features, such as chin, lips, nose, and eyes, provide the visual cues, whereas the vocal tract length and internal muscle tone guide the interpretation of acoustic output [Millar 04].

Social

Social factors temper spoken language to the demands of civil discourse [Cowie 05a]. For example, affective bursts are likely to be constrained in the case of a minor relating to an adult, yet totally unconstrained in a scenario of sibling rivalry. Similarly, a social setting in a library is less likely to yield loud and extroverted displays of affect than a family setting.

Internal state

Internal state has been included in the model for completeness. At the core of affective states is the person and their experiences. Recent events such as winning the lottery or losing a job are likely to influence emotions and their display.

7.3.2 Influences in the Display of Affect

Modulating factors			Production and detection factors		
Cultural	**Social**	**Context**	**Agent Characteristics**	**Physiological**	**Internal State**
Speaker's vis-à-vis listener's age and gender Language Customs Race	Education Familiarity/ rapport with listener Gender	Group situations Ambient conditions Dialogue turn Familiarity with system Sedentary/active Overt/covert Location	Extrovert/ Introvert Authoritarian/ control freak Child vs elderly Appearance, eg spectacles, facial hair, head and eye movement	Voice quality Child vs elderly Gender Illness/ Infirmity Impairment Vocal tract length Skin colour	Recent events, eg lottery wins, losses

Figure 7.4: Use of the model in practice

To help explain the differences between the factors that influence the expression of affect, Figure 7.4 lists some examples. The factors are divided into two groups. On the left, is a list of factors that modulate or influence the speaker's display of affect, i.e. cultural, social and contextual. On the right, are the factors that influence production or detection in the speaker or listener, respectively, i.e. personality type, physiological make-up and internal state.

7.4 A Set of Ontologies

The three ontologies described in this section are a means by which the model described in the previous section could be implemented. Figure 7.5 depicts the relationships between the ontologies and gives examples of each. Formality and rigour increase towards the apex of the diagram.

It needs to be emphasised that this proposal is not confined just to experiments such as those described within this dissertation. It is much broader and the intended users of the set of ontologies extends beyond research exercises. There could be many types of users such as librarians, decision support systems, application developers and teachers.

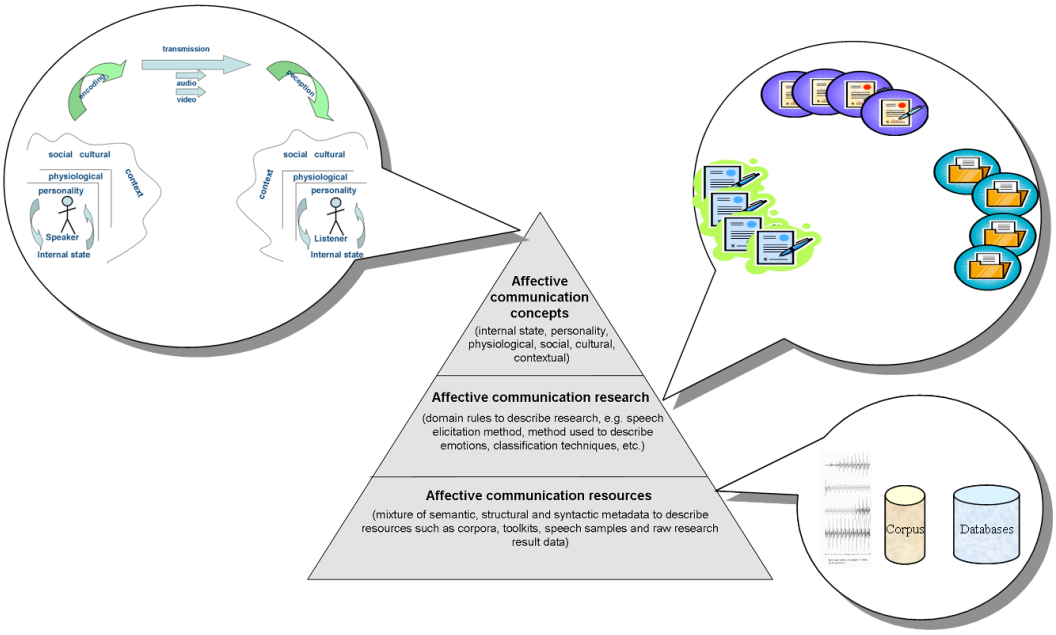


Figure 7.5: A set of ontologies for affective computing

7.4.1 Ontology 1 - Affective Communication Concepts

The top level ontology correlates to the model discussed in Section 7.3 and is a formal description of the domain of affective communication. It contains internal state, personality, physiological, social, cultural, and contextual factors. It can be linked to external ontologies in fields such as medicine, anatomy, and biology. A fragment of the top-level, domain ontology of concepts in shown in Figure 7.6.

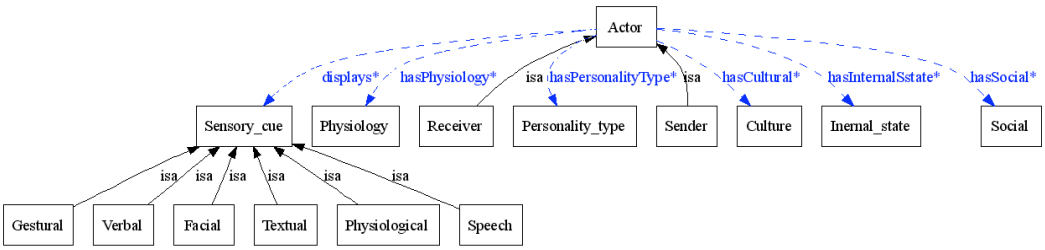


Figure 7.6: A fragment of the domain ontology of concepts

7.4.2 Ontology 2 - Affective Communication Research

This ontology is more loosely defined and includes the concepts and semantics used to define research in the field. It has been left generic and can be further subdivided into an affective computing domain at a later stage, if needed. It is used to specify the rules by which accredited research reports are catalogued. It includes metadata to describe, for example,

- classification techniques used;
- the method of eliciting speech, e.g. acted or natural; and
- manner in which corpora or results have been annotated, e.g. categorical or dimensional.

Creating an ontology this way introduces a common way of reporting the knowledge and facilitates intelligent searching and reuse of knowledge within the domain. For instance, an ontology just based on the models described here could be used to find all research reports where:

SPEAKER(internalState='happy',
physiology='any',
agentCharacteristics='extrovert',
social='friendly',context='public',
elicitation='dimension')

7.4.3 Ontology 3 - Affective Communication Resources

This ontology is more correctly a repository containing both formal and informal rules, as well as data. It is a combination of semantic, structural and syntactic metadata. It contains information about resources such as corpora, toolkits, audio and video samples, and raw research result data.

The next section explains the bottom level, application ontology, in more detail.

7.5 An Exemplary Application Ontology for Affective Sensing

Figure 7.7 illustrates an example application ontology for affective sensing, in a context of investigating dialogues. In the context of the experiments in Chapter 6, a dialogue is the interaction between the participant and the stimuli. During a dialogue, various events can occur, triggered by one of the dialogue participants and recorded by the sensor system. These are recorded as time stamped instances of events, so that they can be easily identified.

In the ontology, the roles for each interlocutor, sender and receiver, are distinguished. This caters for the type of open interview session in the experiments in Chapter 6. At various points in time, each interlocutor can take on different roles. On the sensory side, facial, gestural, textual, speech, physiological and verbal[2] cues are distinguished. The ontology could be extended for other cues and is meant to serve as an example here, rather than a complete list of affective cues. Finally, the emotion classification method used in the investigation of a particular dialogue is also recorded.

[2]The difference between speech and verbal cues here being spoken language versus other verbal utterings.

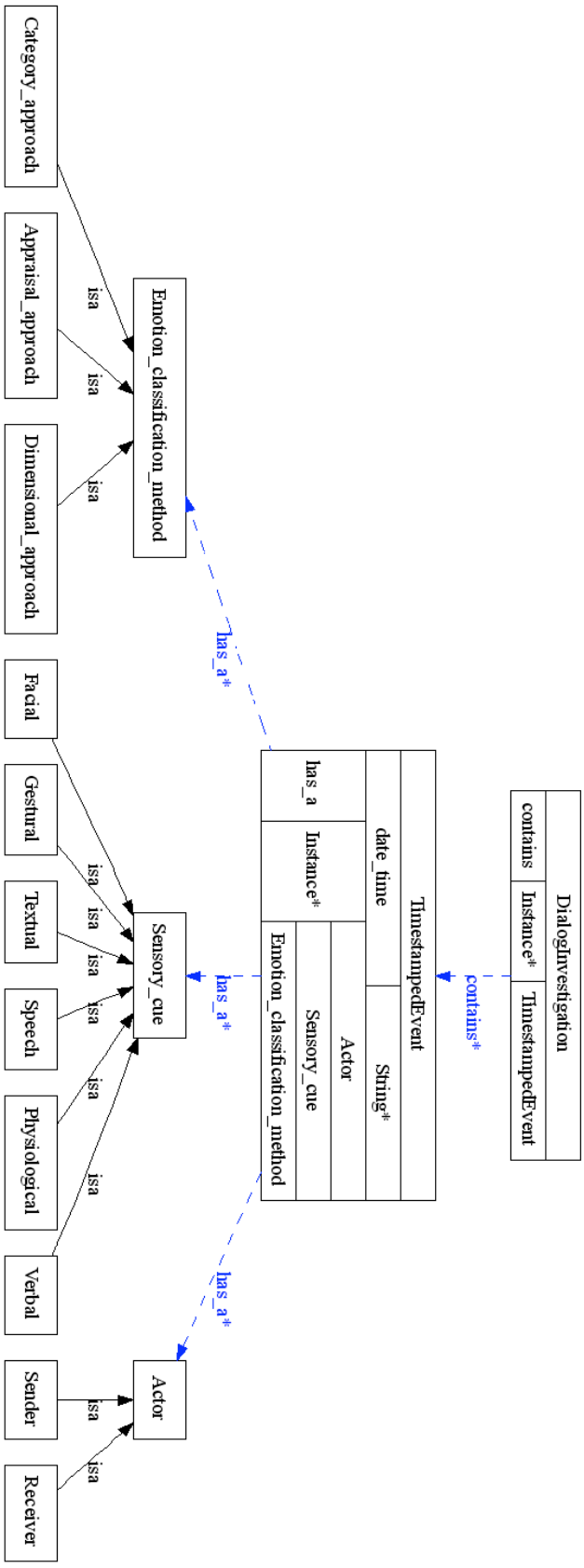


Figure 7.7: An application ontology for affective sensing

8

Conclusions

8.1 Introduction

In this dissertation, state-of-the-art computer vision techniques that apply to Facial Expression Recognition (FER) have been reviewed. After outlining the requirements for building an emotional expression recognition system, the design of such a system, the Any Expression Recognition System (NXS), was presented. Two experiments were devised, with the objectives of 1) proving the concepts of the NXS system and their implementation; and 2) in a much broader sense, establishing if Facial Expression Recognition (FER) techniques could be a applied to more subtle emotional expressions,

such as anxiety and depression. Following the experiments, the practical limitations of the statistical approach to FER were discussed, along with ways in which these could be overcome through the use of a model and ontologies for affective computing.

This chapter concludes the dissertation and is organised as follows:

In Section 8.2, each of the objectives as stated in Chapter 1 are examined, and consideration is given to the results from the experimental work in Chapters 5 and 6. Finally, in Section 8.3, the conclusions and contributions of this dissertation are stated before discussing open issues and future directions.

8.2 Objectives

8.2.1 Objective 1

Explore, though the construction of a prototype system that incorporates AAMs, what would be required in order to build a fully-functional FER system, i.e. one where the system could be trained and then used to recognise new and previously unseen video or images.

Chapter 4 discussed the functional requirements of a system capable of sensing multiple variable inputs from voice, facial expression and movement, making some assessment of the signals of each, and then fusing them to provide some degree of affect recognition. The chapter went on to describe how the requirements had been implemented in a prototype system, *NXS*, which was built to support the experimental aspects of this dissertation. Although the system has been built to cater for *multi-modal* analysis and recognition, its application was confined to FER within the experiments.

The system proved to be flexible and robust throughout the experiments, and dealt with the requirements to process sets of images, as in Chapter 5, as well as the more difficult task of video processing. The underlying software components that are required

to perform FER, were shown to be very stable, i.e. [OpenCV] for face detect and image capture, [VXL] for image processing, DemoLib [Saragih 08] for AAM development, LIBSVM [Chang 01] for SVM classification. Components were easily interchanged, and the MultiBoost classification software [MultiBoost 06] was easily replaced with the LIBSVM implementation of SVM classification.

The system was built using the Qt software from [Qt 09], which will ensure that it can be deployed in Windows, Mac OS or a Unix/Linux variant environments. Anecdotally, NXS's performance was quite adequate for real-time FER, despite little effort being expended in tuning the performance of the software. Overall, it demonstrates that underlying software components are mature enough to be incorporated in a *production-like* system.

8.2.2 Objective 2

Investigate whether FER practices could be applied to non-primary emotional expression such as anxiety. A great deal of experience was gained from the experiment, and, despite the lack of samples and natural data, the results suggest that the recognition of anxious expressions is possible, but becomes more difficult when fearful expressions are also present. The difficulty increases when more primary expressions are added to the classification problem. The exercise demonstrates that facial expression classification is, in general, a difficult task and in some situations, in the absence of contextual information and/or temporal data revealing facial dynamics, may not even be possible. Moreover, even with contextual and temporal evidence present, the fact that a prototypical expression can take many forms, e.g. a 'happy' expression can be portrayed with or without opening the mouth, compounds the degree of difficulty. Any attempt at recognition may not be reliable without the presence of semantic information.

The second part of Experiment 4 demonstrated that, even using two popular and

creditable databases, a classifier built from images from one database, did not achieve a high CA in predicting expressions from the other, despite Gabor filtering being reasonably invariant to lighting conditions. This echoes the preliminary results reported in [Whitehill 09] (albeit, much worse), and, to address this problem, there seems to be two approaches. The first, is to expand the training set of images, including samples with variant conditions, e.g. recorded under different lighting conditions. However, the results in Chapter 5 suggest that this will not be a complete solution. The second, is to incorporate some form of logic or rule processing in expression recognition, and this was discussed in Chapter 7.

8.2.3 Objective 3

Examine whether FER practices could be applied to non-primary emotional expressions, such as those displayed by someone suffering from a MDD. The experiment tested two hypotheses relating to facial activity and expressions and unipolar melancholic depression. Although there were clearly not enough samples to undertake a significant analysis-of-variation or draw conclusions, the results were encouraging.

Anecdotally, even with the small sample size, there were some interesting patterns. For instance, other types of MDD participants seemed to have slightly higher levels of facial activity and expressions than control subjects and patients with unipolar melancholic depression. Another interesting event was that, in one case ethnic background seemed to influence the facial activity and expression responses to the emotion in the video clip. The control subject's responses during the sad clip was similar to other control subjects, whereas his responses during other clips was much lower. Obviously, much more samples would be required to support the notion, but it would make an interesting follow up study.

8.2.4 Objective 4

Identify avenues for improvement in the emotional expression recognition process.

Chapter 7 reflected on some of the lessons learned in the earlier experimental tasks, and the scope was well beyond the field of FER, and set to *affective computing*. Two recurring themes seem inevitable and emerging in the literature, i.e. the need to incorporate rules into the recognition process, and a means to represent the recognition in a standard format. To address these requirements, the use of a model and ontologies of affective computing was proposed.

8.3 Conclusions and Future Work

8.3.1 Summary of Contributions

This is a very broad thesis, which ranges in scope, from computer vision techniques to psychology, taking in knowledge management concepts like ontologies along the way. Such cross-disciplinary dissertations are necessary to advance such a broad and, at times, nebulous, field such as affective computing. This dissertation contributes in several ways.

An artifact of the work is the NXS, which will be evolved and made available to other researchers. It has been used successfully in a collaborative study that investigates the links between facial expressions and MDD, undertaken at the Black Dog Institute, Sydney. The system is easily extendable, and its use is not confined to FER, instead being suitable for experimental work in full multi-modal emotional expression recognition. Its use is not confined to the recognition of primary emotional expression, and it could be used to sense for any other state, e.g. attraction, boredom, level of interest, or even all them. In fact, it is not confined to expression recognition and could be used for *face recognition*.

Even though the sample size was modest, the results in anxiety recognition are very encouraging. It suggests that anxious facial expressions *can* be recognised with FER techniques which could have many applications, extending beyond scientific and medical research, such as interactive games and passenger screening technology.

Similarly, the results of the experiments on depression are encouraging, and once a large enough sample size has been attained, the hypotheses in Chapter 6 can be properly tested. Even at this stage, interesting patterns in the data have emerged, enough to suggest that other hypotheses could be tested within MDD populations, and with other objectives in mind, e.g. cross-cultural responses to affective content.

8.3.2 Future Work

Recognition Based on Temporal Features

The FER within this dissertation was confined to recognition within images and no attempt was made to classify expressions based on temporal features. The raw facial landmark coordinates, already stored within NXS, could be used to train Hidden Markov Models HMMs. The major difficulty with this type of approach is to recognise the apex of the expressions, however, some ensemble of classifiers, perhaps using SVM to detect the peak of the expression and HMM to recongise the temporal expression could be attempted.

A

Presentation of Analysis and Data - Anxiety

A.1 Introduction

Q1.[*]	Fear	1	7.14%
	Anxiety	8	57.14%
	Uncertain	5	35.71%
Q2.[*]	Fear	5	35.71%
	Anxiety	7	50.00%
	Uncertain	2	14.29%
Q3.	Fear	4	28.57%
	Anxiety	9	64.29%
	Uncertain	1	7.14%
Q4.	Fear	6	42.86%
	Anxiety	7	50.00%
	Uncertain	1	7.14%
Q5.	Fear	4	28.57%
	Anxiety	9	64.29%
	Uncertain	1	7.14%
Q6.	Fear	8	57.14%
	Anxiety	4	28.57%
	Uncertain	2	14.29%
Q7.	Fear	11	78.57%
	Anxiety	1	7.14%
	Uncertain	2	14.29%
Q8.	Fear	5	35.71%
	Anxiety	8	57.14%
	Uncertain	1	7.14%
Q9.	Fear	12	85.71%
	Anxiety	0	0.00%
	Uncertain	2	14.29%
Q10.	Fear	1	7.14%
	Anxiety	9	64.29%
	Uncertain	4	28.57%
Q11.	Fear	10	71.43%
	Anxiety	4	28.57%
	Uncertain	0	0.00%
Q12.	Fear	5	35.71%
	Anxiety	4	28.57%
	Uncertain	5	35.71%
Q13.	Fear	8	57.14%
	Anxiety	6	42.86%
	Uncertain	0	0.00%
Q14.	Fear	1	7.14%
	Anxiety	8	57.14%
	Uncertain	5	35.71%
Q15.	Fear	5	35.71%
	Anxiety	4	28.57%
	Uncertain	5	35.71%
Q16.	Fear	11	78.57%
	Anxiety	2	14.29%
	Uncertain	1	7.14%
Q17.	Fear	11	78.57%
	Anxiety	0	0.00%
	Uncertain	3	21.43%
Q18.	Fear	5	35.71%
	Anxiety	4	28.57%

Figure A.1: Experiment 1 - Poll results

B

Presentation of Analysis and Data - Depression

B.1 Introduction

B.2 Old Paradigm

B.3 New Paradigm

	Bill Cosby (Happy)	The Champ (Sad)	Weather (Happy)	Sea of Love (Surprise)	Cry Freedom (Anger)	Total
Co_f_09	94.3759	88.349	32.5771	3.64379	79.0066	297.95239
Co_m_02	66.0356	79.069	29.5169	8.87003	99.3888	282.88033
Co_f_07	76.6297	71.4058	24.4274	2.28335	103.664	278.41025
Co_f_03	68.7369	74.2912	30.0492	5.53615	86.7574	265.37085
Co_m_01	49.9946	58.0265	43.1402	4.2638	80.2981	235.7232
Co_f_05	49.8066	56.9214	33.3114	7.09537	65.8639	212.99867
Co_f_08	60.8075	33.9613	25.0592	8.87829	67.4343	196.14059
Co_f_06	48.9984	27.7255	23.5991	2.13677	66.0349	168.49467
Pa_m_UP-Mel_01	47.1827	48.7715	28.6823	0.939169	34.0904	159.666069
Co_f_10	40.9297	26.9097	15.5889	0.920074	31.2753	115.623674
Pa_m_UP-Mel_03	45.8346	21.2189	17.1226	2.76461	23.1864	110.12711
Co_m_04	15.6498	28.7875	7.7389	1.48216	48.9529	102.61126
Pa_m_UP-Mel_02	18.1763	12.5658	5.02547	0.599598	17.4755	53.842668
Pa_m_UP-Mel_04	9.42419	9.92479	4.89164	0.704307	14.052	38.996927

Table B.1: Old Paradigm - Facial activity

	Bill Cosby (Happy)	The Champ (Sad)	Weather (Happy)	Sea of Love (Surprise)	Cry Freedom (Anger)
Co_f_09	94.3759	182.7249	215.302	218.94579	297.95239
Co_m_02	66.0356	145.1046	174.6215	183.49153	282.88033
Co_f_07	76.6297	148.0355	172.4629	174.74625	278.41025
Co_f_03	68.7369	143.0281	173.0773	178.61345	265.37085
Co_m_01	49.9946	108.0211	151.1613	155.4251	235.7232
Co_f_05	49.8066	106.728	140.0394	147.13477	212.99867
Co_f_08	60.8075	94.7688	119.828	128.70629	196.14059
Co_f_06	48.9984	76.7239	100.323	102.45977	168.49467
Pa_m_UP-Mel_01	47.1827	95.9542	124.6365	125.575669	159.666069
Co_f_10	40.9297	67.8394	83.4283	84.348374	115.623674
Pa_m_UP-Mel_03	45.8346	67.0535	84.1761	86.94071	110.12711
Co_m_04	15.6498	44.4373	52.1762	53.65836	102.61126
Pa_m_UP-Mel_02	18.1763	30.7421	35.76757	36.367168	53.842668
Pa_m_UP-Mel_04	9.42419	19.34898	24.24062	24.944927	38.996927

Table B.2: Old Paradigm - Accumulated facial activity

	Bill Cosby (Happy)
Co_f_09	94.3759
Co_f_07	76.6297
Co_f_03	68.7369
Co_m_02	66.0356
Co_f_08	60.8075
Co_m_01	49.9946
Co_f_05	49.8066
Co_f_06	48.9984
Pa_m_UP-Mel_01	47.1827
Pa_m_UP-Mel_03	45.8346
Co_f_10	40.9297
Pa_m_UP-Mel_02	18.1763
Co_m_04	15.6498
Pa_m_UP-Mel_04	9.42419

(a) Old Paradigm - Facial activity (Bill Cosby)

	The Champ (Sad)
Co_f_09	88.349
Co_m_02	79.069
Co_f_03	74.2912
Co_f_07	71.4058
Co_m_01	58.0265
Co_f_05	56.9214
Pa_m_UP-Mel_01	48.7715
Co_f_08	33.9613
Co_m_04	28.7875
Co_f_06	27.7255
Co_f_10	26.9097
Pa_m_UP-Mel_03	21.2189
Pa_m_UP-Mel_02	12.5658
Pa_m_UP-Mel_04	9.92479

(b) Old Paradigm - Facial activity (The Champ)

	Weather (Happy)
Co_m_01	43.1402
Co_f_05	33.3114
Co_f_09	32.5771
Co_f_03	30.0492
Co_m_02	29.5169
Pa_m_UP-Mel_01	28.6823
Co_f_08	25.0592
Co_f_07	24.4274
Co_f_06	23.5991
Pa_m_UP-Mel_03	17.1226
Co_f_10	15.5889
Co_m_04	7.7389
Pa_m_UP-Mel_02	5.02547
Pa_m_UP-Mel_04	4.89164

(c) Old Paradigm - Facial activity (Weather)

	Sea of Love (Surprise)
Co_f_08	8.87829
Co_m_02	8.87003
Co_f_05	7.09537
Co_f_03	5.53615
Co_m_01	4.2638
Co_f_09	3.64379
Pa_m_UP-Mel_03	2.76461
Co_f_07	2.28335
Co_f_06	2.13677
Co_m_04	1.48216
Pa_m_UP-Mel_01	0.939169
Co_f_10	0.920074
Pa_m_UP-Mel_04	0.704307
Pa_m_UP-Mel_02	0.599598

(d) Old Paradigm - Facial activity (Sea of Love)

	Cry Freedom (Anger)
Co_f_07	103.664
Co_m_02	99.3888
Co_f_03	86.7574
Co_m_01	80.2981
Co_f_09	79.0066
Co_f_08	67.4343
Co_f_06	66.0349
Co_f_05	65.8639
Co_m_04	48.9529
Pa_m_UP-Mel_01	34.0904
Co_f_10	31.2753
Pa_m_UP-Mel_03	23.1864
Pa_m_UP-Mel_02	17.4755
Pa_m_UP-Mel_04	14.052

(e) Old Paradigm - Facial activity (Cry Freedom)

Figure B.1: Old Paradigm - Facial activity for each video

		Sad	**Happy**	Neutral
Bill Cosby	Co_f_06	91	479	24
	Co_f_09	65	425	104
	Co_f_10	88	378	128
	Co_m_02	116	335	143
	Co_m_01	124	333	137
	Pa_m_UP-Mel_03	307	186	101
	Co_f_07	0	185	409
	Co_f_03	185	145	264
	Co_m_04	307	119	168
	Co_f_08	125	110	359
	Co_f_05	93	18	483
	Pa_m_UP-Mel_02	0	14	580
	Pa_m_UP-Mel_01	12	12	570
	Pa_m_UP-Mel_04	0	0	594
The Champ	Co_m_01	428	300	40
	Co_f_07	0	178	590
	Pa_m_UP-Mel_02	0	112	656
	Co_f_09	685	36	47
	Co_f_05	30	14	724
	Co_f_10	707	10	51
	Co_f_06	758	8	2
	Pa_m_UP-Mel_03	669	8	91
	Co_f_03	728	5	35
	Co_m_02	764	0	4
	Co_m_04	756	0	12
	Co_f_08	768	0	0
	Pa_m_UP-Mel_01	0	0	768
	Pa_m_UP-Mel_04	0	0	768
Weather	Co_f_09	46	221	2
	Co_m_01	33	200	36
	Co_m_02	60	198	11
	Pa_m_UP-Mel_02	0	195	74
	Co_f_06	88	181	0
	Co_f_05	20	180	69
	Co_f_10	244	25	0
	Co_f_07	0	24	245
	Co_f_03	244	13	12
	Pa_m_UP-Mel_03	155	9	105
	Co_m_04	257	0	12
	Co_f_08	269	0	0
	Pa_m_UP-Mel_01	0	0	269
	Pa_m_UP-Mel_04	0	0	269
Sea of Love	Co_m_01	7	51	3
	Co_m_02	17	28	16
	Co_f_06	36	25	0
	Co_f_05	6	23	32
	Co_f_07	0	18	43
	Co_f_09	2	2	57
	Co_f_08	22	1	38
	Pa_m_UP-Mel_03	41	1	19
	Co_f_03	61	0	0
	Co_m_04	56	0	5
	Co_f_10	61	0	0
	Pa_m_UP-Mel_01	0	0	61
	Pa_m_UP-Mel_02	0	0	61
	Pa_m_UP-Mel_04	0	0	61
Cry Freedom	Co_m_01	56	446	311
	Co_f_07	0	408	405
	Co_f_06	711	42	60
	Co_f_03	283	36	494
	Co_f_08	612	35	166
	Co_f_10	610	23	180
	Pa_m_UP-Mel_03	609	21	183
	Co_m_02	736	8	69
	Co_f_05	71	4	738
	Co_m_04	613	3	197
	Co_f_09	351	2	460
	Pa_m_UP-Mel_02	0	1	812
	Pa_m_UP-Mel_01	0	0	813
	Pa_m_UP-Mel_04	0	0	813

Table B.3: Old Paradigm - Facial expressions - sorted by happy within video

		Sad	Happy	Neutral
Bill Cosby	Pa_m_UP-Mel_03	307	186	101
	Co_m_04	307	119	168
	Co_f_03	185	145	264
	Co_f_08	125	110	359
	Co_m_01	124	333	137
	Co_m_02	116	335	143
	Co_f_05	93	18	483
	Co_f_06	91	479	24
	Co_f_10	88	378	128
	Co_f_09	65	425	104
	Pa_m_UP-Mel_01	12	12	570
	Co_f_07	0	185	409
	Pa_m_UP-Mel_02	0	14	580
	Pa_m_UP-Mel_04	0	0	594
The Champ	Co_f_08	768	0	0
	Co_m_02	764	0	4
	Co_f_06	758	8	2
	Co_m_04	756	0	12
	Co_f_03	728	5	35
	Co_f_10	707	10	51
	Co_f_09	685	36	47
	Pa_m_UP-Mel_03	669	8	91
	Co_m_01	428	300	40
	Co_f_05	30	14	724
	Co_f_07	0	178	590
	Pa_m_UP-Mel_02	0	112	656
	Pa_m_UP-Mel_01	0	0	768
	Pa_m_UP-Mel_04	0	0	768
Weather	Co_f_08	269	0	0
	Co_m_04	257	0	12
	Co_f_10	244	25	0
	Co_f_03	244	13	12
	Pa_m_UP-Mel_03	155	9	105
	Co_f_06	88	181	0
	Co_m_02	60	198	11
	Co_f_09	46	221	2
	Co_m_01	33	200	36
	Co_f_05	20	180	69
	Pa_m_UP-Mel_02	0	195	74
	Co_f_07	0	24	245
	Pa_m_UP-Mel_01	0	0	269
	Pa_m_UP-Mel_04	0	0	269
Sea of Love	Co_f_03	61	0	0
	Co_f_10	61	0	0
	Co_m_04	56	0	5
	Pa_m_UP-Mel_03	41	1	19
	Co_f_06	36	25	0
	Co_f_08	22	1	38
	Co_m_02	17	28	16
	Co_m_01	7	51	3
	Co_f_05	6	23	32
	Co_f_09	2	2	57
	Co_f_07	0	18	43
	Pa_m_UP-Mel_01	0	0	61
	Pa_m_UP-Mel_02	0	0	61
	Pa_m_UP-Mel_04	0	0	61
Cry Freedom	Co_m_02	736	8	69
	Co_f_06	711	42	60
	Co_m_04	613	3	197
	Co_f_08	612	35	166
	Co_f_10	610	23	180
	Pa_m_UP-Mel_03	609	21	183
	Co_f_09	351	2	460
	Co_f_03	283	36	494
	Co_f_05	71	4	738
	Co_m_01	56	446	311
	Co_f_07	0	408	405
	Pa_m_UP-Mel_02	0	1	812
	Pa_m_UP-Mel_01	0	0	813
	Pa_m_UP-Mel_04	0	0	813

Table B.4: Old Paradigm - Facial Expressions - sorted by sad within video

		Sad	Happy	**Neutral**
Bill Cosby	Pa_m_UP-Mel_04	0	0	594
	Pa_m_UP-Mel_02	0	14	580
	Pa_m_UP-Mel_01	12	12	570
	Co_f_05	93	18	483
	Co_f_07	0	185	409
	Co_f_08	125	110	359
	Co_f_03	185	145	264
	Co_m_04	307	119	168
	Co_m_02	116	335	143
	Co_m_01	124	333	137
	Co_f_10	88	378	128
	Co_f_09	65	425	104
	Pa_m_UP-Mel_03	307	186	101
	Co_f_06	91	479	24
The Champ	Pa_m_UP-Mel_01	0	0	768
	Pa_m_UP-Mel_04	0	0	768
	Co_f_05	30	14	724
	Pa_m_UP-Mel_02	0	112	656
	Co_f_07	0	178	590
	Pa_m_UP-Mel_03	669	8	91
	Co_f_10	707	10	51
	Co_f_09	685	36	47
	Co_m_01	428	300	40
	Co_f_03	728	5	35
	Co_m_04	756	0	12
	Co_m_02	764	0	4
	Co_f_06	758	8	2
	Co_f_08	768	0	0
Weather	Pa_m_UP-Mel_01	0	0	269
	Pa_m_UP-Mel_04	0	0	269
	Co_f_07	0	24	245
	Pa_m_UP-Mel_03	155	9	105
	Pa_m_UP-Mel_02	0	195	74
	Co_f_05	20	180	69
	Co_m_01	33	200	36
	Co_m_04	257	0	12
	Co_f_03	244	13	12
	Co_m_02	60	198	11
	Co_f_09	46	221	2
	Co_f_08	269	0	0
	Co_f_10	244	25	0
	Co_f_06	88	181	0
Sea of Love	Pa_m_UP-Mel_01	0	0	61
	Pa_m_UP-Mel_02	0	0	61
	Pa_m_UP-Mel_04	0	0	61
	Co_f_09	2	2	57
	Co_f_07	0	18	43
	Co_f_08	22	1	38
	Co_f_05	6	23	32
	Pa_m_UP-Mel_03	41	1	19
	Co_m_02	17	28	16
	Co_m_04	56	0	5
	Co_m_01	7	51	3
	Co_f_03	61	0	0
	Co_f_10	61	0	0
	Co_f_06	36	25	0
Cry Freedom	Pa_m_UP-Mel_01	0	0	813
	Pa_m_UP-Mel_04	0	0	813
	Pa_m_UP-Mel_02	0	1	812
	Co_f_05	71	4	738
	Co_f_03	283	36	494
	Co_f_09	351	2	460
	Co_f_07	0	408	405
	Co_m_01	56	446	311
	Co_m_04	613	3	197
	Pa_m_UP-Mel_03	609	21	183
	Co_f_10	610	23	180
	Co_f_08	612	35	166
	Co_m_02	736	8	69
	Co_f_06	711	42	60

Table B.5: Old Paradigm - Facial Expressions - sorted by neutral within video

	Bill Cosby (Happy)	The Champ (Sad)	Weather (Happy)	Silence of the Lambs (Fear)	Cry Freedom (Anger)	The Shining (Fear)	Capricorn One (Surprise)	Total
Pa_f_UP_BP2_08	61.9718	34.5945	67.7825	91.4877	149.151	69.3599	43.5648	517.9122
Pa_f_UP-NonMel_09	74.982	43.9521	27.4592	84.8893	48.3166	29.007	23.8494	332.4556
Co_m_11	47.8463	34.5049	40.8223	51.2466	43.3826	25.3941	14.8772	258.074
Pa_f_UP-NonMel_06	35.8466	41.9455	17.4541	62.5193	49.7708	24.5637	17.7679	249.8679
Co_f_12	58.5707	39.9424	25.3561	43.6596	26.4961	24.3349	18.4008	236.7606
Co_f_15	50.9681	29.948	17.1747	46.0423	28.2632	14.003	15.9124	202.3117
Pa_m_Unkown_07	42.8085	32.9788	19.9293	40.8624	37.639	14.1273	10.4134	198.7587
Co_m_13	30.6725	16.9278	26.0144	73.5504	18.4885	11.9067	12.2761	189.8364
Co_m_16	34.7066	26.5557	17.9017	53.904	29.9717	9.40495	17.0196	189.46425
Co_m_14	40.6977	20.1737	38.1457	31.162	16.2503	11.0059	13.7803	171.2156
Pa_m_PD_11	31.7382	22.6704	17.5711	30.8778	18.0523	4.66098	11.1434	136.71418
Pa_f_UP-Mel_10	21.657	21.373	10.28	15.0949	8.99202	5.11657	3.99677	86.51026
Pa_m_UP-Mel_05	5.82715	6.73422	3.32408	12.3734	8.72638	5.31982	3.20501	45.51006

Table B.6: New Paradigm - Accumulated facial activity

	Bill Cosby (Happy)
Pa_f_UP-NonMel_09	74.982
Pa_f_UP_BP2_08	61.9718
Co_f_12	58.5707
Co_f_15	50.9681
Co_m_11	47.8463
Pa_m_Unkown_07	42.8085
Co_m_14	40.6977
Pa_f_UP-NonMel_06	35.8466
Co_m_16	34.7066
Pa_m_PD_11	31.7382
Co_m_13	30.6725
Pa_f_UP-Mel_10	21.657
Pa_m_UP-Mel_05	5.82715

(a) New Paradigm - Facial activity (Bill Cosby)

	The Champ (Sad)
Pa_f_UP-NonMel_09	43.9521
Pa_f_UP-NonMel_06	41.9455
Co_f_12	39.9424
Pa_f_UP_BP2_08	34.5945
Co_m_11	34.5049
Pa_m_Unkown_07	32.9788
Co_f_15	29.948
Co_m_16	26.5557
Pa_m_PD_11	22.6704
Pa_f_UP-Mel_10	21.373
Co_m_14	20.1737
Co_m_13	16.9278
Pa_m_UP-Mel_05	6.73422

(b) New Paradigm - Facial activity (The Champ)

	Weather (Happy)
Pa_f_UP_BP2_08	67.7825
Co_m_11	40.8223
Co_m_14	38.1457
Pa_f_UP-NonMel_09	27.4592
Co_m_13	26.0144
Co_f_12	25.3561
Pa_m_Unkown_07	19.9293
Co_m_16	17.9017
Pa_m_PD_11	17.5711
Pa_f_UP-NonMel_06	17.4541
Co_f_15	17.1747
Pa_f_UP-Mel_10	10.28
Pa_m_UP-Mel_05	3.32408

(c) New Paradigm - Facial activity (Weather)

	Silence of the Lambs (Fear)
Pa_f_UP_BP2_08	91.4877
Pa_f_UP-NonMel_09	84.8893
Co_m_13	73.5504
Pa_f_UP-NonMel_06	62.5193
Co_m_16	53.904
Co_m_11	51.2466
Co_f_15	46.0423
Co_f_12	43.6596
Pa_m_Unkown_07	40.8624
Co_m_14	31.162
Pa_m_PD_11	30.8778
Pa_f_UP-Mel_10	15.0949
Pa_m_UP-Mel_05	12.3734

(d) New Paradigm - Facial activity (Silence of the Lambs)

	Cry Freedom (Anger)
Pa_f_UP_BP2_08	149.151
Pa_f_UP-NonMel_06	49.7708
Pa_f_UP-NonMel_09	48.3166
Co_m_11	43.3826
Pa_m_Unkown_07	37.639
Co_m_16	29.9717
Co_f_15	28.2632
Co_f_12	26.4961
Co_m_13	18.4885
Pa_m_PD_11	18.0523
Co_m_14	16.2503
Pa_f_UP-Mel_10	8.99202
Pa_m_UP-Mel_05	8.72638

(e) New Paradigm - Facial activity (Cry Freedom)

	The Shining (Fear)
Pa_f_UP_BP2_08	69.3599
Pa_f_UP-NonMel_09	29.007
Co_m_11	25.3941
Pa_f_UP-NonMel_06	24.5637
Co_f_12	24.3349
Pa_m_Unkown_07	14.1273
Co_f_15	14.003
Co_m_13	11.9067
Co_m_14	11.0059
Co_m_16	9.40495
Pa_m_UP-Mel_05	5.31982
Pa_f_UP-Mel_10	5.11657
Pa_m_PD_11	4.66098

(f) New Paradigm - Facial activity (The Shining)

	Capricorn One (Surprise)
Pa_f_UP_BP2_08	43.5648
Pa_f_UP-NonMel_09	23.8494
Co_f_12	18.4008
Pa_f_UP-NonMel_06	17.7679
Co_m_16	17.0196
Co_f_15	15.9124
Co_m_11	14.8772
Co_m_14	13.7803
Co_m_13	12.2761
Pa_m_PD_11	11.1434
Pa_m_Unkown_07	10.4134
Pa_f_UP-Mel_10	3.99677
Pa_m_UP-Mel_05	3.20501

(g) New Paradigm - Facial activity (Capricorn One)

Figure B.2: New Paradigm - Facial activity for each video

		Sad	**Happy**	Neutral
Bill Cosby	Pa_f_UP-NonMel_09	6	549	39
	Co_m_11	7	451	136
	Pa_f_UP-NonMel_06	177	321	96
	Co_f_15	6	310	278
	Co_f_12	179	300	115
	Pa_f_UP_BP2_08	71	281	242
	Pa_m_PD_11	0	182	412
	Co_m_16	0	63	531
	Co_m_14	141	62	391
	Co_m_13	479	50	65
	Pa_m_UP-Mel_05	0	0	594
	Pa_m_Unkown_07	0	0	594
	Pa_f_UP-Mel_10	0	0	594
The Champ	Pa_f_UP-NonMel_09	14	356	398
	Pa_f_UP_BP2_08	479	130	159
	Co_f_12	606	70	92
	Co_m_16	0	46	722
	Pa_f_UP-NonMel_06	721	45	2
	Co_f_15	481	16	271
	Co_m_14	748	15	5
	Co_m_13	592	12	164
	Co_m_11	681	10	77
	Pa_m_UP-Mel_05	0	0	768
	Pa_m_Unkown_07	0	0	768
	Pa_f_UP-Mel_10	0	0	768
	Pa_m_PD_11	0	0	768
Weather	Pa_f_UP-NonMel_06	45	224	0
	Pa_m_PD_11	0	206	63
	Pa_f_UP-NonMel_09	70	147	52
	Co_f_12	133	111	25
	Co_m_11	37	104	128
	Co_m_16	0	104	165
	Co_m_14	81	76	112
	Co_m_13	114	35	120
	Pa_f_UP_BP2_08	23	24	222
	Co_f_15	213	6	50
	Pa_m_UP-Mel_05	0	0	269
	Pa_m_Unkown_07	0	0	269
	Pa_f_UP-Mel_10	0	0	269
Silence of the Lambs	Pa_m_PD_11	0	392	614
	Pa_f_UP_BP2_08	261	293	452
	Pa_f_UP-NonMel_09	334	110	562
	Pa_f_UP-NonMel_06	351	58	597
	Co_f_12	234	40	732
	Co_m_16	0	39	967
	Co_m_14	977	18	11
	Co_m_13	636	10	360
	Co_f_15	515	6	485
	Co_m_11	7	3	996
	Pa_m_UP-Mel_05	0	0	1006
	Pa_m_Unkown_07	0	0	1006
	Pa_f_UP-Mel_10	0	0	1006
Cry Freedom	Pa_f_UP-NonMel_09	36	169	455
	Co_f_12	7	98	555
	Pa_m_PD_11	0	40	620
	Co_m_14	629	29	2
	Co_f_15	629	26	5
	Pa_f_UP_BP2_08	430	24	206
	Co_m_16	0	13	647
	Pa_f_UP-NonMel_06	228	12	420
	Co_m_13	646	10	4
	Co_m_11	111	1	548
	Pa_m_UP-Mel_05	0	0	660
	Pa_m_Unkown_07	0	0	660
	Pa_f_UP-Mel_10	0	0	660
The Shining	Pa_f_UP-NonMel_09	5	110	193
	Co_f_12	244	39	25
	Co_m_16	0	21	287
	Pa_f_UP_BP2_08	97	21	190
	Co_f_15	177	6	125
	Co_m_14	292	5	11
	Co_m_13	305	1	2
	Pa_f_UP-NonMel_06	18	1	289
	Co_m_11	83	0	225
	Pa_m_UP-Mel_05	0	0	308
	Pa_m_Unkown_07	0	0	308
	Pa_f_UP-Mel_10	0	0	308
	Pa_m_PD_11	0	0	308
Capricorn One	Pa_f_UP-NonMel_09	11	124	98
	Co_m_16	0	19	214
	Pa_f_UP_BP2_08	152	16	65
	Co_m_14	203	8	22
	Co_f_12	68	5	160
	Pa_f_UP-NonMel_06	104	2	127
	Co_m_11	0	0	233
	Co_m_13	207	0	26
	Co_f_15	186	0	47
	Pa_m_UP-Mel_05	0	0	233
	Pa_m_Unkown_07	0	0	233
	Pa_f_UP-Mel_10	0	0	233
	Pa_m_PD_11	0	0	233

Table B.7: New Paradigm - Facial expressions - sorted by happy within video

		Sad	Happy	Neutral
Bill Cosby	Co_m_13	479	50	65
	Co_f_12	179	300	115
	Pa_f_UP-NonMel_06	177	321	96
	Co_m_14	141	62	391
	Pa_f_UP_BP2_08	71	281	242
	Co_m_11	7	451	136
	Co_f_15	6	310	278
	Pa_f_UP-NonMel_09	6	549	39
	Co_m_16	0	63	531
	Pa_m_UP-Mel_05	0	0	594
	Pa_m_Unkown_07	0	0	594
	Pa_f_UP-Mel_10	0	0	594
	Pa_m_PD_11	0	182	412
The Champ	Co_m_14	748	15	5
	Pa_f_UP-NonMel_06	721	45	2
	Co_m_11	681	10	77
	Co_f_12	606	70	92
	Co_m_13	592	12	164
	Co_f_15	481	16	271
	Pa_f_UP_BP2_08	479	130	159
	Pa_f_UP-NonMel_09	14	356	398
	Co_m_16	0	46	722
	Pa_m_UP-Mel_05	0	0	768
	Pa_m_Unkown_07	0	0	768
	Pa_f_UP-Mel_10	0	0	768
	Pa_m_PD_11	0	0	768
Weather	Co_f_15	213	6	50
	Co_f_12	133	111	25
	Co_m_13	114	35	120
	Co_m_14	81	76	112
	Pa_f_UP-NonMel_09	70	147	52
	Pa_f_UP-NonMel_06	45	224	0
	Co_m_11	37	104	128
	Pa_f_UP_BP2_08	23	24	222
	Co_m_16	0	104	165
	Pa_m_UP-Mel_05	0	0	269
	Pa_m_Unkown_07	0	0	269
	Pa_f_UP-Mel_10	0	0	269
	Pa_m_PD_11	0	206	63
Silence of the Lambs	Co_m_14	977	18	11
	Co_m_13	636	10	360
	Co_f_15	515	6	485
	Pa_f_UP-NonMel_06	351	58	597
	Pa_f_UP-NonMel_09	334	110	562
	Pa_f_UP_BP2_08	261	293	452
	Co_f_12	234	40	732
	Co_m_11	7	3	996
	Co_m_16	0	39	967
	Pa_m_UP-Mel_05	0	0	1006
	Pa_m_Unkown_07	0	0	1006
	Pa_f_UP-Mel_10	0	0	1006
	Pa_m_PD_11	0	392	614
Cry Freedom	Co_m_13	646	10	4
	Co_m_14	629	29	2
	Co_f_15	629	26	5
	Pa_f_UP_BP2_08	430	24	206
	Pa_f_UP-NonMel_06	228	12	420
	Co_m_11	111	1	548
	Pa_f_UP-NonMel_09	36	169	455
	Co_f_12	7	98	555
	Co_m_16	0	13	647
	Pa_m_UP-Mel_05	0	0	660
	Pa_m_Unkown_07	0	0	660
	Pa_f_UP-Mel_10	0	0	660
	Pa_m_PD_11	0	40	620
The Shining	Co_m_13	305	1	2
	Co_m_14	292	5	11
	Co_f_12	244	39	25
	Co_f_15	177	6	125
	Pa_f_UP_BP2_08	97	21	190
	Co_m_11	83	0	225
	Pa_f_UP-NonMel_06	18	1	289
	Pa_f_UP-NonMel_09	5	110	193
	Co_m_16	0	21	287
	Pa_m_UP-Mel_05	0	0	308
	Pa_m_Unkown_07	0	0	308
	Pa_f_UP-Mel_10	0	0	308
	Pa_m_PD_11	0	0	308
Capricorn One	Co_m_13	207	0	26
	Co_m_14	203	8	22
	Co_f_15	186	0	47
	Pa_f_UP_BP2_08	152	16	65
	Pa_f_UP-NonMel_06	104	2	127
	Co_f_12	68	5	160
	Pa_f_UP-NonMel_09	11	124	98
	Co_m_11	0	0	233
	Co_m_16	0	19	214
	Pa_m_UP-Mel_05	0	0	233
	Pa_m_Unkown_07	0	0	233
	Pa_f_UP-Mel_10	0	0	233
	Pa_m_PD_11	0	0	233

Table B.8: New Paradigm - Facial expressions - sorted by sad within video

		Sad	Happy	**Neutral**
Bill Cosby	Pa_m_UP-Mel_05	0	0	594
	Pa_m_Unkown_07	0	0	594
	Pa_f_UP-Mel_10	0	0	594
	Co_m_16	0	63	531
	Pa_m_PD_11	0	182	412
	Co_m_14	141	62	391
	Co_f_15	6	310	278
	Pa_f_UP_BP2_08	71	281	242
	Co_m_11	7	451	136
	Co_f_12	179	300	115
	Pa_f_UP-NonMel_06	177	321	96
	Co_m_13	479	50	65
	Pa_f_UP-NonMel_09	6	549	39
The Champ	Pa_m_UP-Mel_05	0	0	768
	Pa_m_Unkown_07	0	0	768
	Pa_f_UP-Mel_10	0	0	768
	Pa_m_PD_11	0	0	768
	Co_m_16	0	46	722
	Pa_f_UP-NonMel_09	14	356	398
	Co_f_15	481	16	271
	Co_m_13	592	12	164
	Pa_f_UP_BP2_08	479	130	159
	Co_f_12	606	70	92
	Co_m_11	681	10	77
	Co_m_14	748	15	5
	Pa_f_UP-NonMel_06	721	45	2
Weather	Pa_m_UP-Mel_05	0	0	269
	Pa_m_Unkown_07	0	0	269
	Pa_f_UP-Mel_10	0	0	269
	Pa_f_UP_BP2_08	23	24	222
	Co_m_16	0	104	165
	Co_m_11	37	104	128
	Co_m_13	114	35	120
	Co_m_14	81	76	112
	Pa_m_PD_11	0	206	63
	Pa_f_UP-NonMel_09	70	147	52
	Co_f_15	213	6	50
	Co_f_12	133	111	25
	Pa_f_UP-NonMel_06	45	224	0
Silence of the Lambs	Pa_m_UP-Mel_05	0	0	1006
	Pa_m_Unkown_07	0	0	1006
	Pa_f_UP-Mel_10	0	0	1006
	Co_m_11	7	3	996
	Co_m_16	0	39	967
	Co_f_12	234	40	732
	Pa_m_PD_11	0	392	614
	Pa_f_UP-NonMel_06	351	58	597
	Pa_f_UP-NonMel_09	334	110	562
	Co_f_15	515	6	485
	Pa_f_UP_BP2_08	261	293	452
	Co_m_13	636	10	360
	Co_m_14	977	18	11
Cry Freedom	Pa_m_UP-Mel_05	0	0	660
	Pa_m_Unkown_07	0	0	660
	Pa_f_UP-Mel_10	0	0	660
	Co_m_16	0	13	647
	Pa_m_PD_11	0	40	620
	Co_f_12	7	98	555
	Co_m_11	111	1	548
	Pa_f_UP-NonMel_09	36	169	455
	Pa_f_UP-NonMel_06	228	12	420
	Pa_f_UP_BP2_08	430	24	206
	Co_f_15	629	26	5
	Co_m_13	646	10	4
	Co_m_14	629	29	2
The Shining	Pa_m_UP-Mel_05	0	0	308
	Pa_m_Unkown_07	0	0	308
	Pa_f_UP-Mel_10	0	0	308
	Pa_m_PD_11	0	0	308
	Pa_f_UP-NonMel_06	18	1	289
	Co_m_16	0	21	287
	Co_m_11	83	0	225
	Pa_f_UP-NonMel_09	5	110	193
	Pa_f_UP_BP2_08	97	21	190
	Co_f_15	177	6	125
	Co_f_12	244	39	25
	Co_m_14	292	5	11
	Co_m_13	305	1	2
Capricorn One	Co_m_11	0	0	233
	Pa_m_UP-Mel_05	0	0	233
	Pa_m_Unkown_07	0	0	233
	Pa_f_UP-Mel_10	0	0	233
	Pa_m_PD_11	0	0	233
	Co_m_16	0	19	214
	Co_f_12	68	5	160
	Pa_f_UP-NonMel_06	104	2	127
	Pa_f_UP-NonMel_09	11	124	98
	Pa_f_UP_BP2_08	152	16	65
	Co_f_15	186	0	47
	Co_m_13	207	0	26
	Co_m_14	203	8	22

Table B.9: New Paradigm - Facial expressions - sorted by neutral within video

Bibliography

[Alvinoa 07] C. Alvinoa, C. Kohlerb, F. Barrett, and R. Gurb. *Computerized measurement of facial expression of emotions in schizophrenia*. Journal of Neuroscience Methods, 163(6):350–361, July 2007.

[Annesley 05] J. Annesley and J. Orwell. *On the Use of MPEG-7 for Visual Surveillance*. Technical report, Digital Imaging Research Center, Kingston University, Kingston-upon-Thames, Surrey, UK., 2005.

[Anolli 97] L. Anolli and R. Ciceri. *The Voice of Deception: Vocal Strategies of Naive and Able Liars*. Journal of Nonverbal Behavior, 21:259–284, 1997.

[Ashraf 09] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, and P. Solomon. *The painful face - Pain expression recognition using active appearance models*. Image Vision Computing, 27(12):1788–1796, 2009.

[Asthana 09] A. Asthana, R. Göcke, N. Quadrianto, and T. Gedeon. *Learning Based Automatic Face Annotation for Arbitrary Poses and Expressions from Frontal Images Only*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2009, Miami (FL), USA, June 2009. IEEE Computer Society.

[Athanaselisa 05] T. Athanaselisa, S. Bakamidisa, I. Dologloua, R. Cowie, E. Douglas-Cowie, and C. Cox. *ASR for emotional speech: Clarifying the issues and enhancing performance*. Neural Networks, 18:437–444, 2005.

[AVT] AVT. *Allied Vision Technologies*. `http://www.alliedvisiontec.com/emea/products/cameras.html`. last accessed, 13 April 2010.

[Baggia 08] P. Baggia, F. Burkhardt, J.-C. Martin, C. Pelachaud, C. Peter, B. Schuller, I. Wilson, and E. Zovato. *Elements of an EmotionML 1.0, W3C Incubator Group Report, 20 Nov. 2008*. 2008. Schrder, M. (ed.), 34 p.

[Baker 01] S. Baker and I. Matthews. *Equivalence and Efficiency of Image Alignment Algorithms*. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 1:1090–1097, 2001.

[Baker 02] S. Baker and I. Matthews. *Lucas-Kanade 20 Years On: A Unifying Framework: Part 1*. Technical Report CMU-RI-TR-02-16, Robotics Institute, Pittsburgh, PA, July 2002.

[Baker 03a] S. Baker, R. Gross, and I. Matthews. *Lucas-Kanade 20 years on: A unifying framework: Part 3*. Technical Report CMU-RI-TR-03-35, Robotics Institute, Carnegie Mellon University, Pittsburgh (PA), USA, November 2003.

[Baker 03b] S. Baker, R. Gross, I. Matthews, and T. Ishikawa. *Lucas-Kanade 20 Years On: A Unifying Framework: Part 2*. Technical Report CMU-RI-TR-03-01, Robotics Institute, Pittsburgh, PA, February 2003.

[Baker 04a] S. Baker, R. Gross, and I. Matthews. *Lucas-Kanade 20 Years On: A Unifying Framework: Part 4*. Technical Report CMU-RI-TR-04-14, Robotics Institute, Pittsburgh, PA, February 2004.

[Baker 04b] S. Baker, R. Patil, K. Cheung, and I. Matthews. *Lucas-Kanade 20 Years On: Part 5*. Technical Report CMU-RI-TR-04-64, Robotics Institute, Pittsburgh, PA, November 2004.

[Bartlett 99] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. *Measuring Facial Expressions by Computer Image Analysis*. Psychophysiology, 36:253–263, 1999.

[Bartlett 02] M. Bartlett, G. Littlewort, B. Braathen, T. Sejnowski, and J. Movellan. *A Prototype for Automatic Recognition of Spontaneous Facial Actions*. In Advances in Neural Information Processing Systems, pages 1271–1278, 2002.

[Bartlett 03] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. *Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction*. In In CVPR Workshop on CVPR for HCI, 2003.

[Bartlett 05] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. *Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 568–573, 2005.

[Bartlett 06] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and Javier R. Movellan. *Fully Automatic Facial Action Recognition in Spontaneous Behavior*. In 7th International Conference on Automatic Face and Gesture Recognition, pages 223–230, 2006.

[Benitez 03] A. Benitez and S. Chang. *Automatic Multimedia Knowledge Discovery, Summarization and Evaluation*. Technical report, Department of Electrical Engineering, Columbia University, 2003.

[Bertini 05] M. Bertini, A. Del Bimbo, and C. Torniai. *Video Annotation and Retrieval with Pictorially Enriched Ontologies*. Technical report, Universit`a di Firenze - Italy, 2005.

[beyondblue] beyondblue. `http://www.beyondblue.org.au/index.aspx?link_id=90`. Last accessed 11 December 2009.

[Bhuiyan 07] A. Bhuiyan and C. Liu. *On Face Recognition using Gabor Filters*. In Proceedings of World Academy of Science, Engineering and Technology, 22, pages 51–56, 2007.

[Blanz 99] V. Blanz and T. Vetter. *A Morphable Model for the Synthesis of 3D Faces*. In Special Interest Group on Graphics and Interactive Techniques, pages 187–194, 1999.

[Bower 92] G.H. Bower. The handbook of emotion and memory: Research and theory, chapter How Might Emotions Affect Learning?, pages 3–31. Lawrence Erlbaum Associates, Inc, 365 Broadway, Hillsdale, New Jersey 07642, 1992.

[Brierley 07] B. Brierley, N. Medford, P. Shaw, and A. Davidson. *Emotional memory for words: Separating content and context*. Cognition & Emotion, 21 (3):495–521, 2007.

[Buchanan 02] H. Buchanan and N. Niven. *Validation of a Facial Image Scale to assess child dental anxiety.* International Journal of Paediatric Dentistry, 12(1):47–52, January 2002.

[Buchheim 07] A. Buchheim and C. Benecke. *Affective facial behavior of patients with anxiety disorders during the adult attachment interview: a pilot study*. Psychotherapie Psychosomatik Medizinische Psychologie, 8(57):343–347, March 2007.

[Burges 98] C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, 2(2):121–167, 1998.

[Casagrande 06] N. Casagrande. *Multiboost: An open source multi-class adaboost learner*. 2005-2006. `http://www.iro.umontreal.ca/casagran/multiboost.html`, last accessed 20 August 2008.

[Chang 01] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chen 03] W. Chen, T. Chiang, M. Hsu, and J. Liu. *The validity of eye blink rate in Chinese adults for the diagnosis of Parkinsons disease*. Clinical Neurology and Neurosurgery, 105:90–92, 2003.

[Chen 05] P. Chen, C. Lin, and B. Schölkopf. *A tutorial on V-support vector machines: Research Articles*. Applied Stochastic Models in Business and Industry, 21(2):111–136, 2005.

[Chen 07] F. Chen and K. Kotani. *Facial Expression Recognition by SVM-based Two-stage Classifier on Gabor Features*. In Machine Vision Applications, pages 453–456, 2007.

[Chiariglione 01] L. Chiariglione. *Introduction to MPEG-7: Multimedia Content Description Interface*. Technical report, Telecom Italia Lab, Italy, 2001.

[Cootes 92] T. Cootes and C. Taylor. *Active Shape Models - Smart Snakes*. British Machine Vision Conference, pages 266–275, 1992.

[Cootes 95] T. Cootes, C. Taylor, D. Cooper, and J. Graham. *Active Shape Models—their training and applications*. Computer Vision and Image Understanding, 61(1):38–59, 1995.

[Cootes 98] T. Cootes, G. Edwards, and C. Taylor. *Active Appearance Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1407:484–498, 1998.

[Cootes 01] T. Cootes and C. Taylor. *Statistical Models of Appearance for Computer Vision*. Technical report, University of Manchester, 2001.

[Cornelius 96] R. Cornelius. *The science of emotion*. New Jersey: Prentice Hall, 1996.

[Cover 67] T. Cover and P. Hart. *Nearest neighbor pattern classification*. Information Theory, IEEE Transactions on, 13(1):21–27, 1967.

[Cowie 03] R. Cowie and R. Cornelius. *Describing the emotional states that are expressed in speech*. Speech Communication, 40:5–32, 2003.

[Cowie 05a] R. Cowie, E. Douglas-Cowie, and C. Cox. *Beyond emotion archetypes: Databases for emotion modelling using neural networks*. Neural Networks, 18:371–388, 2005.

[Cowie 05b] R. Cowie, E. Douglas-Cowie, J. Taylor, S. Ioannou, M. Wallace, and S. Kollias. *An Intelligent System For Facial Emotion Recognition*. IEEE International Conference on Multimedia and Expo, 2005.

[cvG] *cvGabor c++ source codeDownload*. `http://www.personal.rdg.ac.uk/~sir02mz/`. Last accessed 24 January 2010.

[Daugman] Daugman. *Computer Science Tripos: 16 Lectures by J G Daugman*. `http://www.cl.cam.ac.uk/teaching/0910/CompVision/LectureNotes2010.pdf`. Last accessed 28 January 2010.

[Daugman 85] J. Daugman. *Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters*.

Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2(7):1160–1169, 1985.

[Davidson 04] R. Davidson, J. S. Maxwell, and A. J. Shackman. *The privileged status of emotion in the brain*. Proceedings of the National Academy of Sciences USA, 101:11915–11916, August 2004.

[Dellaert 96] F. Dellaert, T. Polzin, and A. Waibel. *Recognizing Emotion in Speech*. International Conference on Spoken Language Processing, October 1996.

[Devillers 05] L. Devillers, L. Vidrascu, and L. Lamel. *Challenges in real-life emotion annotation and machine learning based detection*. Neural Networks, 18:407–422, 2005.

[Donato 99] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. *Classifying Facial Actions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21:974–989, 1999.

[Dunn 95] D. Dunn and W. Higgins. *Optimal Gabor filters for texture segmentation*. Image Processing, IEEE Transactions on, 4(7):947–964, 1995.

[Edwards 98a] G. Edwards, C.Taylor, and T Cootes. *Interpreting Face Images using Active Appearance Models*. 3rd. International Conference on Face & Gesture Recognition, 1998.

[Edwards 98b] G. Edwards, C.J. Taylor, and T.F. Cootes. *Interpreting Face Images Using Active Appearance Models*. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition FG'98, pages 300–305, Nara, Japan, April 1998. IEEE.

[Ekman 71] P. Ekman and W. Friesen. *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology, 17(2):124–129, 02 1971.

[Ekman 75] P. Ekman and W. Friesen. *Unmasking the Face*. Prentice Hall, Englewood Cliffs NJ, 1975.

[Ekman 82] P. Ekman and H. Oster. *Emotion in the Human Faces*. New York: Cambridge University Press, 2nd edition, 1982.

[Ekman 97] P. Ekman and E.L. Rosenberg. *What the Face Reveals*. Series in Affective Science. Oxford University Press, Oxford, UK, 1997.

[Ekman 99] P. Ekman. Handbook of cognition and emotions, chapter Basic Emotions, pages 301–320. Wiley, New York., 1999.

[Ekman 03] P. Ekman. *Darwin, Deception, and Facial Expression*. Annals New York Academy of Sciences, pages 205–221, 2003.

[Ellgring 96] H. Ellgring and K. Scherer. *Vocal indicators of mood change in depression*. Journal of Nonverbal Behavior, 20:83–110, 1996.

[Ellgring 05] K. Scherer & H. Ellgring. *Multimodal markers of appraisal and emotion*. In Paper presented at the ISRE Conference, Bari, 2005.

[Ellgring 08] H. Ellgring. *Nonverbal communication in depression*. Cambridge University Press, Cambridge, UK, 2008.

[EMFACS] EMFACS. `http://www.face-and-emotion.com/dataface/facs/emfacs.jsp`. Last accessed 2 April 2009.

[Ezust 06] Alan Ezust and Paul Ezust. *An Introduction to Design Patterns in C++ with Qt 4 (Bruce Perens Open Source)*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2006.

[Fasel 02] I. Fasel, M. Bartlett, and J. Movellan. *A Comparison of Gabor Filter Methods for Automatic Detection of Facial Landmarks*. Fifth IEEE International Conference on Automatic Face and Gesture Recognition, page 242, 2002.

[Fasel 03] B. Fasel and J. Luettin. *Automatic facial expression analysis: a survey*. Pattern Recognition, 36(1):259–275, 2003.

[Flint 93] A. Flint, S. Black, I. Campbell-Taylor, G. Gailey, and C. Levinton. *Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression*. Journal of Psychiatric Research, 27:309–319, 1993.

[Frank 93] M. Frank, P. Ekman, and W. Friesen. *Behavioral markers and recognizability of the smile of enjoyment*. Journal of Personality and Social Psychology, 64(1):83–93, 1993.

[Freund 99] Y. Freund and R. Schapire. *A short introduction to boosting*. Journal of Japanese Society for Artificial Intelligence, 14(5):771–780, 1999.

[Fridlund 83] A. Fridlund and J. Izard. Social psychophysiology: A sourcebook, chapter Electromyographic studies of facial expressions of emotions and patterns of emotion, pages 163–218. Academic Press, New York, 1983.

[FRV] *Face Recognition Vendor Test*. Last accessed 24 November 2009.

[Fry 79] D. B. Fry. *The Physics of Speech*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, United Kingdom, 1979.

[Fu 08] C.H.Y. Fu, S.C.R. Williams, A.J. Cleare, J. Scott, M.T. Mitterschiffthaler, N.D. Walsh, C. Donaldson, J. Suckling, C. Andrew, H. Steiner, and R.M. Murray. *Neural Responses to Sad Facial Expressions in Major Depression*

Following Cognitive Behavioral Therapy. Biological Psychiatry, 64(6):505–512, 2008.

[Gao 09] X. Gao, Y. Su, X. Li, and D. Tao. *Gabor texture in active appearance models*. Neurocomputing, 72(13-15):3174–3181, 2009.

[Goeleven 06] E. Goeleven, R. De Raedt, S. Baert, and E.Koster. *Deficient inhibition of emotional information in depression*. Journal of Affective Disorders, 93(1-3):149–157, 2006.

[Grana 05] C. Grana, D. Bulgarelli, and R. Cucchiara. *Video Clip Clustering for Assisted Creation of MPEG-7 Pictorially Enriched Ontologies*. Technical report, University of Modena and Reggio Emilia, Italy, 2005.

[Grinker 61] R. Grinker, N. Miller, M. Sabshin, R. Nunn, and J. Nunnally. *The phenomena of depressions*. Harper and Row, New York, 1961.

[Gross 95] J. Gross and R. Levenson. *Emotion elicitation using films*. Cognition & Emotion, 9:87–108, 1995.

[Gross 05] R. Gross, I. Matthews, and S. Baker. *Generic vs. person specific active appearance models*. Image Vision Computing, 23(12):1080–1093, 2005.

[Hajdinjak 03] M. Hajdinjak and F. Mihelič. *Wizard of Oz Experiments*. In EUROCON, Ljubljana, Slovenia, 2003.

[Harrigan 96] J.A. Harrigan and D.M. O'Connell. *How do you look when feeling anxious? Facial displays of anxiety*. Personality and Individual Differences, 21:205–212, August 1996.

[Harrigan 97] J. Harrigan and K. Taing. *Fooled by a Smile: Detecting Anxiety in Others*. Journal of Nonverbal Behaviour, 21:203, 1997.

[Harrigan 04] J. Harrigan, K. Wilson, and R. Rosenthal. *Detecting state and trait anxiety from auditory and visual cues: a meta-analysis*. Personality and Social Psychology Bulletin, 30(1):56–66, 2004.

[Hirschberg 05] J. Hirschberg, S. Benus, and J. Brenierand F. Enos. *Distinguishing Deceptive from Non-Deceptive Speech*. In Interspeech, 2005.

[Hsu 03] C. Hsu, C. Chang, and C. Lin. *A Practical Guide to Support Vector Classification*. Bioinformatics, 2003.

[HUMAINE 06] HUMAINE. `http://emotion-research.net/`, 2006. Last accessed 24 January 2010.

[Hunter 02] J. Hunter. *Enhancing the Semantic Interoperability of Multimedia through a Core Ontology*. Technical report, Harmony Project, funded by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC), 2002.

[Jaimes 03] A. Jaimes and J. Smith. *Semi-Automatic Data-driven Construction of Multimedia Ontologies*. Iternational Conference on Multimedia and Expo, 2003.

[James 90] W. James. *Principles of Psychology*. Harvard University, 1890.

[Joormann 06] J. Joormann and I. Gotlib. *Is This Happiness I See? Biases in the Identification of Emotional Facial Expressions in Depression and Social Phobia*. Journal of Affective Disorders, 93(1-3):149–157, 2006.

[Joormann 07] J. Joormann and I. Gotlib. *Selective Attention to Emotional Faces Following Recovery From Depression*. Journal of Abnormal Psychology, 116(1):80–85, 2007.

[Kaiser 98] S. Kaiser, T. Wehrle, and S. Schmidt. *Emotional Episodes, Facial Expressions, and Reported Feelings in Human-Computer Interactions.* In ISRE Publications, pages 82–86, 1998.

[Kamarainen 06] J. Kamarainen, V. Kyrki, and H. Klviinen. *Invariance properties of Gabor filter-based features-overview and applications.* IEEE Transactions on Image Processing, 15(5):1088–1099, 2006.

[Kanade 00] T. Kanade, Y. Tian, and J. Cohn. *Comprehensive Database for Facial Expression Analysis*. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pages 46–53, 2000.

[Kessler 05] R. Kessler, W. Chiu, O. Demler, K. Merikangas, and E. Walters. *Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication.* Arch Gen Psychiatry, 62(6):617–27, 2005.

[Kohavi 95] R. Kohavi. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. International Joint Conference on Artificial Intelligence, pages 1137–1143, 1995.

[Koike 98] K. Koike, H. Suzuki, and H. Saito. *Prosodic Parameters in Emotional Speech*. In International Conference on Spoken Language Processing, pages 679–682, 1998.

[Kvaal 05] K. Kvaal, I. Ulstein, I. Nordhus, and K. Engedal. *The Spielberger State-Trait Anxiety Inventory (STAI): the state scale in detecting mental disorders in geriatric patients*. International Journal of Geriatric Psychiatry, 20:629–634, 2005.

[Lades 93] M. Lades, J.. Vorbrggen, J. Buhmann, J. Lange, C. Malsburg, R. Wrtz, and W. Konen. *Distortion Invariant Object Recognition in the Dynamic Link Architecture*. IEEE Trans. Computers, 42:300–311, 1993.

[Ladouceur 06] C. Ladouceur, R. Dahl, D. Williamson, B. Birmaher, D. Axelson, N. Ryan, and B. Casey. *Processing emotional facial expressions influences performance on a Go/NoGo task in pediatric anxiety and depression.* Journal of Child Psychology and Psychiatry and Allied Disciplines, 47(11):1107–1115, Nov 2006.

[Lagoze 01] C. Lagoze and J. Hunter. *The ABC Ontology and Model*. Technical report, Cornell University Ithaca, NY and DSTC Pty, Ltd. Brisbane, Australia, 2001.

[Lang 05] P. Lang, M. Bradley, and B. Cuthbert. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-6, University of Florida, Gainesville, FL, 2005.

[Lazarus 91] R. Lazarus. *Emotion and adaptation*. Oxford University Press, New York :, 1991.

[Lee 96] T. Lee. *Image Representation Using 2D Gabor Wavelets*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(10):959–971, 1996.

[Lee 07] B. Lee, S. Cho, and H. Khang. *The neural substrates of affective processing toward positive and negative affective pictures in patients with major depressive disorder*. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 31(7):1487–1492, 2007.

[Li 07] F. Li and K. Xu. *Optimal Gabor Kernel's Scale and orientation selection for face classification*. Optics & Laser Technology, 39(4):852–857, 2007.

[Lien 98] J. Lien, J. Cohn, T. Kanade, and C. Li. *Automated Facial Expression Recognition Based on FACS Action Units*. In Third IEEE International Conference on Automatic Face and Gesture Recognition, pages 390–395, April 1998.

[Liscombe 05] J. Liscombe, G. Riccardi, and D. Hakkani-Tür. *Using Context to Improve Emotion Detection in Spoken Dialog Systems*. In EUROSPEECH'05, 9th European Conference on Speech Communication and Technology, pages 1845–1848, September 2005.

[Littlewort 06] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. *An automatic system for measuring facial expression in video*. Computer Vision and Image Understanding, Special Issue on Face Processing in Video, 24(6):615–625, 2006.

[Littlewort 07] G. Littlewort, M. Bartlett, and K. Lee. *Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain*. 9th International Conference on Multimodal Interfaces, pages 15–21, 2007.

[Liu 04] C. Liu. *Gabor-based Kernel PCA with Fractional Power Polynomial Models for Face Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26:572–581, 2004.

[Liu 06] W. Liu and Z. Wang. *Facial Expression Recognition Based on Fusion of Multiple Gabor Features*. 18th International Conference on Pattern Recognition, 3:536–539, 2006.

[Lucas 81] B. Lucas and T. Kanade. *An Iterative Image Registration Technique with an Application to Stereo Vision*. International Joint Conferences on Artificial Intelligence, pages 674–679, April 1981.

[Lucey 06] Simon Lucey, Iain Matthews, Changbo Hu, Zara Ambadar, Fernando de la Torre, and Jeffrey Cohn. *AAM Derived Face Representations for Robust Facial Action Recognition*. Automatic Face and Gesture Recognition, IEEE International Conference on, 0:155–162, 2006.

[Luo 06] Hangzai Luo and Jianping Fan. *Building concept ontology for medical video annotation*. In MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, pages 57–60, New York, NY, USA, 2006. ACM.

[Manjunath 96] B. Manjunath and W. Ma. *Texture Features for Browsing and Retrieval of Image Data*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8):837–842, August 1996.

[Martin 08] C. Martin, U. Werner, and H. Gross. *A real-time facial expression recognition system based on Active Appearance Models using gray images and edge images*. 8th IEEE International Conference on Automatic Face & Gesture Recognition, pages 1–6, Sept. 2008.

[Matthews 04] I. Matthews and S. Baker. *Active Appearance Models Revisited*. In International Journal of Computer Vision, Volume 60, pages 135–164, 2004.

[McIntyre 06] G. McIntyre and R. Göcke. *Researching Emotions in Speech*. In 11th Australasian International Conference on Speech Science and Technology, pages 264–369, Auckland, New Zealand, December 2006. ASSTA.

[Mcintyre 07] G. Mcintyre and R. Göcke. *Towards Affective Sensing*. In Proceedings of the 12th International Conference on Human-Computer Interaction HCII2007, Volume 3 of *Lecture Notes in Computer Science LNCS 4552*, pages 411–420, Beijing, China, July 2007. Springer.

[McIntyre 08a] G. McIntyre and R. Göcke. *A Composite Framework for Affective Sensing*. In Proceedings of Interspeech 2008, pages 2767–2770. ISCA, 22-26 September 2008.

[McIntyre 08b] G. McIntyre and R. Göcke. Affect and Emotion in Human-Computer Interactions, chapter The Composite Sensing of Affect, pages 104–115. Lecture Notes in Computer Science LNCS 4868. Springer, August 2008.

[McIntyre 09] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear. *An Approach for Automatically Measuring Facial Activity in Depressed Subjects*. In 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, September 2009. DOI 10.1109/ACII.2009.5349593.

[Millar 04] J. B. Millar, M. Wagner, and R. Göcke. *Aspects of Speaking-Face Data Corpus Design Methodology*. In International Conference on Spoken Language Processing 2004, Volume II, pages 1157–1160, Jeju, Korea, October 2004.

[Mogg 05] K. Mogg and B. Bradley. *Attentional Bias in Generalized Anxiety Disorder Versus Depressive Disorder*. Cognitive Therapy and Research, 29:29–45, 2005.

[Monk 08] C. Monk, R. Klein, and E. Telzer et al. *Amygdala and Nucleus Accumbens Activation to Emotional Facial Expressions in Children and Adolescents at Risk for Major Depression*. American Journal of Psychiatry, 165(3):90–98, Jan 2008.

[Moore 08] E. Moore, M. Clements, J. Peifer, and L. Weisser. *Critical analysis of the impact of glottal features in the classification of clinical depression*

in speech. IEEE Transactions on Biomedical Engineering, 55(1):96–107, 2008.

[Movellan 08] J. Movellan. *Tutorial on Gabor Filters*. Tutorial paper `http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf`, 2008.

[MPEG-7] MPEG-7. *Multimedia Content Description Interface*. `http://www.darmstadt.gmd.de/mobile/MPEG7`. Last accessed 23 April 2010.

[MultiBoost 06] MultiBoost. *An open source multi-class AdaBoost learner*. `http://www.iro.umontreal.ca/~casagran/multiboost.html`, 2005-2006.

[Murray 93] I. Murray and L. Arnott. *Toward the simulation of emotion in synthetic speech*. Journal Acoustical Society of America, 93(2):1097–1108, 1993.

[Muscles] Facial Muscles. `http://www.csupomona.edu/~jlbath/LabPics/facial.htm`. Last accessed 30 Jun 2009.

[N. Medforda et al 05] N. Medforda et al. *Emotional memory: Separating content and context*. Psychiatry Research: Neuroimaging, 138:247–258, 2005.

[Navigli 03] R. Navigli, P. Velardi, and A. Gangemi. *Ontology Learning and Its Application to Automated Terminology Translation*. IEEE Intelligent Systems, pages 22–31, 2003.

[NBS] NBS. *Neurobehavioral Systems*. `http://www.neurobs.com/`. last accessed 14 April 2010.

[NIMH] NIMH. *National Institute of Mental Health*. `http://www.nimh.nih.gov/health/publications/anxiety-disorders/index.shtml`. Last accessed 10 april 2010.

[Nixon 01] M. Nixon and A. Aguado. *Feature Extraction and Image Processing*. MPG Books Lrd, Brodmin, Cornwall, 2001.

[Obrenovic 05] Z. Obrenovic, N. Garay, J. Lpez, I. Fajardo, and I. Cearreta. *An Ontology for Description of Emotional Cues*. Technical report, Laboratory for Multimodal Communications, University of Belgrade, 2005.

[Okada 09] T. Okada, T. Takiguchi, and Y. Ariki. *Pose robust and person independent facial expressions recognition using AAM selection*. IEEE 13th International Symposium on Consumer Electronics, pages 637–638, May 2009.

[OpenCV] OpenCV. *Image Processing Library Download*. `http://sourceforge.net/projects/opencvlibrary/files/`. Last accessed 10 November 2009.

[Pantic 00] Maja Pantic and Leon J.M. Rothkrantz. *Automatic Analysis of Facial Expressions: The State of the Art*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1424–1445, 2000.

[Pantic 04a] M. Pantic and L. Rothkrantz. *Case-based reasoning for user-profiled recognition of emotions from face images*. Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan, 2004.

[Pantic 04b] M. Pantic and L. Rothkrantz. *Facial action recognition for facial expression analysis from static face images*. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 34(3):1449–1461, 2004.

[Pantic 07] M. Pantic and M. Bartlett. Face recognition, chapter Machine Analysis of Facial Expressions, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, July 2007.

[Parker 02] G. Parker and K. Roy. *Examining the utility of a temperament model for modelling non-melancholic depression*. Acta Psychiatrica Scandinavica, 106(1):54–61, 2002.

[Parker 06] G. Parker, V. Manicavasagar, J. Crawford, L. TULLY, and G. Gladstone. *Assessing personality traits associated with depression: the utility of a tiered model*. Psychological Medicine, 36(8):1131–1139, August 2006.

[Picard 97] R.W. Picard. *Affective Computing*. MIT Press, Cambridge (MA), USA, 1997.

[Polydoros 06] P. Polydoros, C.Tsinaraki, and S. Christodoulakis. *GraphOnto: OWL-Based Ontology Management and Multimedia Annotation in the DS-MIRF Framework*. Technical report, Lab. Of Distributed Multimedia Information Systems, Technical University of Crete (MUSIC/TUC) University Campus, Kounoupidiana, Chania, Greece,, 2006.

[Pree 95] Wolfgang Pree. *Design Patterns for Object-Oriented Software Development*. Addison Wesley Longman, 1st edition, 1995.

[PRISM] PRISM. `http://research.cs.tamu.edu/prism/`. Last accessed on 26 February 2010.

[Protégé] Protégé. `http://protege.stanford.edu/`. last accessed, 22 April 2010.

[Qt 09] Qt. *A cross-platform application and UI framework*. `http://qt.nokia.com/`, November 2009. Last accessed 10 November 2009.

[Rahman 05] A. Rahman, I. Kiringa, and A. Saddik. *An Ontology for Unification of MPEG-7 Semantic Descriptions*. Technical report, School of Information Technology and Engineering Univeristy of Ottawa, Ontario, Canada, 2005.

[RapidMiner] RapidMiner. http://rapid-i.com/content/view/181/190/. last accessed 04 August 2010.

[Reed 07] L. Reed, M. Sayette, and J. Cohn. *Impact of depression on response to comedy: A dynamic facial coding analysis.* Journal of Abnormal Psychology, 117(4):804–809, May 2007.

[Rege 05] M. Rege, M. Dong, F. Fotouhi, M. Siadat, and L. Zamorano. *Using MPEG-7 to build a Human Brain Image Database for Image-guided Neurosurgery*. Medical Imaging 2005: Visualization, Image-Guided Procedures, and Display, pages 512–519, 2005.

[Renneberg 05] Babette Renneberg, Katrin Heyn, Rita Gebhard, and Silke Bachmann. *Facial expression of emotions in borderline personality disorder and depression*. Journal of Behavior Therapy and Experimental Psychiatry, 36(3):183–196, 2005.

[Ro 01] Yong Man Ro, Munchurl Kim, Ho Kyung Kang, and B. S. Manjunath. *MPEG-7 Homogeneous Texture Descriptor*. Electronics and Telecommunications Research Institute Journal, 23:41–51, 2001.

[Saatci 06] Y. Saatci and C. Town. *Cascaded classification of gender and facial expression using active appearance models*. In Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, pages 393–398, April 2006.

[Salembier 01] P. Salembier and J. Smith. *MPEG-7 Multimedia Description Schemes*. IEEE Transactions on Circuits and Systems for Video Technology, VOL. 11, NO. 6:748–759, 2001.

[Sander 05] D. Sander, D. Grandjean, and K. Scherer. *A systems approach to appraisal mechanisms in emotion*. Neural Networks, 18:317–352, 2005.

[Saragih 06] J. Saragih and R. Göcke. *Iterative Error Bound Minimisation for AAM Alignment*. Pattern Recognition, International Conference on, 2:1192–1195, 2006.

[Saragih 08] J. Saragih. *The Generative Learning and Discriminative Fitting of Linear Deformable Models*. PhD thesis, Research School of Information Sciences and Engineering, The Australian National University, Canberra, Australia, 2008.

[Saragih 09] J. Saragih and R. Göcke. *Learning AAM fitting through simulation*. Pattern Recognition, 42(11):2628–2636, 2009.

[Scherer 99] K. R. Scherer. Handbook of cognition and emotion, chapter Appraisal theory. New York: John Wiley, 1999.

[Scherer 03] K. R. Scherer. *Vocal communication of emotion: A review of research paradigms*. Speech Communication, 40:227–256, 2003.

[Scherer 04] K. R. Scherer. *HUMAINE Deliverable D3c: Preliminary plans for exemplars: theory. Retrieved 26 October, 2006 from.* `http://emotion-research.net/publicnews/d3c/`, 2004.

[Schröder 05] M. Schröder and R. Cowie. *HUMAINE project: Developing a Consistent View on Emotion-oriented Computing. Retrieved 26 October, 2006 from.* `http://emotion-research.net/aboutHUMAINE`, 2005.

[Schröder 07] M. Schröder, L. Devillers, K. Karpouzis, J. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson. *What Should a Generic Emotion Markup Language Be Able to Represent?* In Ana Paiva, Rui Prada,

and Rosalind W. Picard, editors, Affective Computing & Intelligent Interaction, Volume 4738 of *Lecture Notes in Computer Science*, pages 440–451. Springer, 2007.

[Sebe 03] N. Sebe and M. Lew. *Robust Computer Vision Theory and Applications*. Springer, 2003.

[Sebe 05] N. Sebe, I. Cohen, A. Garg, and Th. Huang. *Machine Learning in Computer Vision*. Springer, 2005.

[Shen 05] L. Shen. *Recognizing Faces — An Approach Based on Gabor Wavelets*. PhD thesis, School of Computer Science, University of Nottingham, 2005.

[Shen 06] L. Shen and L. Bai. *A review on Gabor wavelets for face recognition*. Pattern Analysis & Applications, 9(2-3):273–292, 2006.

[Shen 07] L. Shen, L. Bai, and Z. Ji. Advances in visual information systems, Volume 4781, chapter A SVM Face Recognition Method Based on Optimized Gabor Features, pages 165–174. Springer Berlin / Heidelberg, 2007.

[Shigeno 98] S. Shigeno. *Cultural Similarities and Differences in the Recognition of Audio-Visual Speech Stimuli*. In 5th International Conference on Spoken Language Processing, Volume 1057, pages 281–284. International Conference on Spoken Language Processing, 1998.

[Song 05] D. Song, H. Lie, M. Cho, H. Kim, and P. Kim. Image and video retrieval, Volume 3568/2005, chapter Domain Knowledge Ontology Building for Semantic Video Event Description, pages 267–275. Springer Berlin / Heidelberg, 2005.

[S.Strupp 08] S.Strupp, N. Schmitz, and K. Berns. *Visual-Based Emotion Detection for Natural Man-Machine Interaction*. In KI '08: Proceedings of the 31st

annual German conference on Advances in Artificial Intelligence, pages 356–363, Berlin, Heidelberg, 2008. Springer-Verlag.

[Stegmann 02] M. Stegmann and D. Gomez. *A Brief Introduction to Statistical Shape Analysis*. Technical report, University of Denmark, DTU, March 2002.

[Stibbard 01] R. Stibbard. *Vocal expression of emotions in non-laboratory speech: An investigation of the Reading/Leeds Emotion in Speech Project annotation data.* PhD thesis, University of Reading, UK, 2001.

[Sung 08] J. Sung and D. Kim. *Pose-Robust Facial Expression Recognition Using View-Based 2D 3D AAM*. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 38(4):852–866, July 2008.

[ten Bosch 00] L. ten Bosch. *Emotions: What is Possible in the ASR Framework*. SpeechEmotion, 2000.

[Tian 01] Y. Tian, T. Kanade, and J. Cohn. *Recognizing Action Units for Facial Expression Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2):97–115, Feb 2001.

[Tong 09] Y. Tong, X. Liu, and F. Wheeler P. Tu. *Automatic Facial Landmark Labeling with Minimal Supervision*. Computer Vision and Pattern Recognition, pages 2097–2104, June 2009.

[Tsai 01] D. Tsai, S. Wu, and M. Chen. *Optimal Gabor filter design for texture segmentation using stochastic optimization*. Image Vision Computing, 19(5):299–316, 2001.

[Tsinaraki 03] C. Tsinaraki, P. Polydoros, F. Kazasis, and S. Christodoulakis. *Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content*. Technical report, Lab. of Distributed Multimedia Information

Systems and Applications (MUSIC/TUC), Technical University of Crete Campus, 2003.

[Tsinaraki 07] C. Tsinaraki, P. Polydoros, and S. Christodoulakis. *Interoperability Support between MPEG-7/21 and OWL in DS-MIRF*. IEEE Transactions on Knowledge and Data Engineering, 19(2):219–232, 2007.

[Valstar 06a] M. Valstar and M. Pantic. *Biologically vs. Logic Inspired Encoding of Facial Actions and Emotions in Video*. In ICME, pages 325–328, 2006.

[Valstar 06b] M. Valstar and M. Pantic. *Fully Automatic Facial Action Unit Detection and Temporal Analysis*. In Conference on Computer Vision and Pattern Recognition Workshop, 2006.

[Vapnik 95] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

[Velten 68] E. Velten. *A laboratory task for induction of mood states*. Behaviour Research and Therapy, 6:473–482, 1968.

[Vinciarelli 08] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. *Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain*. In 16th ACM International Conference on Multimedia, pages 1061–1070, New York, NY, USA, 2008. ACM.

[Vinciarelli 09] A. Vinciarelli, M. Pantic, and H. Bourlard. *Social signal processing: Survey of an emerging domain*. Image Vision Computing, 27(12):1743–1759, 2009.

[Viola 01] P. Viola and M. Jones. *Robust real-time face detection*. Proceedings Eighth IEEE International Conference on Computer Vision ICCV 2001, 2:747, 2001.

[VXL] VXL. `http://vxl.sourceforge.net/`. Last accessed 10 November 2009.

[Wallhoff] F. Wallhoff. *Database with Facial Expressions and Emotions from Technical University of Munich (FEEDTUM)*. `http://www.mmk.ei.tum.de/~{}waf/fgnet/feedtum.html`. Last accessed on 26 February 2010.

[Wang 02] X. Wang and H. Qi. *Face Recognition Using Optimal Non-Orthogonal Wavelet Basis Evaluated by Information Complexity*. In 16th International Conference on Pattern Recognition, Volume 1, page 10164, 2002.

[Whissell 89] C. Whissell. *The Dictionary of Affect in Language*. In Emotion: Theory, Research and Experience, 1989.

[Whitehill 09] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. *Toward Practical Smile Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31:2106–2111, 2009.

[Wiskott 97] L. Wiskott, J. Fellous, N. Krger, and C. Malsburg. *Face Recognition by Elastic Bunch Graph Matching*. IEEE Trans. Pattern Analysis and Machine Intelligence, 19:775–779, 1997.

[Wu 04] B. Wu, H. Ai, and R. Liu. *Glasses Detection by Boosting Simple Wavelet Features*. International Conference on Pattern Recognition, 1:292–295, 2004.

[Yacoub 03] S. Yacoub, S. Simske, X.Lin, and J. Burns. *Recognition of Emotions in Interactive Voice Response Systems*. Technical report, HP Laboratories Palo Alto, 2003.

[Zhou 06] M. Zhou and H. Wei. *Face Verification Using GaborWavelets and AdaBoost*. In International Conference on Pattern Recognition, pages 404–407, 2006.

[Zhou 09] M. Zhou and H. Wei. *Facial Feature Extraction and Selection by Gabor Wavelets and Boosting*. In 2nd International Congress on Image and Signal Processing, pages 1–5, 2009.