Technical Report

UCAM-CL-TR-793
ISSN 1476-2986

Number 793

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Affect inference in learning environments: a functional view of facial affect analysis using naturalistic data

Shazia Afzal

December 2010

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted May 2010 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Murray Edwards College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This research takes an application-oriented stance on affective computing and addresses the problem of automatic affect inference within learning technologies. It draws from the growing understanding of the centrality of emotion in the learning process and the fact that, as yet, this crucial link is not addressed in the design of learning technologies. This dissertation specifically focuses on examining the utility of facial affect analysis to model the affective state of a learner in a one-on-one learning setting.

Although facial affect analysis using posed or acted data has been studied in great detail for a couple of decades now, research using naturalistic data is still a challenging problem. The challenges are derived from the complexity in conceptualising affect, the methodological and technical difficulties in measuring it, and the emergent ethical concerns in realising automatic affect inference by computers. However, as the context of this research is derived from, and relates to, a real-world application environment, it is based entirely on naturalistic data. The whole pipeline - of identifying the requirements, to collection of data, to the development of an annotation protocol, to labelling of data, and the eventual analyses – both quantitative and qualitative; is described in this dissertation. In effect, a framework for conducting research using natural data is set out and the challenges encountered at each stage identified.

Apart from the challenges associated with the perception and measurement of affect, this research emphasises that there are additional issues that require due consideration by virtue of the application context. As such, in light of the discussed observations and results, this research concludes that we need to understand the nature and expression of emotion in the context of technology use, and pursue creative exploration of what is perhaps a qualitatively different form of emotion expression and communication.

Acknowledgements

This has been a long journey but perhaps the most transformative and rewarding experiences of my life. It certainly would not have been possible without the love, inspiration and support that I received from numerous people around me. I am indebted to my supervisor, Peter Robinson, for believing in me and providing me with this wonderful opportunity. I thank him for his guidance and encouragement in all these years and remain deeply obliged to him for helping me grow as a researcher. I thank Alan Blackwell for involving me in several interesting discussions and his critical questioning that urged me to remain focussed on the practicalities. I also thank David Good for his helpful insights during experiment design and Roddy Cowie for his valuable comments on this dissertation.

I thank my dear friends, Aisha, Aurelie, Beenish, Catherine, Cecily, Ilaria, Imran, Laurel, Munazah, Nicole, Qurat, Ramla, and Ubaier, as well as members of the Rainbow Group – Daniel, Ian, Laurel, Luke, Metin, Phil, Pradipta, and Tal, for their excellent company and most cherished conversations. Thank you: Catherine, Ilaria, Nicole and Ramla for being the most wonderful housemates I could wish for; Cecily, for being the most valuable companion especially in the final year; Ian and Phil, for being such pleasant officemates; Ubaier, for hearing me out for hours; and Imran, for your patient wait and unfailing support.

I thank the staff at the Computer Laboratory, Murray Edwards College, Betty and Gordon Moore Library, and the University Library; especially, Lise, Megan, Fiona, Michelle, Sian, Graham and Nicholas, for their help and assistance on more occasions than I can remember.

This work would not have been possible without the gracious funding by the Gates Cambridge Trust, the Overseas Research Studentship, and numerous other awards from the Murray Edwards College, the Cambridge Philosophical Society and of course, the Computer Laboratory. I am also grateful to them for their financial support towards my several conference and workshop attendances where I could present and share my work with the research community. I thank the numerous reviewers and co-researchers whose feedback has tremendously helped in this final presentation of my research.

Finally, this dissertation is a tribute to the efforts of my parents for providing me with opportunities of education and growth that few can imagine in the conflict-stricken valley of Kashmir. I thank my Dad for a disciplined upbringing and his never-ending insistence on learning and accomplishment; my Mum for her love, strength and prayers; my sisters, Lubna and Saniya, for their support and care; my nieces, Aaidah and Aairah, and my nephews, Musa and Eesa, for being the final motivation to finish up.

To my Parents, With Love.

Contents

List of figures

List of tables

1. Introduction

This research falls within the domain of Affective Computing (Picard, 1997) which aims to represent, detect and analyse nonverbal behaviour in an attempt to model affective phenomena in human-computer interactions (HCI). It has a special relevance in applications where computers take on a social role like an instructor, a helper or a companion. This research investigates one such aspect of affective behaviour in the context of computer-assisted learning environments. This chapter outlines the specific research aims and identifies the associated challenges.

1.1 Motivation

Computer-based learning now encompasses a wide array of innovative learning technologies ranging from adaptive hypermedia systems to sophisticated tutoring environments, educational games, virtual environments or just simply online tutorials. These continue to enrich the learning process in numerous ways. Keen to emulate the effectiveness of human tutors in the design and functioning of learning technologies, researchers have continually looked at the strategies of expert human teachers for motivation and are making directed efforts to make this machine-learner interaction more natural and instinctive. Since learning with computers is essentially self-paced, assessing the learners' experience becomes important. Detection of learners' affect states can give better insight into a learners' overall experience which can be helpful in adapting the tutorial interaction and strategy. Such a responsive interface can also alleviate fears of isolation in learners and facilitate learning at an optimal level. To enhance the motivational quality and engagement value of instructional content, affect recognition needs to be considered in light of its implications to learning technologies.

Effective tutoring by humans is an interactive yet guided process where learner engagement is constantly monitored to provide remedial feedback and to maximise the motivation to learn (Merill, Reiser, Trafton, & Ranney, 1992). Indeed, formative assessment and feedback is an important aspect of effectively designed learning environments and should occur continuously and unobtrusively, as part of the instruction (Bransford, Brown, & Cocking, 1999). In naturalistic settings, the availability of several channels of communication facilitates the constant monitoring necessary for such an interactive and flexible learning experience (Picard, et al., 2004; de Vicente & Pain, 1998). One of the biggest challenges for computer tutors then is to achieve the mentoring capability of expert human teachers (van Vuuren, 2006). To give such a capability to a machine tutor entails giving it the ability to infer affect.

Learning has a strong affective quality that impacts overall performance, memory, attention, decision-making and attitude. Recent research provides compelling evidence to support the multiplicity and functional relevance of emotions for the situational and ontogenetic development of learners' interest, motivation, volition, and effort (Pekrun, 2005). It reflects the growing understanding of the centrality of emotion in the teaching-learning process and the fact that as yet this crucial link has not been addressed in machine-learner interactions (O'Regan, 2003).

Despite this recognition of affect as a vital component of learning processes and a context for cognition to occur, computer-based learning environments have long ignored this aspect and have concentrated mostly in modelling the behaviour of a learner in response to a particular instructional strategy (Picard, et al., 2004; du Boulay & Luckin, 2001). This relative bias towards the cognitive dimension of learning is now being criticised and the inextricable linkage between affective and cognitive functions is being stressed. This comes at a time when advances in the field of affective computing have opened the possibility of envisioning integrated architectures by allowing for formal representation, detection, and analysis of affective phenomena. This increasing interest in building affect sensitive human-computer interactions thus finds an important application in learning technologies (Cowie, et al., 2001).

1.2 Aims and potential challenges

In the learning context, automatic measurement of affect has been approached as a three stage problem: understanding affect in learning, developing reliable sensing techniques, and finally, intelligent adaptation (Craig, Graesser, Sullins, & Gholson, 2004; Kort, Reilly, & Picard, 2001). This conceptual division reduces the complexity of an overlapping and wide domain into specific focus areas. All the three areas look at traditional learning theories and natural teacher-learner interactions for inspiration and insight.

The first phase involves exploring underlying theories and models of affect in learning to derive some sort of emotion subset or taxonomy to start with. It requires understanding what we should aim to detect in machine-learner interactions and why. In other words, what are the significant affective phenomena linked to learning and how these should be conceptualised in computer-based learning environments. The second stage looks at how these affect states are manifested and involves development of appropriate sensing techniques for their reliable detection. The last stage is concerned with how this information about the learners' affect state can be utilised in context of the knowledge state for appropriate feedback or diagnosis, altering the pace of learning or adapting a tutoring strategy. The desirable stage of intelligent adaptation relies on satisfactory progress in the initial two phases. To develop an understanding of the nature of data and the limitations in application of affect-sensing technology, an emphasis on the practicalities is therefore necessary. This research aims to contribute towards this understanding.

The transition from our intuitive understanding of affect perception to formalising it in a grammar for computational models, presents itself with a number of challenging issues. In order to contextualise the research objectives, these are briefly introduced below. A more detailed and thorough discussion appears in several key publications like Peter and Herbon (2006), Cowie, Douglas-Cowie, and Cox (2005), Porayska-Pomsta and Pain (2004), Pantic (2003), Cowie et al. (2001), and Picard (1997), to name a few.

Definitional Issues

The study of emotion has an extensive and diverse literature ranging in perspectives - evolutionary, behaviourist, componential, socio-cultural and also neuro-scientific approaches. An apt indicator of the terminological confusion itself is the compilation of 92 emotion definitions and 9 statements by Kleinginna and Kleinginna way back in 1981. This complexity in the general understanding of what emotion means is a major impediment to researchers investigating affective user-interfaces. Furthermore, the absence of a theory or model of affect in learning, at a level of detail that is amenable to implementation magnifies this difficulty (Porayska-Pomsta & Pain, 2004).

It has been recommended that researchers adopt a working definition of emotion in order to plan, communicate and identify the scope of a project (Larsen & Fredrickson, 1999). Assuming such a working definition allows one to construct a framework to ask experimental questions, design methodologies and interpret results without getting embroiled in a psychological debate over how emotion is defined. The model definition proposed by Kleinginna and Kleinginna (1981) following their extensive review, is relevant for such a purpose and is reproduced here:

“Emotion is a complex set of interactions amongst subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labelling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behaviour that is often, but not always, expressive, goal-directed, and adaptive” (Kleinginna & Kleinginna, 1981, p. 355).

This captures two important and consensual aspects of emotion: one, as reaction to events deemed relevant to the needs, goals or concerns of an individual; and two, encompassing physiological, affective, behavioural and cognitive components (Brave & Nass, 2002).

Descriptive Issues

Descriptive issues are derived from the choice of conceptualising affective content within an experimental framework and can be broadly classified along three main representation schemes: categorical, continuous and appraisal-based (Cowie, Douglas-Cowie, & Cox, 2005).

Categorical schemes conform to the everyday use of language terms for verbalising emotional experiences and are naturally suited for their descriptive ease and familiarity. There is an enormous variety of emotion words ranging from the six basic emotions - happiness, sadness, fear, anger, disgust and surprise (Ekman & Freisen, 1971), to more elaborate lexical taxonomies that allow representation of more complex affect states beyond the prototypical ones (Baron-Cohen, Golan, Wheelwright, & Hill, 2004). However, the range of possible descriptors together with issues about their cross-cultural compatibility and meaning can be problematic (Peter & Herbon, 2006; Wierzbicka, 2003).

An alternative to categorical description is based on the dimensional view of emotions whereby an affect state is represented as a point in a space of two or more dimensions defined by psychological concepts. One of the most popular models is Russell's Circumplex Model of Affect (Russell, 1980) which posits that emotions conform to a circular or radix arrangement with the coordinates of this circular space defined along two orthogonal dimensions of arousal (activation) and valence (pleasantness). The primary advantage of dimensional models is the ease in charting an emotional experience without explicit articulation of a specific emotion. However, collapsing an emotional state to dimensional constructs inevitably leads to a loss of discriminative information and makes some emotions like fear, anger and disgust indistinguishable.

Appraisal based schemes like Scherer's Component Process Model (2005) and Ortony, Clore, & Collins' (1988) Cognitive Theory of Emotion, provide a powerful predictive framework by specifying emotions as valenced reactions to critical events/objects arising out of an appraisal or evaluation of the situation. Although appraisal based methods enable a cause-effect reasoning of emotions and are attractive as an operational model, only a few affect recognition systems have adopted this method due to the level of detail required as well as the reliance on subjective appraisal accounts. On the other hand, although the measurement issue between categorical and dimensional views has been a subject of discussion for over a hundred years (Larsen & Fredrickson, 1999; Izard, 1993; Lazarus, 1991; Darwin, 1872), both have been adopted with considerable success in numerous affect recognition systems.

In all, the choice of a description strategy is an important one that effectively determines the design and functionality of an affect-sensitive system. Defining a representation scheme implies characterising the construct to be measured, and therefore, the output of the system itself. Given its importance, the issue of representation is further elaborated in Section 1.2.1.

Methodological Issues

The methodological issues in automatic affect inference arise from the sensitivity of emotion to the form and type of measurement method and are related to the validity, accuracy, timing and context of emotion assessment (Larsen & Fredrickson, 1999). Since emotions are dynamic processes that unfold over time, measures obtained with minimum latency or better still, concurrently during the emotional experience, maximise the chances of validity and accuracy. In order to ensure ecological validity of a measurement technique, the relevant

contextual factors also need to be taken into account. Moreover, establishing the reliability of emotion assessment often requires the use of trained experts which can be costly and time consuming.

The recent emphasis on the use of naturalistic databases, as against the posed/acted ones, for the real-world application of affective systems, reflects all these concerns. Naturalism is, however, at odds with the signal processing quality required for efficient and robust analysis by computers (Cowie, Douglas-Cowie, & Cox, 2005). In general, identifying an emotional episode in an ongoing interaction is a complex task that is intricately related to the representation scheme used, type and level of emotion measurement as well as the empirical agenda.

Technical Issues

Emotions or affect states are multi-dimensional constructs manifested across the visual, auditory and physiological channels, often at varying time scales. As such, the robust tracking of behavioural signals associated with changes across these modalities is an active problem in computer vision research. For example, issues related to pose, scale, resolution, lighting and occlusions are still relevant in the analysis of visual input. Independence of a technique from users' physiognomies, gender, age and ethnicity; sensitiveness to temporal dynamics; fusion of information from multiple channels as well as context-sensitive interpretation are challenging factors to be considered (Pantic & Rothkrantz, Toward an Affect-Sensitive Multimodal Human-Computer Interaction, 2003). A further concern is the training and validation of automatic affect analysers across the diverse experimental conditions, databases and annotation protocols used. While it is generally agreed that a reliable assessment of affect state requires concurrent use of multiple channels, the suitability of a modality or combinations of more, depends on the application requirements, types of emotions to be detected, technical feasibility, ethical concerns and real-time requirements (Hudlicka, 2003).

Ethical Issues

Ethical issues in automatic affect sensing are concerned with user awareness and control over affect monitoring, the use of non-invasive sensors, reliability of results, and issues of comfort and privacy (Picard & Klein, 2002). Amidst all these critical concerns is the broad question of how to negotiate a balance between automatic adaptation versus what could be perceived as manipulation or a feeling of 'surreptitiousness' in users (Brave & Nass, 2002).

1.2.1 Integrated discussion on representational issues

The issue of representation is at the core of emotion research and therefore affective computing. For practical use in technological contexts, it is necessary to understand how emotion is expressed in everyday interactions, to be able to suitably represent it as well as to enable inter-operability. This is because handling of emotion data by machines requires programmed representations of affect and a clear structure that will perform real-time

interaction with a user. Selection of an appropriate descriptive framework embodies the way affect is conceptualised within a system, the way it is observed and assessed, and consequently, the way it is processed (Peter & Herbon, 2006). However, the question of representation is not a simple one as it requires an understanding of the typology and semantics of the whole range of emotion-related phenomena like short-lived, intense emotions; moods; long-lasting established emotions; stances; attitudes/preferences, traits/affect dispositions, etc (Cowie & Cornelius, 2003). These differ in terms of duration, intensity as well as frequency of occurrence and can be crucial in providing the necessary context for meaningful interpretation of behaviour (Cowie, 2009; Larsen & Fredrickson, 1999). For emotional intelligence in the true sense, for example, one should be able to judge if an emotion reaction was based on an instantaneous situation or due to the underlying mood or that based on the affect disposition of a user. Of further concern are issues derived from the very nature of naturally occurring emotional behaviour like co-occurrence of emotions, regulation effects like simulation and attenuation, cultural and inter-personal differences, and the inherent ambiguity in signs of emotions. All this complicates the task of describing emotional content and while no single best representation scheme exists, there are established psychological traditions that have been used effectively to formalise the behaviour of interest.

The most long-standing way by which affect has been described by psychologists is in terms of discrete categories – an approach rooted in everyday language and driven by historical tradition around the existence of universal emotions. The most popular example of this description is the list of six prototypical emotion categories by Ekman, with claims of their cross-cultural understanding and recognition. The main advantage of the categorical scheme is that people use it to describe emotional displays in everyday interactions and is therefore intuitive. However, assignment of emotions into discrete categories or words is often considered arbitrary because of the social and cultural differences in semantic descriptions of emotion and for a designer of an HCI system, the requirement of an exclusive unambiguous representation. Linguistic labels can be imprecise and capture only a specific aspect of the phenomena with an associated uncertainty in the perceived meaning of a category. Nevertheless, this approach has had a dominating influence on the field of affective computing and most of the existing systems focus on recognising a list of basic emotions. Traditional psychological lists of emotions are mostly oriented to archetypal emotions and these are not the states that appear in most naturalistic data, especially in HCI contexts. As such, they do not represent the full range of emotions that can occur in natural communication settings. To overcome the intractable number of emotion terms and to ensure relevance in potential applications, the strategy of preselecting context-relevant word lists or cumulating relevant categories to derive pragmatic lists as in the HUMAINE database (Douglas-Cowie, et al., 2007) or the more principled taxonomy of complex mental states by Baron-Cohen (2004), has been advocated and applied effectively (Cowie, 2009; Zeng et al. 2009).

An alternative representation is offered by the psychological concept of dimensions. According to this view, emotions can be characterised and differentiated in terms of two or more cardinal dimensions to get a simplified representation of the affective space. Although the exact number and nature of these dimensions continues to be a subject of disagreement, there is some consensus on an underlying two-dimensional affective space charted by the orthogonal dimensions of activation and valence (Russell 1980; Watson and Tellegen 1985). Dimensional representation allows description of emotional states in a more tractable manner but is by definition an approximate tool. It collapses the structured, high-dimensional space of possible emotional states into a homogenous space of minimal dimensions. This inevitably results in information loss and inconsistent formulation in the event of different ways of achieving the collapse (Cowie & Cornelius, 2003). It cannot distinguish between all emotions, for instance, fear and anger require an additional dimension involving power or control to be distinguished, while politeness and interest lie outside its scope. In terms of assessment, this representation is not intuitive and raters need special training to use dimensional labelling systems.

In search for an optimal dimensional framework for structuring emotions, most early researchers have suggested at least three dimensions, commonly evaluation-pleasantness, potency-control, and activation-arousal (e.g. Osgood, May & Miron, 1975). Some argue that pleasure and arousal constitute major orthogonal dimensions (e.g. Russell 2003), while others advocate positive and negative affect as dimensions oriented at 45 degrees to pleasure and avoidance behaviours, respectively (Westerman, Gardner, & Sutherland, 2006). More recently, Fontaine et al. (2007)propose a fourth dimension of unpredictability to satisfactorily represent the semantic space of emotions. They argue that simple two-dimensional models, such as the valence-arousal model, miss major sources of variation in the emotion space and advocate using at least four dimensions to get a fairly comprehensive account of the emotional experience. So while the level of abstractness and simplicity offered by the dimensional scheme is fairly attractive, the choice of relevant dimensions depends on the emotion distinctions sought in a specific research view and the impact that this choice is likely to make in the way affect is handled in a system.

One of the most influential approaches in recent times is based on the appraisal theory which is based on the notion that emotional processes are elicited and dynamically patterned as the individual continuously and recursively appraises objects, behaviours, events, and situations with respect to their effect on his/her values, goals, and general well-being (Sander, Grandjean, & Scherer, 2005). In simple terms, it propounds that distinct types of emotions correspond to distinct ways of appraising the situation that evokes the emotion. As such, appraisal theories like Scherer's Component Process Model (2005), offer a fine theoretical tool to explain both the elicitation and multimodal reaction patterning as well as explain the high degree of qualitative differentiation of emotional experience. Consequently, these offer a powerful logical framework for emotion representation where an emotional state is described through a set of stimulus evaluations checks including novelty, intrinsic pleasantness, goal significance, coping potential and compatibility. While this makes it

attractive as an operational model for computational purposes, the high degree of specificity involved in appraisal judgements makes it extremely challenging and tedious to translate this scheme into an engineering framework for the purposes of automatic emotion recognition (Zeng et al. 2009).

Each of the representations offers a distinctive abstraction level and eventually the choice of an appropriate description depends on the theoretical and empirical agenda of a research. From a computational perspective, it does not imply choosing one theory over another but rather a practical evaluation of the proposed application and a careful assessment of the effect of these choices on the design and functioning of a system. The W3C incubator group addresses these issues from a technological standpoint and seeks to develop a standard motion markup language known as EmotionML (Schröder, 2008), designed to be usable in a broad range of technological contexts while reflecting concepts from the affective sciences. A markup language is a general term for a formalised system used to annotate descriptions of events and in this context, useful to annotate emotional content and structure of natural corpora. Emotion-related mark-up languages can be defined at several abstraction levels. At the highest level, they may denote the mental state of the message producers, in terms of their beliefs, desires, intentions and affective state. Going down in this hierarchy, they may denote message content: its communicative goal, focus, relation with other parts of the message, and so on. At the lowest level, tokens provide a symbolic representation of the surface structure of the message: linguistic features, facial expressions, gestures, speech characteristics and other forms that people perceive as the means by which other people express the mental state of the sender and the message content (Douglas-Cowie, 2004). The idea behind EmotionML is not to standardise emotions or to unify emotion theories but instead to transfer and make available the descriptions of emotion-related states in application-oriented technological contexts.

1.2.2 Research agenda

In view of the preceding discussion, this work is an attempt to discover the feasibility of automatic affect inference in computer-based learning environments and to explore its potential in alleviating the gap in realising effective affect-sensitive interaction. The objective is to explore methods that can reasonably approximate the human perceptual skills of affect inference by restricting the bandwidth of communication to what is practically available in an average computer-based learning environment.

The practical contributions of this research are aligned along the main objectives outlined as follows:

- To review the problem space and develop an understanding of the state of art.
- To identify an appropriate framework for conceptualising affect in the context of affect-sensitive learning systems.
- To consider the key issues that may influence the selection and application of emotion measurement techniques in the target scenario.

- To compile a corpus of representative emotional behaviour in the relevant context.
- To explore the performance of different classification methods on the naturalistic data collected and to characterise the selected affect states; and
- Finally, to assess the most practical way of modelling affect in light of the results obtained and implications envisaged.

In order to study the dynamics of machine-learner interaction as closely and naturally as possible, this work is based on detailed observational and machine analysis of empirically collected data.

1.3 Dissertation outline

This dissertation is structured as follows:

Chapter 2: Background presents a discussion of the problem and why it is important. It surveys related work in the area to give an idea of the state-of-art and also defines the scope of this dissertation.

Chapter 3: Representative Data describes the data collection exercise and the annotation procedures in detail. It also provides a comparison of the collected database against existing visual databases.

*Chapter 4: Facial Affect Analysis e*valuates the performance of different classification algorithms and strategies on the collected natural corpus. Both supervised and unsupervised methods are used to uncover the structure of affect states.

Chapter 5: Emotional Information in Facial Feature Points reports results from an experiment analysing the emotional information encoded in facial feature points using state-of-art automatic facial feature tracking.

Chapter 6: Conclusions presents the synopsis of the research highlighting the main contributions while setting out directions for future research, in the light of more qualitative observations from the data.

Throughout this dissertation the terms emotion, affect and mental states are used interchangeably, as are the differing terms for computer-based learning environments like computer-assisted learning, intelligent tutoring systems, computer tutors, learning technologies, etc.

1.4 Publications

Excerpts from this research have appeared in the following peer-reviewed publications:

1. Afzal, S., Robinson, P. (2010). Measuring Affect in Learning - Motivation and Methods, 10th IEEE International Conference on Advanced Learning Technologies (ICALT), Tunisia.

(Best Paper Award)

2. Afzal, S., Robinson, P. (2009). Natural Affect Data - Collection & Annotation in a Learning Context, In Proceedings of Affective Computing & Intelligent Interaction (ACII), Amsterdam.
3. Afzal, S., Sezgin, T.M., Gao, G. & Robinson, P. (2009). Perception of Emotional Expressions in Different Representations Using Facial Feature Points, In Proceedings of Affective Computing & Intelligent Interaction (ACII), Amsterdam.
4. Afzal, S., Morrison, C. & Robinson, P. (2009). Intentional affect: An alternative notion of affective interaction with a machine, In Proceedings of British HCI, Cambridge, UK.
5. Riek, L., Afzal, S. & Robinson, P. (2009). Do Affect-Sensitive Machines Influence User Behavior?, In Proceedings of the AISB Symposium on The Social Understanding of Artificial Intelligence (SSoAI), Edinburgh, UK.
6. Afzal S., Robinson P. (2008). Dispositional Expressivity and HCI, In Proceedings of Workshop on Emotion In HCI, British HCI, Liverpool, UK.
7. Riek, L., Afzal S., & Robinson P. (2008). Affect Decoding Measures and Human-Computer Interaction, In Proceedings of Measuring Behaviour (MB), Maastricht.
8. Afzal S., Robinson P. (2008). An Interface to Simplify Annotation of Emotional Behaviour, In Proceedings of HUMAINE Workshop on Corpora in Emotion Research (LREC), Morocco.
9. Afzal S. (2007). Evaluating Learner State from Affective Cues, In Proceedings of Doctoral Consortium of the Intl. Conf. on Affective Computing and Intelligent Interaction (ACII), Lisbon.
10. Afzal S., Robinson P. (2007). A Study of Affect in Intelligent Tutoring, In Proceedings of Workshop on Modelling and Scaffolding Affective Experiences to Impact Learning, Intl. Conf. on Artificial Intelligence in Education (AIED), Los Angeles.

2. Background

Emotions are crucial for healthy cognitive functioning and have direct relevance to learning and achievement. Not surprisingly then, affective diagnoses constitute a significant aspect of expert human mentoring. Computer-based learning environments aim to model such social dynamics to make learning with computers more immersive, engaging and hence more effective. This chapter highlights the surge of interest in studying emotions in learning, introduces techniques for accessing emotions, and surveys recent efforts to automatically measure emotional experience in learning environments. It attempts to bring together the motivation, methodological issues, and modelling approaches for affect inference in the learning context in order to contribute to an understanding of the research objectives and the current state-of-art. It identifies issues and describes the problem space to form the foundation of this dissertation and the work undertaken.

2.1 Learning and emotions

The neurobiology of emotions suggests that not only are learning, attention, memory, decision-making and social functioning affected by emotional processes but also that our repertoire of behavioural and cognitive options has an emotional basis. This relationship underscores the importance of the ability to perceive and incorporate social feedback in learning (Immordino-Yang & Damasio, 2007). Indeed, recent evidence from educational research supports the relationship of emotion with cognitive, motivational and behavioural processes (Pekrun, 2005; Turner, Husman, & Schallert, 2002). The seminal works of Boekaerts (2003), Pekrun, Goetz, Titz, and Perry (2002) and Meyer and Turner (2002) have pioneered the renewed surge of interest in affect and learning.

In a series of qualitative case-studies, Pekrun et al. (2002) explored the 'occurrence and phenomenological structures of academic emotions'. They demonstrated that learners experience a rich diversity of positive and negative emotions; the most frequently reported being: anxiety, enjoyment, hope, pride, and relief, as well as anger, boredom and shame. Developing a multidimensional instrument, the Academic Emotions Questionnaire [AEQ], they conducted quantitative studies to test assumptions underlying Pekrun's cognitive-motivational model (Pekrun, 1992). Using dimensions of valence (positive vs. negative) and activation (activating vs. deactivating) they distinguished four groups of emotions with reference to their performance effects – positive activating emotions (such as enjoyment of learning, hope, or pride); positive deactivating emotions (e.g., relief, relaxation after success, contentment); negative activating emotions (such as anger, anxiety, and shame); and

negative deactivating emotions (e.g., boredom, hopelessness). Accordingly, they studied the effects of these emotions on learning and achievement with cognitive and motivational mechanisms like motivation to learn, strategies of learning, cognitive resources, and self-regulation. Instances of these mechanisms like interest and effort, learning strategies like elaboration or rehearsal, task irrelevant thinking diverting cognitive resources and self-regulated learning as compared to reliance on external guidance may all occur in the course of learning with a computer tutor and are thus directly relevant to this study.

To evaluate the dynamic and interactive effects of affect and motivation on learning processes like task engagement and appraisal, Boekaerts (2003) conducted several longitudinal studies using the On-line Motivation Questionnaire (Boekaerts, 2002) and found evidence for the existence as well as relevance of two separate, parallel processing pathways – *the cold cognition pathway* and *the hot cognition pathway*. The cold cognition pathway consists of meaning-generating processes that are the building blocks of learning comprehension and problem-solving. The hot cognition pathway on the other hand comprises of the emotional evaluations of learning opportunities that are triggered by emotions and moods in the actual learning episode. In her Model of Adaptive Learning (Boekaerts, 1992), these represent the mastery and the well-being path respectively. Boekaerts asserts that the evaluative information of the hot cognition path is situation specific and initiates concern-related monitoring, thereby influencing both decision-making (short-term effect) as well as value attribution (long-term effect).

Based on a decade of research on motivation and a diverse study of learner-teacher interactions, Meyer and Turner (2002) highlight the inseparability of emotion, motivation and cognition; and argue for integrated approaches to treat these as equal components in the social process of learning. They report their findings as *serendipitous*, thus emphasising the presence of emotion in instructional interactions. Although the context of their research is classroom based, they provide a reflective account on the obvious nature of emotion in learning interaction.

Kort, Reilly and Picard (2001) highlight the importance of continuous affect monitoring as a critical mentoring skill. They propose a spiral model that combines the phases of learning to emotion axes by charting out quadrants that map different stages occurring in the learning process. The horizontal emotion axes range from negative to positive across different emotion sets like anxiety-confidence, boredom-fascination, frustration-euphoria, dispirited-encouraged and terror-enchantment. The vertical axis forms the learning axis that represents the transition between *constructive learning* and *un-learning*. This model assumes that the learning experience involves a range of emotions in the space of the learning task and visualises the movement of a learner from one quadrant to another. Figure 2.1-1 illustrates this model.

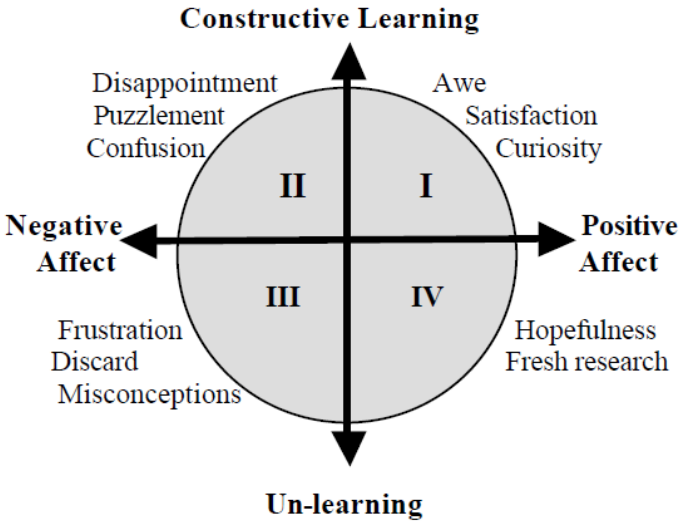


Figure 2.1-1: Model proposed by Kort, Reilly and Picard (2001) to relate phases of learning with emotions

In an attempt to understand the emotional dimension of online learning in qualitative terms, O'Regan (2003) explored the *lived experience* of students taking online learning courses. The study identifies both positive and negative emotions experienced by students, significantly - frustration, fear/anxiety, shame/embarrassment, enthusiasm/excitement and pride. These had a variable effect on the learning process depending on the strength and nature of the emotion, as well as the associated learning context.

In another study, using a manual affect coding system, Craig, Graesser, Sullins, and Gholson (2004) observed the occurrence of six affect states during learning with an intelligent tutoring system. They analysed frustration, boredom, flow, confusion, *eureka* and neutral, and found significant relationships between learning and the affective states of boredom, flow and confusion.

More recently, Jarvenoja and Jarvela (2005) and Wosnitza and Volet (2005) provide empirical evidence from participants in social online learning to categorise sources of emotional experience along self, task, context or social directedness to highlight the impact of students' emotions on their motivation and engagement in the learning process.

In essence, learning has a strong affective quality that impacts overall performance, memory, attention, decision-making and attitude (Kort, Reilly, & Picard, 2001; Lisetti & Schiano, 2000). We know from a multitude of studies in different educational contexts that learners experience a wide range of positive and negative emotions. These emotions are situated and have social and instructional antecedents. For the discourse to be effective, it is imperative then to have access to and ensure the emotional well-being of learners. Since learning with computers is essentially self-paced, assessing the learner's experience becomes important. The aim is to reasonably emulate the social dynamics of human teacher-learner interactions in models that capture the essence of effective learning strategies like one to one tutoring (Bloom, 1984; van Vuuren, 2006).

2.2 Measuring emotions

Current methods for measuring emotions can be broadly categorised as Subjective/Objective and Qualitative/Quantitative. In the context of learning, an additional categorisation as Snapshot/Continuous can be defined based on the timing of the emotion measurement (Wosnitza & Volet, 2005). Snapshot type measurements are done immediately before/after the learning process while continuous measurements are process-oriented and give access to the ongoing emotional experience. Consequently, snapshot measures provide only a limited window into the anticipated or reflected emotions at the end of the learning experience as against the continuous measures that provide direct access to emotions as they unfold during learning. Table 2.2-1 categorises some common methods for measuring emotional experience during learning.

Table 2.2-1: Methods for measuring emotional experience during learning

	Snapshot Type *(Before / After Learning)*		**Continuous Type** *(During Learning)*	
	Qualitative	**Quantitative**	**Qualitative**	**Quantitative**
Subjective	Open Interviews Emotional Probes Stimulated Recall	Questionnaires Surveys	Emotional Diaries Think-aloud	Experience / Time-Sampling
Objective	Structured Interviews	Transcripts Analysis Video Analysis	Observational Analysis	Interactional Content Physiology / Nonverbal Behaviour Analysis

For intervention to be effective, remedial action has to be immediate - particularly in the case of strong emotions. Given the complex and transient nature of emotions, any retrospective accounts are problematic because of issues related to the potential for multiple levels of awareness, reappraisals and reconstruction of meanings during recall (Schutz, Hong, Cross, & Obson, 2006). This necessitates dynamic evaluation of emotions but without disrupting the learning task itself. Ideally then, an unobtrusive, quantitative, and continuous account of emotional experience is a suitable method of enquiry. Amongst the methods listed in Table 2.2-1, analysis of nonverbal behaviour in the lower right quadrant, offers a reasonable fit to this requirement (Pekrun, 2005; Picard, et al., 2004; Hudlicka, 2003).

Analyses of tutoring sessions have indeed revealed that affective diagnoses, as an important aspect of expert human mentoring, depend heavily on inferences drawn from facial expressions, body language, intonation, and paralinguistic cues (Lepper, Woolverton, Mumme, & Gurtner, 1993). Advances in the field of affective computing have opened the possibility of emotion recognition from its nonverbal manifestations like facial expressions, head pose, body gestures, voice and physiology. The field is promising, yet in a formative stage as current technologies need to be validated for reliability outside controlled experimental conditions.

2.2.1 Automatic measurement of affect

The semantics and manifestation of affective phenomena have been extensively studied across the disciplines of psychology, cognitive science, computer vision, physiology, behavioural psychology, etc. In spite of this, it still remains a challenging task to develop reliable affect recognition technologies. The reasons are varied. Expression and measurement of affect, and specifically its interpretation, is person, time and context dependent. Sensory data is ambiguous and incomplete as there are no clear criterions to map observations onto specific affect states. Lack of such ground-truths makes validation of developed techniques difficult and worse still, application-specific. Consequently, we do not know whether a system that achieves higher classification accuracy than another is actually better in practice (Pantic & Rothkrantz, Toward an Affect-Sensitive Multimodal Human-Computer Interaction, 2003). Affect modelling in real-time is thus a challenging task given the complexity of emotions, their personal and subjective nature, the variability of their expression across, and even within, individuals, and frequently, lack of sufficient differentiation among associated visible and measurable signals (Hudlicka, 2003).

However, despite the difficulties, a whole body of research is persevering to give computers at least as much ability as humans have in recognising and interpreting affective phenomena that enables them to carry out intelligent behaviour and dialogue with others. This optimistic vision has already produced some commendable results and the following section reviews how machine perception of affect is being realised within learning environments. The interested reader is referred to Zeng, Pantic, Roisman, and Huang (2009) for a survey of general affect recognition methods using audio-visual modalities.

2.2.2 Previous work

Despite the prospects, there are relatively few studies on automatic affect sensing in learning environments. Table 2.2-2 compares these in chronological order based on the affect construct they measure, the information source they use, the learning context in which the study was done, and the specific computational approach adopted. Most of the works reviewed here measure discrete emotion categories like confusion, interest, boredom, etc (Mavrikis, Maciocia, & Lee, 2007; Kapoor & Picard, 2005; D'Mello, Picard, & Graesser, 2007; and Sarrafzadeh, Fan, Dadgostar, Alexander, & Messom, 2004); while a few use appraisal-based models of emotion (Jaques & Vicari, 2007; Heylen, Ghijsen, Nijholt, & Akker, 2005; Conati, 2002). Related constructs like difficulty, stress, fatigue and motivation have also received some attention (Whitehall, Bartlett, & Movellan, 2008; Liao W, Zhang, Zhu, Ji, & Gray, 2006; de Vicente & Pain, 1998).

Based on the modelling approach used, affect inference methods can be broadly categorised as (Liao et al. 2006; Alexander, Hill, & Sarrafzadeh, 2005):

- Predictive - those that predict emotions based on an understanding of their causes
- Diagnostic - those that detect emotions based upon their physical effects, and
- Hybrid - those that combine causal and diagnostic approaches

Table 2.2-2: Affect modelling in learning environments

Citation	Affect Construct	Information Source	Learning Context	Method
Whitehall, Bartlett & Movellan (2008)	Difficulty level and speed of content	Facial expressions	Lecture videos	Support Vector Machines and Gabor filters
Zakharov, Mitrovic & Johnston (2008)	Positive and negative valence	Facial expressions	Pedagogical agent-based educational environment	Rule-based system
Baker (2007)	Off task behaviour	Interaction Log files	Cognitive Tutor software	Latent response model
Jaques & Vicari (2007)	OCC Cognitive Theory of Emotions	User's actions & interaction patterns	Pedagogical agent based educational environment	Belief-Desire-Intention (BDI) reasoning; appraisal based inference
D'Mello, Picard & Graesser (2007)	Flow, confusion, boredom, frustration, eureka & neutral	Posture, dialogue and task information	Dialogue based ITS-Auto Tutor	Comparison of multiple classifiers
Kapoor, Burleson & Picard (2007)	Pre-frustration & Not pre-frustration	Facial expressions, posture, mouse pressure, skin conductance, task state	Automated Learning Companion	Gaussian process classification; Bayesian inference
Mavrikis, Maciocia & Lee (2007)	Frustration, confusion, boredom, confidence, interest & effort	Interaction logs & situational factors	Interactive Learning Environment-WALLIS	Rule induction
Liao et al. (2006)	Stress & fatigue	Physical appearance, physiological, behavioural and performance measures	Maths and audio based experimental tasks	Influence Diagram; Ensemble of classifiers
Amershi, Conati & Maclaren (2006)	Affective reactions to game events	Skin conductance, heart rate, EMG	Educational game-Prime Climb	Unsupervised clustering
Kapoor & Picard (2005)	Interest, Disinterest, break-taking behaviour	Facial expressions, posture patterns & task state	Educational Puzzle	Ensemble of classifiers
Heylen et al (2005)	Scherer's Component Process Model	Facial Expressions, task state	Agent-based ITS for nurse education-INES	Appraisal using stimulus evaluation checks
Sarrafzadeh et al (2004)	Happiness/success surprise/happiness sadness/disappointment, confusion frustration/anger	Facial expressions	Elementary Maths ITS	Fuzzy-rule based classification
Litman & Forbes (2003)	Negative, neutral & positive emotions	Acoustic-prosodic cues, discourse markers	Physics Intelligent Tutoring Spoken Dialogue System - ITSPOKE	Comparison of multiple classifiers
Conati (2002); Conati & Zhou (2004)	OCC Cognitive Theory of Emotions	Interaction patterns, personality, goals	Educational game-Prime Climb	Dynamic decision network; Appraisal based inference
de Vicente & Pain (2002; 1998)	Motivation	User actions and interaction patterns; Experience sampling	Japanese numbers ITS-MOODS	Motivation Diagnosis Rules

The predictive approach takes a top-down causal view to reason from direct input behaviour like state knowledge, self-reports, navigation patterns or outcomes to actions. It is generally based on sound psychological theories like Scherer's Component Process Model (Scherer, 2005) or the OCC Cognitive Theory of Emotions (Ortony, Clore, & Collins, 1998). The appraisal theory provides a detailed specification of appraisal dimensions along emotion-antecedent events like novelty, pleasantness, goal-relevance, coping potential and norm/self compatibility; but suffers from the methodological problem of reliance on an accurate self-appraisal. The OCC theory on the other hand defines 22 emotions arising as valenced reactions to situations consisting of events, actors and objects. It does not however include some important affect states like boredom, interest and surprise which are relevant to learning scenarios (Picard, et al., 2004).

Conati (2002) and Conati and Zhou (2002) implement the OCC theory to assess learner emotions during interaction with an educational game. They use a dynamic decision network to model affect states but do not establish the accuracy of the model empirically. In another study, de Vicente and Pain (2002) were able to formalise inference rules for diagnosis of motivation using screen capture of learner interactions with a tutoring system. This work is significant in that it relies only on the concrete aspects of learner interactions such as mouse movements and quality of performance for motivation inference. These rules however, have not been implemented and hence remain a theoretical assumption. Heylen et al. (2005) describe an attempt to relate facial expressions, tutoring situation and the mental state of a student interacting with an intelligent tutoring system. They do not infer affect states automatically from facial expressions but use Scherer's Component Process Model (2005) of emotion appraisal using stimulus evaluation checks. Their results are inconclusive and specific to the tutoring system used in their study.

Diagnostic methods on the other hand take a bottom-up approach and are based on inference from sensory channels using traditional pattern classification techniques to approximate or estimate affective behaviour. These rely on the understanding that non-verbal behaviour through bodily gestures, facial expressions, voice, etc, is instinctively more resourceful and aims to infer affective cues with the aid of sensors. Notable in this category is the Affective Computing Group at MIT which is involved in a series of projects towards the building of a *Learning Companion*. Kapoor et al. (2007) use a novel method of self-labelling to automatically classify data observed through a combination of sensors, into 'pre-frustration' or 'not-pre-frustration'. In related work, Kapoor and Picard (2005) use multi-sensor classification to detect interest in children solving a puzzle by utilising information from the face, posture and current task of the subjects. The high recognition rates on these classification techniques are achieved for a single distinct affect state using sophisticated and fragile equipment. These do not as yet perform real-time classification.

D'Mello and Graesser (2007) use posture patterns along with dialogue, to discriminate between affect states during interaction with an intelligent tutoring system called Auto-Tutor. This is a dialogue based system achieving recognition of affect states like flow,

confusion, boredom, frustration, eureka and neutral. Interestingly however, the ground truth used for validating their classification is mainly the facial action coding of recorded interaction by FACS experts. FACS or the Facial Action Coding System is the anatomic classification devised by Ekman and Friesen (1978) that defines 44 Action Units to describe any human facial expression and will be described further in Chapter 4.

Amershi et al. (2006) use unsupervised clustering to analyse students' biometric expressions of affect that occur within an educational game. Their approach is quite interesting and different from the usual supervised classification techniques normally applied for automatic sensing. However, lack of a benchmark or standard to compare performance makes it difficult to evaluate the efficiency of this method.

Sarrafzadeh et al. (2004) employ a fuzzy approach to analyse facial expressions for detecting a combination of states like happiness/success, surprise/happiness, sadness/disappointment, confusion and frustration/anger. They do not, however, give a measure of the accuracy of their method and focus more on the stage after detection. Litman and Forbes (2003) propose a method of affect modelling from acoustic and prosodic elements of student speech. Their study is particularly relevant for dialogue based systems.

Recent works of Zakharov, Mitrovic, and Johnston (2008) and Whitehall, Bartlett, and Movellan (2008) that use facial expression analysis techniques to measure valence and difficulty level, respectively, also fall within this category.

Finally, models of hybrid approaches, as in Conati (2002) and Liao et al. (2006), leverage the top-down and bottom-up evidence in an integrated manner for improved recognition accuracy. This involves using dynamic probabilistic approaches to model uncertainty in affect and its measurement, while explicitly modelling the temporal evolution of emotional states. Such frameworks are promising as they can allow context-sensitive interpretation of affective cues. However, specification and fusion of information from the multiple channels still remains a significant challenge for actual implementation.

2.3 Discussion and scope of this dissertation

Ideally, automatic sensing should be able to function in real-time; measure multiple and co-occurring emotions unobtrusively and without causing disruption in the actual learning process. As reviewed in the previous section, numerous efforts are being made towards this goal to give computer-based tutoring some semblance of emotional intelligence. Table 2.2-2 lists the relevant works and categorises these according to their specific focus and approach. It highlights the variety in modelling techniques that range from rule-based systems to complex probabilistic models; the different ways in which affect is conceptualised in these systems based on whether a dimensional, discrete or appraisal-based stance is adopted; the array of interactional as well as behavioural measures used to infer affect; and importantly, the nature and focus of the learning setup used. Given this diversity in the measured affect constructs, the specific learning environments and the channels used as information sources;

it is difficult to comment on the overall performance of a system and determine its efficiency in a broad sense. This inability to make generalisable claims is an acknowledged limitation of affect sensing technologies (Pantic & Rothkrantz, 2003) and makes it challenging to establish the merit and success of a particular system satisfactorily and with confidence. Nevertheless, what is apparent is a growing understanding of the importance of affect modelling in learning and this substantiates further research in the area. The following sections lay out some design choices that set the scope of this dissertation and define the problem space.

2.3.1 Conceptualisation of affect

The definitional debate on the rather fuzzy concept of emotion poses a fundamental problem for distinguishing affect states in learning. In absence of a standard theory or model of affect in learning, the choice of conceptualising emotion is principally ad-hoc. A pragmatic approach is recourse to domain relevant folk concepts of emotions derived from natural languages (Scherer, 2005; Lisetti & Schiano, 2000).

Such a lexical taxonomy underlies the Mind Reading DVD (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) which is an interactive computer-based guide to emotions. Based on a taxonomic classification by Baron-Cohen, it groups 412 emotion concepts into 24 mutually exclusive emotion groups. Each group encompasses the finer shades of that emotion concept and therefore gives the flexibility of choosing the right level of semantic distinction. The DVD in itself is a rich corpus of labelled video and can serve as a powerful tool for preliminary analysis. Using this taxonomy, I selected a set of affect categories that are representative of some important affective states linked to learning (See Section 2.1). These are – Afraid, Angry, Bored, Happy, Interested and Unsure. These constitute a challenging set of complex mental states that have been studied extensively for their relevance to learning. Figure 2.3-1 illustrates some of the emotion concepts that they encompass. Although each emotion group encompasses more emotion concepts than the ones shown, I have selected for the purpose of illustration only a subset here. These encompass representative emotions from each of Kort, Reilly and Picard's (2001) emotion axes as well as those of Pekrun et al.'s (2002) academic emotions with the exception of hope, pride and shame which have more complex social antecedents and meanings and are therefore excluded from this study.

2.3.2 Learning context

We know that emotions are situated, have contextual antecedents and are influenced by social consequences. Knowledge of the learning setting is important then to ground a research work in a specific context and help assess its generalisation ability. The nature and dynamics of emotions in a solo learning setting e.g., Conati and Zhou (2002), Conati (2002), will no doubt differ from those generated within an agent-based learning environment like in Jaques and Vicari (2007), Heylen, Ghijsen, Nijholt, and Akker (2005), Kapoor, Burleson, and Picard (2007), or with those that involve dialogue, as in D'Mello, Picard, & Graesser (2007), Litman and Forbes (2003). The nature of affect and its dependence on context thus makes the choice of a learning environment an important one. As such, I decided to use a solo, one to one learning setting for my study. By focusing on a self-regulated learning model my

objective was to minimise the potential effects of design variables like instructional strategy, pedagogy integration, learning theory, process of communication, collaboration, presence of an embodied agent, etc; in the assessment and interpretation of emotional experience.

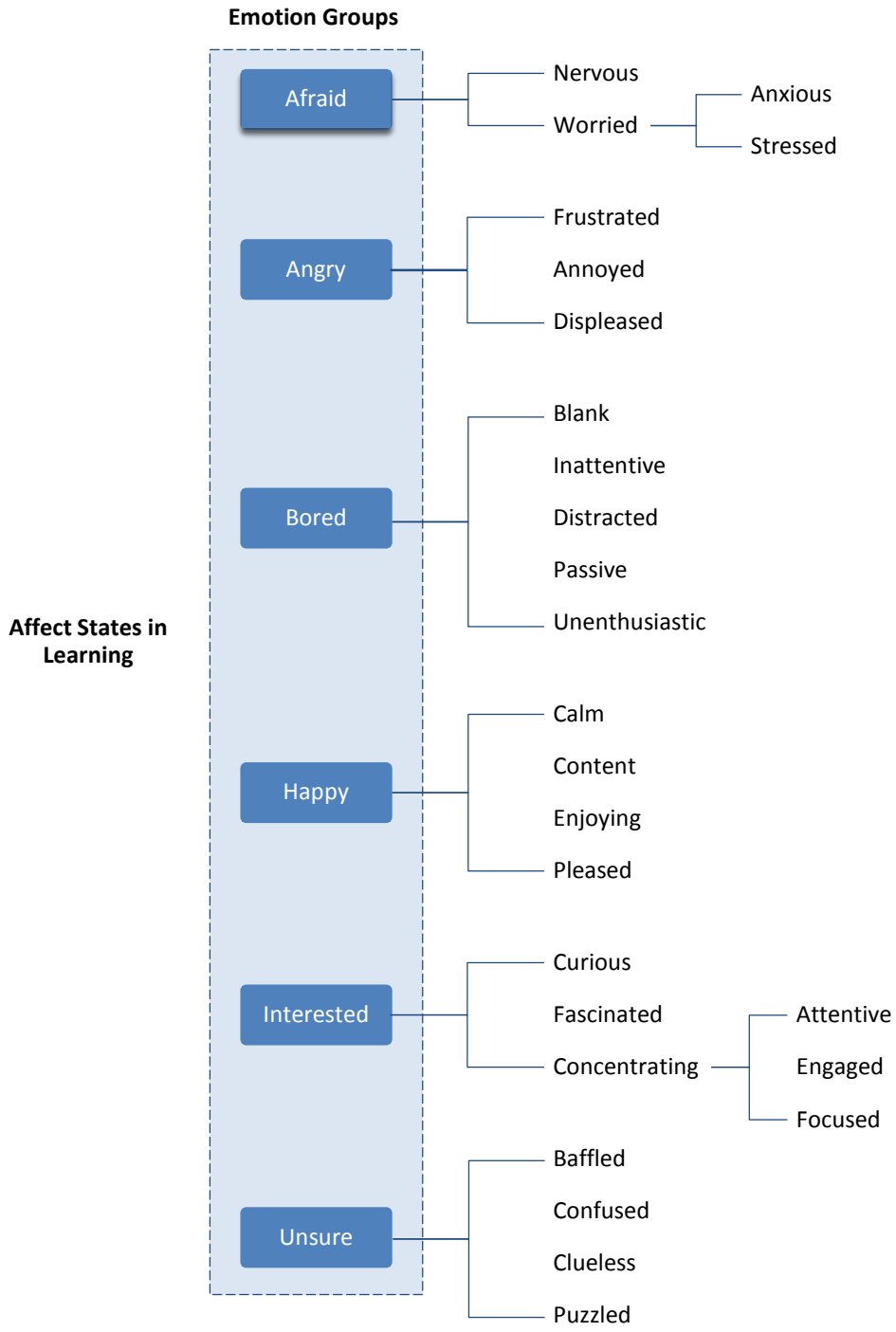


Figure 2.3-1: Selected view of the emotion groups chosen for this study and the emotion concepts they encompass (derived from Baron-Cohen, Golan, Wheelwright, and Hill, 2004)

2.3.3 Choice of modality

Emotion is expressed through visual, vocal and physiological channels. The visual channel includes facial expressions, body gestures, eye-gaze and head pose; the vocal channel focuses on measures of intonation and prosody; while the physiological channel includes measures of skin conductance, blood volume pressure, heart rate, temperature, etc. Lack of a consistent mapping between observable aspects of behaviour and actual affective states, technical feasibility, and practical issues complicate the choice of modality for sensing in a learning setting. Issues of ethics, privacy and comfort further constrain the design, use and deployment of appropriate sensing technologies. The use of physiological sensing in particular is challenging. Though relatively easy to detect and reasonably unobtrusive now, physiological sensing has some inherent shortcomings like requirement of specialised equipment, controlled conditions, baseline determination and normalising procedures, possible discomfort in usage, expertise in use of sensing apparatus and issues of privacy and comfort (Scherer, 2005; Hudlicka, 2003). Speech analysis may not always be suitable as not all learning environments are dialogue based. Table 2.3-1 below gives a brief comparative overview.

Table 2.3-1: Overview of the three dominant channels of nonverbal behaviour

Visual	Vocal	Physiological
Facial expressions, Head pose, Body gestures, Eye-gaze	Speech, Prosody and Intonation	Skin conductance, Blood volume pressure, Heart rate, Breathing rate, Temperature, Muscle tension
▪ Natural and observable ▪ Unobtrusive ▪ Practically deployable ▪ Does not require specialised equipment; exception for gestures and eye-gaze ▪ Behavioural coding required to set ground-truth	▪ Natural, discernable ▪ Unobtrusive ▪ Practically deployable ▪ Limited to dialogue based systems ▪ Manual annotation required to set ground-truth	▪ Unobservable ▪ Unobtrusive but has issues with comfort and privacy ▪ Requires tightly controlled environmental conditions ▪ Specialised and fragile equipment ▪ Easy to access the bio-signals but difficult to interpret

As reviewed in previous works listed in Table 2.2-2, multiple channels are currently being probed for emotional signs ranging from facial expressions, posture, pressure patterns, prosody, interaction patterns and even trait factors like personality. Combination of one or more channels is likely to improve accuracy of emotion but is a challenging problem and a research avenue in itself. An important issue here is to understand redundancy and variation in the time course of the different information channels to inform purposeful fusion of relevant information. Works like that of D'Mello, Picard, and Graesser (2007) who analyse relative contributions of information channels are important for viable design and implementation of such systems.

It is interesting to note from previous works in affect recognition as to how labelling of recorded behaviour relies on manual facial expression analysis and how this is used as the gold-standard, or benchmark, for evaluating the accuracy of a method, irrespective of the sensed modality. This indicates that while judging someone's affective state humans predominantly rely on facial expressions. Given this pre-eminence of facial signs in human communication the face is a natural choice for inferring affective states. Facial information can be detected and analysed unobtrusively and automatically in real-time requiring no specialised equipment except a simple video device. Although recent studies have looked at the divergence in emotional information across modalities (Cowie, 2009; Cowie & McKeown, 2009), affect inference from facial expressions has been found to be consistent with other indicators of emotion (Cohn, 2006).

However, facial expressions are not simple read-outs of mental states and their interpretation being context-driven is largely situational. Computer tutors can exploit this aspect to infer affective states from observed facial expressions using the knowledge state and navigation patterns from the learning situation as supporting evidence. Given the requirements of an affective computer tutor, the visual modality thus has a great potential for evaluating learner states thereby facilitating an engaging and optimal learning experience. It is for these reasons that the visual modality was selected for affect analysis in this work.

2.4 Visual affect recognition

Following Darwin's seminal work on *The Expression of Emotion in Man and Animals*, the study of facial expressions has been a subject of rigorous scientific enquiry. Lately, it has found enthusiastic support in the HCI community and is in fact considered to be indispensable for affective HCI design (Pantic & Bartlett, 2007). Automatic facial expression analysis has made considerable progress over the past decade, excellent reviews of which can be found in Fasel and Luettin (2003), Pantic and Rothkrantz (2000), and Zeng, Pantic, Roisman, and Huang (2009). Figure 2.4-1 outlines the sequence of steps a typical facial expression recognition system undergoes. The three steps involve:

1. face acquisition, in which the face is detected or located in the scene of interest;
2. facial feature extraction, in which the shape of facial components and/or the texture of the facial area is described from the detected face region; and
3. facial expression classification, in which the facial features and/or the changes in the appearance of facial features is/are analysed and classified into some facial expression interpretative categories like facial muscle activations or emotion categories.

Currently, most automatic facial affect analysers support a limited number of basic emotion states, ignore context while performing inference, rely on posed data that does not generalise to naturalistic data and in most cases adopt strong assumptions that limit robust feature extraction and variation in appearance (Pantic, Pentland, Nijholt, & Huang, 2007;

Schwaninger, Wallraven, Cunningham, & Chiller-Glaus, 2006). As a result, recent technologies of affect perception from facial expressions do not achieve recognition in a manner as would be suitable for application within a computer tutor (Picard et al, 2004; Pantic, 2003).

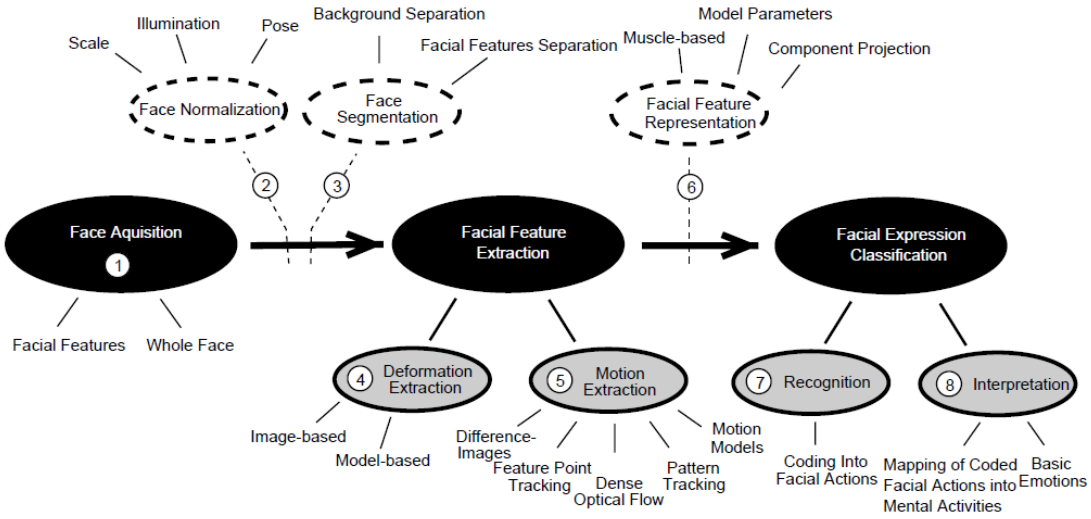


Figure 2.4-1: Generic facial expression analysis framework (from Fasel and Luettin, 2003)

2.4.1 MindReader – a mental state inference tool

Perception and interpretation of facial expressions is inherently complex, and therefore, computationally challenging. Adopting a modular perspective by separating measurement of facial behaviour from interpretation of affect states can significantly reduce this complexity (Ekman, 1982). This separation can also facilitate subjective evaluations of data to be abstracted into measurable features for implementation. The mental state inference tool developed by El Kaliouby (2005) for her doctoral dissertation at the University of Cambridge supports such a modular approach and represents the state-of-art technology in the field. The decision to study the visual modality was in fact also influenced by the availability of this tool even though, as will be discussed in the ensuing chapters, the framework was not eventually adopted. Figure 2.4-2 depicts the MindReader system overview. Video input is abstracted spatially and temporally into head and facial events at different granularities. On each level, more than one event can occur simultaneously. The three levels are briefly defined here:

Head and facial actions

This level models the basic spatial and motion characteristics of the face and head pose. The motions are described by the FACS (Ekman & Friesen, 1978) – an objective system for measuring facial and head motions. By tracking feature points over an image sequence and analysing their displacements over multiple frames, a characteristic motion pattern for various action units (AUs) like lip-pull, brow-raise, head tilt, etc can be calculated. The head and facial actions are abstracted as spanning five video frames, that is, approximately every 166ms at 30fps.

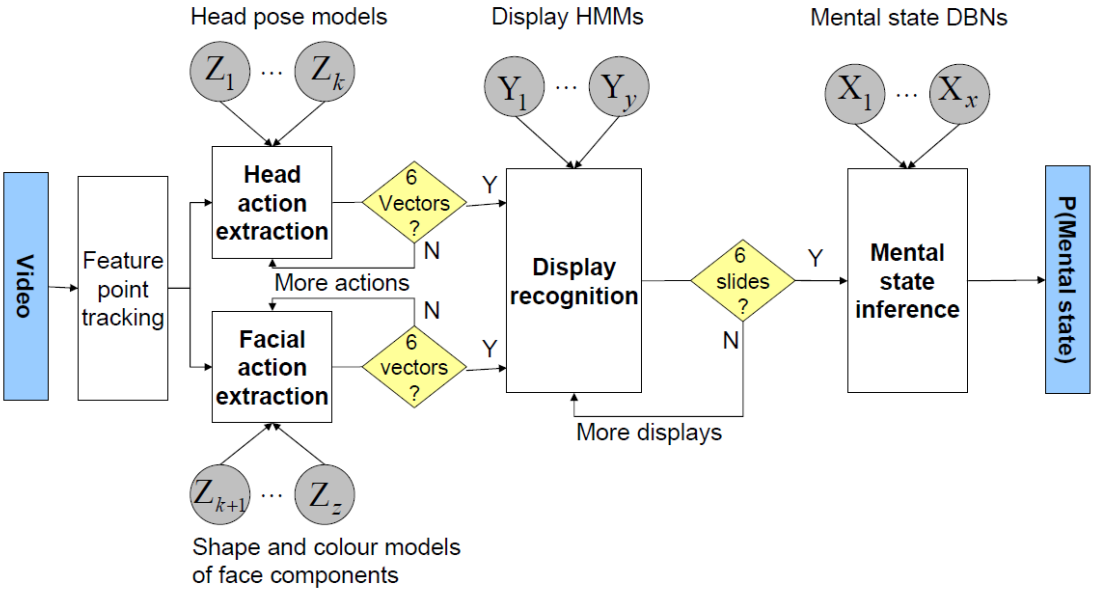


Figure 2.4-2: Procedural description of inference in the MindReader (from El Kaliouby, 2005)

Head and facial displays

Displays serve as the intermediate step between tracked AUs and the inferred mental states. The input to this level is the running sequence of head and facial actions. The MindReader supports facial and head displays like head nod, head shake, head tilt, head turn, lip corner pull, lip pucker and mouth open. The displays are modelled using discrete Hidden Markov Models (HMMs). Individual displays span 30 frames, that is, every one second at 30fps.

Mental states

The top-most level of the model represents six mental states – *agreeing*, *disagreeing*, *thinking*, *interested*, *concentrated* and *unsure*. The probability of a mental state is conditioned on the most recently observed displays and previous inferences of the mental state. A separate Dynamic Bayesian Network (DBN) models the probability of each mental state allowing the system to represent mental states that may co-occur. Each mental state is modelled as spanning 60 frames or 2 seconds at 30fps.

2.4.2 Working with the MindReader

MindReader is a computational tool that has been validated to perform real-time inference of complex mental states from head and facial displays in a video stream (El Kaliouby & Robinson, 2005; El Kaliouby & Robinson, 2004). It combines bottom-up vision-based processing of the face with top-down predictions of mental state models to interpret complex mental states like agreeing, disagreeing, concentrated, interested, thinking and unsure. It requires a standard video- or web-camera to perform inference in an unobtrusive manner with no manual intervention or prior calibration. These properties made it suitable for modelling of affect in an environment such as learning with a computer tutor. But since a

different set of affect states were identified for the purpose of this study, the system models had to be re-trained for recognition of affect states I was interested in. As the three levels of the MindReader have been trained and developed independently of others, only the mental state level of the MindReader was trained with examples of the relevant affect states. The two bottom levels, that recognise the facial actions and the facial displays respectively, correspond to an objective measurement of the facial expression changes and were therefore, not re-trained. The training process of the mental states can be invoked from within the MindReader by specifying the new emotion samples. The output DBN files are then replaced with the original ones to estimate the performance of the re-trained models.

I used the same corpus as El Kaliouby (2005), for initial training of the models. This corpus - the Mind Reading DVD (Baron-Cohen, Golan, Wheelwright, & Hill, 2004), is based on the same lexical taxonomy from which affect categories were selected for this study. The DVD has six videos for each emotion concept ranging in duration between 5-8 seconds. All videos are frontal with a uniform white background and have a resolution of 320x240. The mental states were acted out by 30 trained actors using example scenarios but with no specific instructions on how to perform them. All actors were looking into the camera and none of them were talking. The captured videos were validated by a panel of 10 judges who were asked the question 'could this be *the emotion name*?' When 8 out of 10 judges agreed, the video was included in the library.

The videos from the DVD constituted the training set representing the selected affect states and included a total of 32 mental state concepts in all. Out of these, 13 videos had to be discarded as the tracker failed to initialise on the initial frames. The training set was effectively reduced to 179 video clips. The performance of the trained models when tested on the training data itself reduced drastically from the originally reported 77% to approximately 25%. Generally, if a classifier is trained and tested on the same dataset, recognition results are significantly higher than those obtained using cross-validation. However, there was no point in doing cross-validation or generalisation tests because the recognition accuracy on the training data itself was too low to merit the effort. The detailed confusion matrix for the inference outputs is presented in Table 2.4-1.

The confusion matrix gives a succinct visualisation of classifier performance. Affect states along the rows represent the actual class labels and along the columns they represent the predicted ones. The number of correctly classified instances thus appears along the diagonal. Specifically, row i of the matrix describes the classification results for class i while column i lists the number of times class i is recognised. The totals column gives the total number of sample videos that are labelled as i. The last column lists the true positive rate (TP) or the classification rate for the class i. It is computed as the ratio of videos correctly classified as class i to the total number of videos labelled as i. The totals row on the other hand, yields the total number of videos that are classified as belonging to class i. Finally, the bottom row yields the false positive rate (FP) for class i which is the ratio of videos falsely classified as i to the total number of videos that are labelled as anything but i.

Table 2.4-1: Detailed confusion matrix of inference results showing classification accuracy of approx. 25% with a false positive rate of 15%

Predicted →

← *Actual*								
Affect State	Bored	Angry	Happy	Afraid	Interested	Unsure	**Total**	**TP %**
Bored	**8**	0	0	4	1	12	25	32.0
Angry	4	**5**	1	8	5	4	27	18.5
Happy	3	0	**1**	3	2	14	23	4.4
Afraid	4	1	0	**17**	2	11	35	48.6
Interested	3	0	0	6	**0**	26	35	0.0
Unsure	0	1	3	9	5	**16**	34	47.1
Total	22	7	5	47	15	83	**179**	**25.1**
FP %	9.1	1.3	2.6	20.8	10.4	46.2	**15.1**	

2.4.3 Discussion

It was difficult to estimate the underlying cause for the drop in recognition accuracy of the MindReader when trained on a different set of emotions. The obvious deduction is that the MindReader was conceived and optimised for a specific set of mental states and its generalisation to an entirely new set of categories had not been determined. As such, the assumptions and thresholds that were true for the original affect states may not have been relevant for the new affect states. Specifically with regard to the facial and head displays, the MindReader does not support some significant ones that are characteristic to the affect states under study such as brow-lowering and asymmetric mouth movements in case of confusion. This obviously reduces the recognition accuracy for affect states that may contain these as strong discriminators.

Moreover, a manual observation of the affect states' videos revealed differences in the dynamics of displays across affect states. Dynamic features include measures like duration, intensity and velocity and are known to differentiate between morphologically similar but psychologically different facial configurations (Cohn & Schmidt, 2004). Background literature in human perception of emotions reveals that dynamic features play an important role in attributing meaning to observed behaviour. An occurrence of a head-tilt for instance can be common to both boredom and interest but a difference in duration and speed can discriminate between the two. Augmenting the MindReader with dynamic features would undoubtedly increase the accuracy. However, the videos in the DVD mostly include peak expressions and do not contain the natural sequence of onset, apex and offset which is necessary to compute the dynamic features. As such, a more natural representation of the affect states was required.

Overall, this highlights the importance of low-level structural characteristics in building affect recognition systems and the intricate dependence of higher level inference on such lower level measurements. In order to address these issues, the two bottom levels modelling the facial actions and the facial displays needed to be re-trained for the current set of affect states. This would require extracting relevant samples at each of these levels, labelling them,

and then determining the model parameters anew. In effect, this would involve re-building the entire MindReading framework. Since the focus of this research was on naturalistic data, this was not pursued further at this stage as the generalisation ability would still have remained questionable, as has been found by other researchers when faced with naturalistic data (Batliner, Fischer, Huber, Spilker, & Noth, 2003).

However, two important strands emerged from this exercise. Firstly, I observed that for a realistic study it was more appropriate to focus on context-relevant naturalistic corpora and design appropriate modelling approaches based on that understanding. The reduced generalisation of the MindReader indicated differences in the underlying characteristics of emotions at a lower structural level. If the thresholds and assumptions validated for identifying a certain set of emotions do not hold for a different set of emotions even within the same database, it is unlikely that they will do so for an entirely different, and importantly, naturalistic database. Although the DVD is an excellent corpus of labelled data it is recorded in an entirely different functional context with a strong social directedness or orientation. Moreover, the use of actors in the recording of videos makes the expressions extremely artistic and powerful. This I believe makes the emotional tone of the DVD too removed from what one would expect to arise in a standard learning environment. This, along with the recent emphasis on research using naturalistic data for viable applications of affective computing (Pantic, Pentland, Nijholt, & Huang, 2007; Cowie, Douglas-Cowie, & Cox, 2005; Douglas-Cowie et al., 2004), supported a modelling approach based on data collected in a relevant context. An in-depth account of the data collection and subsequent analysis is the theme of Chapters 3 and 4.

Secondly, I was interested in determining the adequacy of facial feature points in encoding relevant expression changes for affect perception. Facial feature point tracking has numerous advantages that make it desirable for use in real-time applications as compared to other alternatives. It can be compared to the point-light technique used in psychology for studying various phenomena related to human perception of biological motion and findings where the movement of points was found to be a good predictor of emotions (Bassilli, 1978). As facial feature point tracking is the primary input taken by facial affect analysers, this makes determining its efficiency crucial for eventual classifier performance. To evaluate the information value of automatically tracked feature points an experiment was conducted details of which appear in Chapter 5.

2.5 Summary and conclusions

A consistent theme that emerges from education research is that teaching and learning are essentially ‘emotional practices’. We know that learners experience a wide range of both positive and negative emotions, and that these influence their cognitive functioning and behaviour. Access to emotions is then important to ensure optimal learning, more so in the case of computer-based learning environments where the learner’s motivation is an important determinant of engagement and success. However, automatic measurement of

affect is a challenging task. Emotions consist of multiple components that may include intentions, action tendencies, appraisal, other cognitions, central and peripheral changes in physiology, and subjective feelings. As a result they are not directly observable and can only be inferred from expressive behaviour, self-report, physiological indicators, and context (Cohn, 2006).

This chapter has outlined the problem space with respect to application of affect-sensitive technologies in computer-based learning. Building on a discussion of studies highlighting the relevance of emotions in learning, the different techniques for measuring emotions and recent advances in automatic recognition and/or prediction of affect in learning contexts were discussed. Six categories of pertinent affect states were identified; the visual modality for affect modelling was preferred given the requirements of a viable measurement technique; and a bottom-up analysis approach based on context-relevant data was deemed appropriate. The next chapter will describe the data collection and annotation procedures and discuss some preliminary observations.

3. Representative Data

Automatic inference using machine learning relies on extensive training data which serves as the ground-truth for development and evaluation of appropriate algorithms. For viable applications of affect sensitive technology the use of naturalistic over posed data is being increasingly emphasised. Creating a repository of naturalistic data is however a very challenging task. This chapter reports results from the collection and subsequent annotation of data obtained in a learning scenario. The conceptual and methodological issues encountered during data collection are discussed, and problems with labelling and annotation are identified. A comparison of the compiled database with some standard databases is also presented.

3.1 Introduction

As emotion research gradually integrates with HCI studies and matures in application from mere prevention of usability problems to promoting richer user experiences, the need to capture 'pervasive emotion' (Cowie, Douglas-Cowie, & Cox, 2005) and also its context of occurrence is becoming an increasing concern. Existing databases are often oriented to prototypical representations of a few basic emotional expressions, being mostly posed or recorded in scripted situations. Such extreme expressions of affect rarely occur, if at all, in HCI contexts. The applicability of such data therefore becomes severely limited because of its observed deviation from real-life situations (Batliner, et al., 2003; Cowie, Douglas-Cowie, & Cox, 2005) and for my purpose its relevance to a learning situation like one-on-one interaction with a computer tutor. There is evidence that naturalistic head and facial expressions of affect differ in configuration and dynamics from posed/acted ones and are, in fact, mediated by separate neural pathways (Cohn & Schmidt, 2004; Pantic & Patras, 2006). Ekman (Ekman & Rosenberg, 1997) identifies at least six characteristics that distinguish spontaneous from posed facial actions: morphology, symmetry, duration, speed of onset, coordination of apexes and ballistic trajectory. Moreover, there is an increasing emphasis on the role of situational context in the nature and meaning of emotion (Russell & Fernandez-Dols, 1997).

Ideally then, a database should depict naturalism, limited or no experimental control, and be contextually relevant. Since existing databases mostly include deliberately expressed emotions and are recorded in contexts that differ from their eventual application, their

relevance to a naturalistic situation like learning with a computer is debatable conceptually, and as found practically (Batliner, et al., 2003; Cowie, Douglas-Cowie, & Cox, 2005; Ekman & Rosenberg, 1997). Consequently, for developing applications that are generalisable to real-life situations, there is now an increasing shift from easier to obtain posed data to more realistic naturally occurring data in the target scenarios.

Handling the complexity associated with naturalistic data, is however, a significant problem. Nonverbal behaviour is rich, ambiguous and hard to validate, making naturalistic data collection and labelling a tedious, expensive and time-consuming exercise. In addition, lack of a consistent model of affect makes the abstraction of observed behaviour into appropriate labelling constructs very arbitrary. However, the need for representative data is essential in order to carry out realistic analysis, to develop appropriate methods and eventually perform validation of inferences. Thus, to ensure ecological validity and assist in a more meaningful interpretation, it was deemed necessary to study affect patterns as they occur naturally in context. The eventual purpose was to abstract this behaviour in terms of features that can enable automatic prediction and reliable computational modelling of affect states. This motivated a data collection exercise, details of which are presented in the following sections.

3.2 Data Collection

Before conducting a formal data collection exercise, a trial was undertaken to get an idea of what sort of data to expect. Three participants of mean age 25 years took part in a pilot study that involved video recording them in their habitual work-place while doing two computer-based learning tasks, namely a map-based interactive tutorial and a card sorting puzzle. Observations revealed some interesting points of inquiry and highlighted some technical issues to be considered for the formal data collection:

- Overall there was significant variability in the emotional behaviour of subjects even though they were of comparable age and ethnicity, and were doing the same task in the same setup. A marked distinction in the behaviour style was observed in that the facial and head displays although subtle, were consistent for individual subjects. This indicated a strong personality factor in emotional behaviour and was followed up by exploring relevant personality indicators of emotional expressivity.
- The frequency and range of displays across the two tasks were remarkably different indicating a clear influence of task on nonverbal behaviour. While the puzzle evoked quick jerky movements, the tutorial elicited more engaging and sustained gestures. It is difficult to establish whether the difference was due to the nature of the stimuli or in the perceived value of the task but the tasks were retained for the formal data collection to retain variety.
- The pilot recordings were not done in a formal lab setup in order to maintain naturalism. I found however that a certain degree of experimental control in terms of lighting and camera setup was necessary to allow video processing and analysis. To

achieve a compromise between naturalism and video quality a usability lab replicating a typical work-place setup was used for the formal data collection.

The pilot trial was useful to get preliminary insights into the feasibility of using visual cues and helped assess the best way to collect data in a realistic application setting. Based on the observations a formal data collection was conducted as described below.

3.2.1 Encoders

Eight participants, three males and five females in the age group of 21 to 32, were recruited to serve as encoders of emotional behaviour. The term encoder is used to denote these participants as being the source or examples for affective data obtained (Ekman & Rosenberg, 1997). All were regular and proficient computer users (μ=20 hrs of computer usage per week) so there was no effect of comfort level or exposure to the task requirements. Two of the participants wore glasses while one sported a beard. All participants recorded being happy, relaxed or in anticipation at the onset of experiment. They were informed that they would be video recorded during the interaction but remained naïve to the actual purpose of the experiment until after the experiment finished. Table 3.2-1 gives the profile of these participants based on their responses to a pre-experiment questionnaire.

Table 3.2-1: Profile of participants that served as encoders of emotional behaviour

Encoders	A	B	C	D	E	F	G	H
Origin	British	asian	asian	icelandic	asian	british	british	irish
Gender	Female	female	female	male	female	male	male	female
Age-group	27-32	21-26	21-26	27-32	21-26	21-26	21-26	21-26
Weekly computer usage (hrs)	20+	20+	20+	20+	20+	11-20	11-20	11-20
Mood at onset	Happy	relaxed	relaxed	pretty good	happy	relaxed anticipation	good	normal
Other				beard	glasses		glasses	

3.2.2 Setup

The recording setup was based on guidelines in Frank et al. (Frank, Juslin, & Harrigan, 2005). The experiment was conducted in a usability lab with a mock living room or personal office environment effect. It was chosen to facilitate video recording without compromising the naturalism of the desired behaviour. Standard computing equipment, that is, a desktop computer with a mouse and keyboard was used for the experiment. A video camera was mounted on top of the computer screen to allow video recording of the participants' upper body focusing mainly on the face. Additionally, a screen-capture utility, Camtasia™ Studio (2006), was used to obtain a complete interaction record for reference.

3.2.3 Measuring Expressivity

As observed in the pilot study, nonverbal behaviour research shows that there are indeed individual differences in the manner and intensity by which people express their felt emotions. Riggio and Riggio (2005) emphasise that emotional expressiveness as a personal style is relatively consistent across situations. As such, it should be interesting to observe if and how this dispositional expressiveness translates to HCI settings and what implications this could have for HCI in general and affective computing applications in particular.

Emotional or Dispositional Expressivity

Emotional expressivity is defined in two ways: to denote skill in sending messages nonverbally and facially - also known as nonverbal encoding ability (Riggio, 1986), and as a general expressive style and a central component of individual personality (Friedman, et al. 1980). Behavioural assessments and self-report measures are the two ways of measuring nonverbal expressiveness. Lack of standardised observation tests together with cost, time and reliability issues have made researchers turn to self-report means of assessing nonverbal or emotional expressiveness and have had good success with these (Riggio and Riggio, 2005). Self-report measures of nonverbal expressiveness assess individual differences in the generation and/or expressions of emotions and a more general tendency to display affect spontaneously and across a wide range of situations. Some popular measures are compared in Table 3.2-2 and include: Perceived Encoding Ability (PEA), Affective Communication Test (ACT), Berkeley Expressivity Questionnaire (BEQ), Emotional Expressivity Scale (EES), Emotional Expressivity Questionnaire (EEQ), Social Skills Inventory-Emotional Expressivity Sub-Scale (SSI-EE), Test of Attentional & Interpersonal Style (TAIS), Affect Intensity Measure (AIM) and Emotional Intensity Scale (EIS).

Based on how the construct of emotion is conceptualised, which component of emotion is assessed, the target population and administration time, availability, and psychometric properties like reliability and internal consistency, three self-report tests for measuring individual expressivity were selected. These are:

- Affective Communication Test - ACT (Friedman, Prince, Riggio, & DiMatteo, 1980). This is a 13-item measure of dynamic expressive style and gives a measure of the individual differences in nonverbal emotional expressiveness. It is strongly related to personality traits like charisma and eloquence.
- Emotional Expressivity Scale - EES (Kring, Smith, & Neale, 1994). This is a 17-item scale that conceptualises expressivity as a stable, individual-difference variable. It captures the general disposition towards the outwardly display of emotions regardless of valence or channel. Thus it presumes that expressivity is consistent across situations and across communication channels.

Table 3.2-2: Self-Report measures of Nonverbal Expressivity

Measure	Items	Construct	Psychometric properties	Availability	Citation
Test of Attentional and Interpersonal Style (TAIS)	144	More of a personality measure; attentional and interpersonal characteristics; 2 subscales:negative and positive affective expression; similar to EEQ & BEQ subscales	Test-retest reliability coefficients ranged from .60 to .93 (2 week interval)	www-enhanced-performance .com	(Nideffer, 1976)
Perceived Encoding Ability (PEA)	-	Perceived encoding ability	Questionable scale validity (Riggio. Widaman & Friedman, 1985)	-	(Zuckerman & Larrance, 1979)
Affective Communication Test (ACT)	13	Individual differences in nonverbal expressiveness or charisma	Good internal consistency alpha coeff. .77; Test-retest correlation .90 & .91 (2-month & 1- week interval);	Available from author	(Friedman, Prince, Riggio, & DiMatteo, 1980)
Affect Intensity Measure (AIM)	40	Individual difference variable;general temperament dimension of emotional reactivity and variability; people scoring high have more intense emotional reactions	alpha coefficient .90;Test-retest reliability coefficients ranged from .80,.81, .81 & .75 (1-,2-,3-,month & 2-year interval)	Published ->	(Larsen & Diener, 1985) (1987)
Social Skills Inventory (SSI) Emotional Expressivity Sub-scale	105	Expressivity as a dimension of social and interpersonal skill	alpha coefficient .75 to .88; test-retest reliability .81 to .96 (2-week interval)	Nominal fee - www.mindgarden.com	(Riggio, 1986)
Emotional Expressivity Questionnaire (EEQ)	16	Alternative to ACT but focusing more narrowly on emotional expressiveness. Developed as adjunct to a measure of ambivalence over emotional strivings-AEQ	Good internal consistency alpha coeff. ranging from .78	Published ->	(King. & Emmons, 1990)
Emotional Intensity Scale (EIS)	30	Intensity of positive and negative emotional states-also gives an overall score; similar to AIM but independent of the frequency with which states are experienced;	internal consistency alpha .90; test-retest correlation .83 (9-week interval)	Published ->	(Bachorowski & Braaten, 1994)
Emotional Expressivity Scale (EES)	17	Expressivity as a stable, individual-difference variable; captures emotional expressivity as a trait-like construct	Very good internal consistency alpha coeff. .91; Test-retest correlation .90 (4 week interval)	Published ->	(Kring, Smith, & Neale, 1994)
Berkeley Expressivity Questionnaire (BEQ)	16	Trait-like construct; Emphasises observable behavioural reactions. 3 sub-scales that measure occurrence of positive emotions (PEX), negative emotions (NEX) and strength of emotions (STR).	Good internal consistency alpha coeff. ranging from .71 to .76; Test-retest correlation .86 (2-month interval)	Published ->	(Gross & John, 1995)

- Berkeley Expressivity Questionnaire - BEQ (Gross & John, 1995). This is 16-item instrument that conceptualises expressivity as the behavioural changes associated with the experience of emotions. It emphasises observable behavioural reactions and has three subscales: BEQ-PEX that is a measure of expressivity for positive emotions, BEQ-NEX that is a measure of expressivity for negative emotions and BEQ-STR that gives an indication of the intensity or strength of emotional reactions.

Measuring Expressivity in HCI

In absence of a standard measure in HCI, I used a somewhat eclectic approach to measure the overall expressivity in an interaction sequence by using six global dimensions of expressivity together with the number of non-neutral emotional episodes observed for each participant. The idea was to see if there were any global indicators in terms of quantity and quality of movements and gestures that could give an overall estimate of expressivity in HCI.

Based on a global level speech annotation method (Martin, Abrilian, Devillers, Lamolle, Mancini, & Pelachaud, 2005), the following six dimensions or parameters were used to characterise expressivity:

- **Overall activation**: the amount of activity - {Static/Passive, Neutral, Animated/Engaged}
- **Spatial extent**: the amplitude of movements - {Contracted, Normal, Expanded}
- **Temporal extent**: duration of movements - {Slow/Sustained, Normal, Quick/Fast}
- **Fluidity**: continuity and smoothness of movement - {Smooth, Normal, Jerky}
- **Power**: strength and dynamics of movements - {Weak/Relaxed, Normal, Strong/Tense}
- **Repetitivity**: repetition of same expression/gesture several times - {Low, Normal, High}

A global measure, G, obtained by summing the scale values of these six parameters, was hypothesised to predict expressivity in HCI for comparison with the self-report tests.

3.2.4 Procedure

Participants were run individually in the usability lab and were observed via a one-way mirror from the adjoining room. This ensured that they were alone during the tasks and were not disturbed by an additional presence. Formal consent for recording was taken in writing from all subjects prior to the experiment. Subjects were video recorded while doing two tasks: an interactive map-based geography tutorial and a card matching activity. See Figure 3.2-1 for illustration. The session finished by completion of the expressivity test questionnaires described above, the self-annotation of videos, and subsequently, a semi-structured interview.

The tutorial enabled participants to study the countries and landscapes of different continents followed by a test of their learning. It served as a platform to observe facial affect signs when the learner is in complete control of the pace and strategy of the learning task. There was no time limit on this task but participants took on average about 20 minutes to complete this activity.

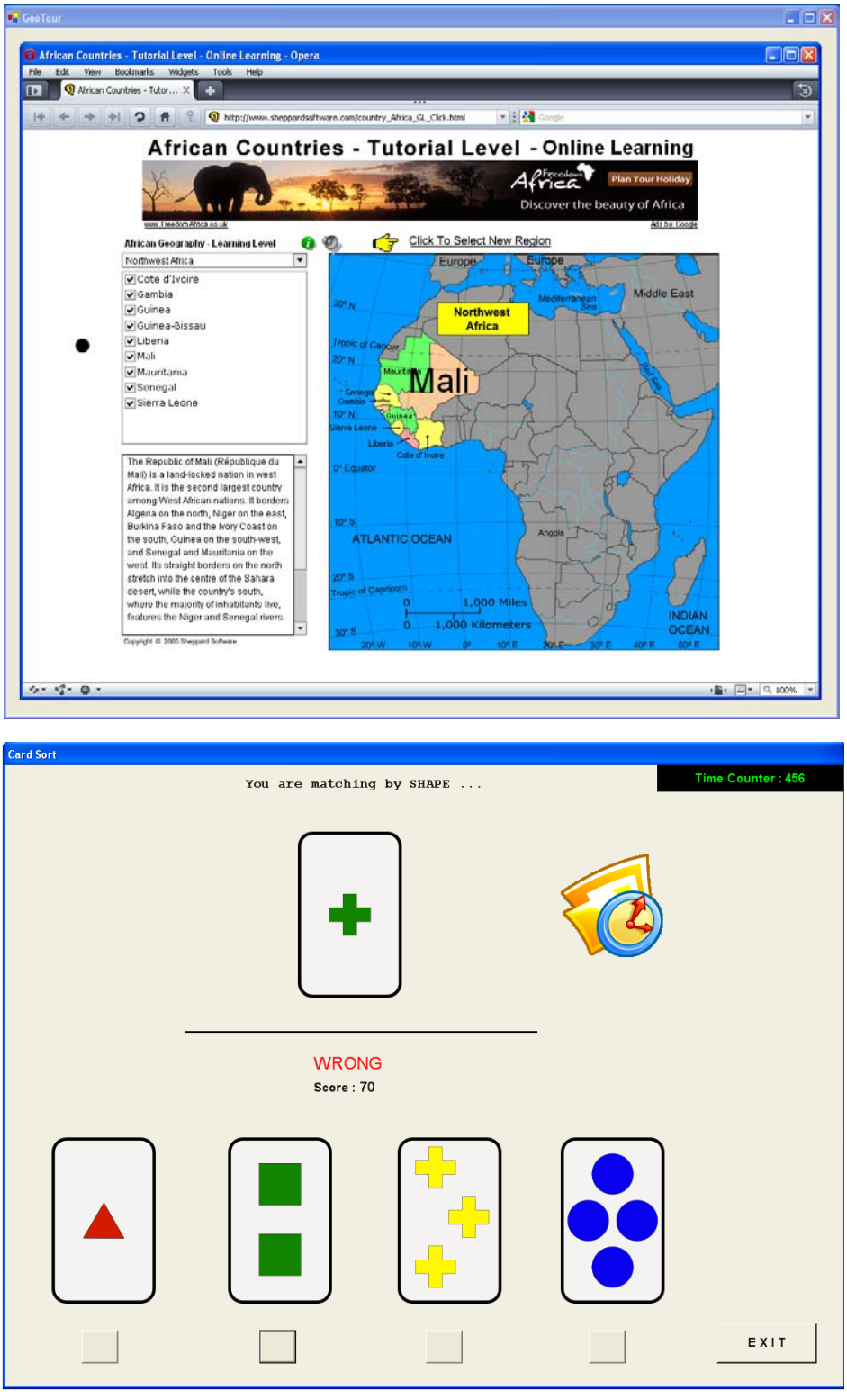


Figure 3.2-1: Screenshots of the two learning tasks used for inducing emotions

The second task was an adaptation of a card sorting activity meant to demonstrate the effect of situational anxiety on higher mental activities (Skemp, 1971). Cards having one, two, three or four of either squares, circles, crosses or triangles in red, green, blue or yellow were used - all figures on a card being alike and of the same colour. Participants had to sort the cards against four category cards based on a changing criterion. The four category cards - one red triangle, two green squares, three yellow crosses and four blue circles, are laid out and the subject is asked to sort the remaining cards first by colour, then by shape, then by number and finally in consecutive changing order of first by colour, second by shape, third by number and so on. This task is supposed to inhibit reflective intelligence leading to lowered performance and thus, decreased motivation. I was interested to see if this was accompanied by observable changes in facial expressions.

The two tasks were chosen to encourage a variety of emotion expressions and were sequentially varied across subjects. The card activity contained three triggers/events presented only once per participant during the game interaction (order-varied): the screen blanking out for five seconds, a match not being possible and variation in feedback/scoring. These are not dramatic deviations from the task and were used to induce reaction to some common interaction events. In the first case, the screen went to sleep for five seconds without warning but recovered immediately after. In the second event, the participants were asked to match a dummy card against the category cards. The dummy card had five diamond shapes in black and white and thus could not be matched according to any of the game rules. The change in feedback event was implemented by replacing the correct/incorrect answer text with a happy/sad smiley respectively.

3.2.5 Discussion

Approximately four hours of video data were collected from the eight participants. As before, there was a significant variability in the emotional behaviour of participants. Individual differences in expressivity were in fact quite striking. Some participants were animated and displayed a wide range of expressions while others were notably inexpressive. There was difference even in the way subjects reacted to the triggered events. Consistently across the encoder (participant) group, more emotional expressions occurred during the card game than during the tutorial substantiating the impact of task difference on the nonverbal behaviour of individuals. As individual and task differences seem to influence perceived emotional behaviour, it is reasonable to suggest that emotion inference technology will need to address these in design and function.

Global level annotation for expressivity was completed for all subjects. In addition, the number of non-neutral emotional episodes observed during an interaction sequence was also recorded. As mentioned in Section 3.2.3, the purpose of measuring the expressivity dimension was to observe if people display similar emotional behaviour in HCI settings as they do in real-life. If they do, then it would give us a sort of intransient stable factor to account for personal motion bias while doing automatic inference (Bernhardt & Robinson, 2007). If not, then it

would be an interesting result and may indicate re-evaluating assumptions that we make about affective behaviour and its manifestation in applications of affective computing.

In fact, no clear relationship of the subjects' dispositional expressivity and manifested behaviour was observed. The non-parametric correlation coefficient Kendall's Tau, τ, was computed to test the correlations. Kendall's Tau, τ, is similar to Spearman's r_s but is supposed to give a better estimate of correlation and therefore allow more accurate generalisations when the sample size is small (Field, 2009), as in this case. No significant correlations were observed between the test scores and the global expressivity measures although reassuringly, a significant correlation was found between the global expressivity index G and the number of non-neutral emotional episodes, $\tau = 0.69$, p (two-tailed) < 0.05; which essentially measures the same construct but at a different level of detail. This seems to suggest that the nature of overall emotional expressivity does not remain consistent across interaction contexts. In other words, the frequency and nature of emotional behaviour that occurs during social interactions may not be similar to that observed in human-computer interaction and this could have implications in the design, development and deployment of affect recognition technologies.

3.3 Annotation & Labelling

Automatic prediction using machine learning relies on extensive training data which in this case implies preparation of labelled representative data. This requires observational assessments on data to be represented in a quantifiable manner via annotation. It involves developing a protocol to catalogue observations and to represent the behaviour of interest using an appropriate coding scheme in terms of desired labelling constructs. The annotation method I used evolved from various domain relevant decisions related to the choice of labelling constructs and modality, anticipated technical constraints in the target scenario, relation to context and ease of interpretation. To achieve a compromise between descriptive detail and economy of annotation effort (Kipp, Neff, & Albrecht, 2007), this annotation scheme is tailored to this research but also applicable to similar areas. It is designed to map spontaneous interpretation of recorded behaviour onto affect states. Before elaborating on the annotation process itself, I will outline the choices and practices from nonverbal behaviour research that provides the framework for the annotation procedure.

Coding scheme

Coding schemes are theoretical stances that embody the behaviours or distinctions that are important for exploring the data. It is possible to locate these along a continuum, with one end anchored by physically-based schemes – schemes that classify behaviour with clear and well-understood roots in physiology, and the other end by socially-based schemes – schemes that deal with behaviour whose very classification depends far more on the mind of the investigator (and others) than on the mechanisms of the body (Bakeman & Gothman, 1997). Relevant examples of physiologically-based coding schemes are FACS, MAX and MPEG-4 which, although more standardised and comprehensive, are complex, require extensive

training and involve specialised procedures. Socially-based coding schemes on the other hand, are observational systems that are rooted in social processes and follow from cultural tradition or negotiation amongst observers as to a meaningful way to view and categorize behaviour. As a result, they require considerably more inference and potentially sensitive observers. To contrast with physically-based schemes, these examine behaviour or messages that have more to do with social categories of interaction like smiling or happiness rather than with physiological elements of behaviour like amplitude or a specific facial configuration (Manusov, 2005). Relative to facial affect analysis, the distinction is akin to what Ekman (1997) defines as the component versus judgement methods (Cohn, 2006). Since my goal is to quantify behaviour into the different affect categories, a socially based coding scheme was deemed more appropriate.

Level of Measurement

Determining the level of measurement is an important choice when examining nonverbal behaviour using a socially based coding scheme. It concerns the amount of behaviour examined and the extent to which the assessment involves more concrete indicators of behaviour's occurrence or more abstract assessments of the social meaning of behaviour (White & Sargent, 2005). The distinction can be referred to as macro vs. micro level of measurement and is in general related to the level of abstraction adopted. I reconciled two abstraction levels by following a hierarchical labelling process where an inferential level coding of extracting emotionally salient segments is followed by two levels of more focused coding along the pre-selected affect states.

Coding Unit

The coding unit refers to the decisions about when to code within an interaction and the length of time the observation should last. It has two broad variants - event based and interval based. Event based coding involves decision making triggered by a behavioural event of interest while interval based coding assesses pre-determined intervals of time within an interaction. Event based coding provides a realistic way of segmenting behaviours but it may result in loss of time information unless precise onset and offsets are noted. Interval based coding on the other hand is easy to use but requires selecting an optimal time interval and may truncate behaviour unnaturally. Choosing one over the other depends upon the research view and the level of accuracy required, complexity of the coding scheme and the frequency of behaviour occurrence (Bakeman & Gothman, 1997). I used interval-based coding to allow an easy and systematic observation in the first annotation round, but as discussed further on in Section 3.3.1, had to replace it with an event based coding.

Labelling Construct

Annotation schemes commonly employ either categorical, dimensional or appraisal based labelling approaches (Cowie, Douglas-Cowie, & Cox, 2005). In addition, free-response labelling may also be used for richer descriptions. As discussed in Section 2.3.1, I use a variant of

categorical labelling in which raters are asked to choose from pre-selected domain relevant emotional descriptors namely: *afraid*, *angry*, *bored*, *happy*, *interested* and *unsure*. These descriptors refer to non-basic affective-cognitive states and are pertinent in learning situations.

To familiarise the raters with their meaning and scope, a list of these emotion groups along with the emotion concepts they encompass (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) was provided at the beginning of the coding session. To reduce the bias of forced choice on selected affect labels - an often listed drawback in categorical methods (Russell & Fernandez-Dols, 1997), the scheme allows the rater to define his/her own category or label under a residual '*Other'* option if the perceived state is not represented by the provided categories. This ensures a degree of flexibility in coding and allows raters to express their responses in their preferred vocabulary or response mode.

Raters

Selecting raters or coders is an important aspect of designing annotation studies as they should be able to discern meaning from behaviour and make judgements effectively. I attempted three modes of annotation with respect to raters: self-annotation by encoders themselves, experts and non-experts. I use the term expert to denote raters who have some degree of formal experience as opposed to non-experts whose skills of emotion perception come from experience in day to day social interaction.

Reliability Measures

Inter-rater reliability measures for nominal data include raw-agreement, Scott's pi, Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha (Hayes & Krippendorff, 2007). Since the approach used here involves multiple raters rating multiple categories I use Fleiss' kappa to report inter-rater reliability (Fleiss, Levin, & Paik, 2003). Kappa is a statistical measure that calculates the degree of agreement in classification over that expected by chance and is scored as a number between 0 and 1, where 1 indicates perfect agreement. For practical purposes, values greater than 0.75 signify excellent agreement beyond chance and kappa values less than 0.40 represent poor agreement (Landis & Koch, 1977, as cited in Fleiss et al. 2003, pg. 604).

Having set the scope of the annotation framework in terms of general methodological decisions, I will now describe the three iterations of annotation that the data underwent.

3.3.1 First Annotation

Design

The very first annotation was performed by subjects themselves immediately after the experiment. The objective was to use this self-annotation as a triangulation method when comparing felt emotions and observed behaviour. Given the specific research setup and the

type of labelled data sought, none of the standard self-report instruments were found suitable (Isomursu, Tahti, Vainamo, & Kuutti, 2007). As such, self-annotation was implemented using an interval-based coding system through fixed-time slots. Subjects were prompted to rate their agreement on each of the pre-selected categories based on a Likert scale ranging from Strongly Agree to Strongly Disagree after every 20 seconds of elapsed video. A free-response option to allow subjective descriptions as well as the 'Other' option was also provided. Annotation was implemented to allow a split-screen viewing of recorded behaviour with the time synchronised interaction record obtained via screen capture to encourage context-sensitive judgment in a sequential manner. The idea was to retain the natural evolution of the behaviour and preserve the temporal dynamics of interaction. Figure 3.3-1 shows a snapshot of the annotation interface implemented as a stand-alone application using Visual Basic.NET.

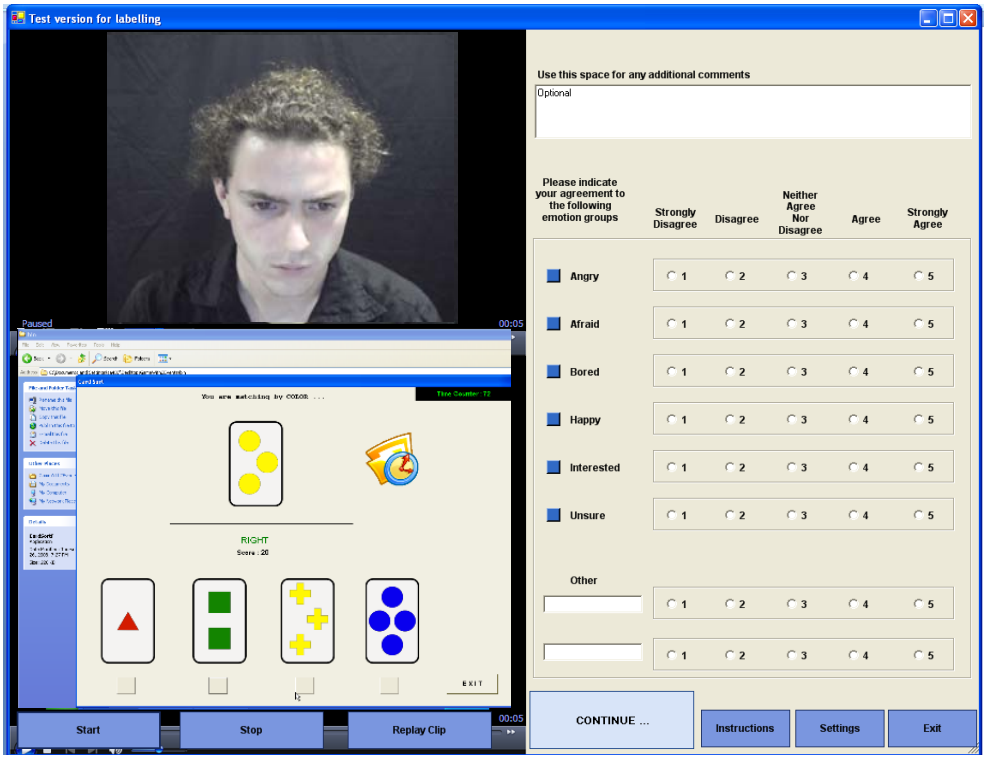


Figure 3.3-1: Snapshot of the interval based self-annotation

Results

The purpose of obtaining self-report was to get a subjective account of emotional behaviour. Observation of the labelling process however, indicated otherwise. Although participants responded differently to watching their own expressions - some surprised, mimicking and laughing at themselves, and others embarrassed, rushing through the video; the reactions did not suggest that they associated a subjective feeling with these but rather interpreted the expression as they might if it belonged to another in a social setting. This level of cognitive mediation was perceived as confounding the self-labelling purpose. It seemed that participants were more interested in 'watching' themselves and rushed through the coding part. They also complained that 20 seconds was a very tiny interval and that 'nothing major' was happening.

Three participants left the coding mid-way complaining of boredom. None of the participants had a problem with the annotation interface or the procedure itself but found watching themselves and 'creating' a meaning from their videos hard and uneventful. For these reasons, the self-annotation was considered unreliable and was discarded. Although the self-annotation was not successful in itself it helped re-assess certain choices in light of the data and shaped the next level of annotation:

- Emotional behaviour in the videos was subtle and gradual making interval-based coding extremely tedious. Deciding on an optimal time interval relative to the observed behaviour in such a case was difficult. The 20 second interval was chosen after trials with 5, 10 and 15 second intervals were not successful either. As such, switching to event-based coding was deemed appropriate for maximising the annotation value and effort.
- Although easy to use, interval-based coding artificially truncates behaviour resulting in information loss and arbitrary segments. As such, it would fail to account for emotional transitions occurring at the periphery of time intervals and depending on the frequency of such occurrences could severely affect the quality of training data required for machine learning. This further endorsed switching to event-based coding.
- Finally, labelling using multiple external raters was adopted to improve reliability of annotation. In the application context, this corresponds to taking a tutor-centric view.

3.3.2 Second Annotation

Design

Using event-based coding, the original videos were segmented into 105 non-neutral segments using ELAN[1]. ELAN is a free, multimodal annotation tool providing multi-layer video annotation features and complete control to an expert user for identifying and annotating segments of interest. It supports multileveled transcription and complies with output standards like XML and CSV which are helpful for exporting annotations for further analyses. ELAN also allows navigation through different time steps which is useful during behavioural coding. It supports user definable vocabularies for coding and provides easy navigation across the annotation levels. Figure 3.3-2 demonstrates the ELAN annotation using a trial session on one of the videos. A single application window gives powerful playback options along with flexible annotation modes to give an overall view of the annotation density. However, as with other video annotation tools, it is accompanied with a strong learning curve and requires considerable practice to achieve proficiency in use. It is therefore suitable only for an expert annotator, unless an appropriate level of training is provided.

As the eventual purpose was to compile a database of training clips, extraction of video segments corresponding to the transcribed annotations was required. The process of video extraction is time-consuming and computationally expensive when dealing with large amounts

[1] http://www.lat-mpi.eu/tools/elan/

of data. As such, the annotations or labels assigned in ELAN were exported to text files and processed using VirtualDub[2]. VirtualDub is a free video processing utility which is streamlined for fast linear operations over video and also allows batch processing. It provides a powerful and versatile scripting framework called Sylia which can be used to program the entire video extraction process efficiently using scripts. As such, the annotation files generated in ELAN were parsed to produce Sylia scripts which were then batch-processed in VirtualDub to produce the annotation based video clips. With input from Zuo (2007), Figure 3.3-3 outlines the steps in the video extraction process following the ELAN annotation.

The mean duration of extracted clips was 3.4 seconds ($\sigma = 2.5$), ranging from a minimum of 0.6 seconds to a maximum of 16 seconds. The segmentation was based on changes in the blanket expression where behaviour seemed consistent over a period of time. This essentially meant extracting portions of video that contained emotional behaviour as against portions with no observed changes (El Kaliouby & Teeters, 2007; Abrilian, et al., 2005). During the annotation, care was taken to preserve the temporal boundaries while demarcating the emotional segments. The manual annotation process followed by the corresponding automatic video extraction reduced the original video corpus of approximately 4 hours to less than 6 minutes at 30 fps. While this was a substantial gain in required annotation effort for subsequent labelling it highlighted how scantily the interaction was accompanied by changes in the observed visual modality.

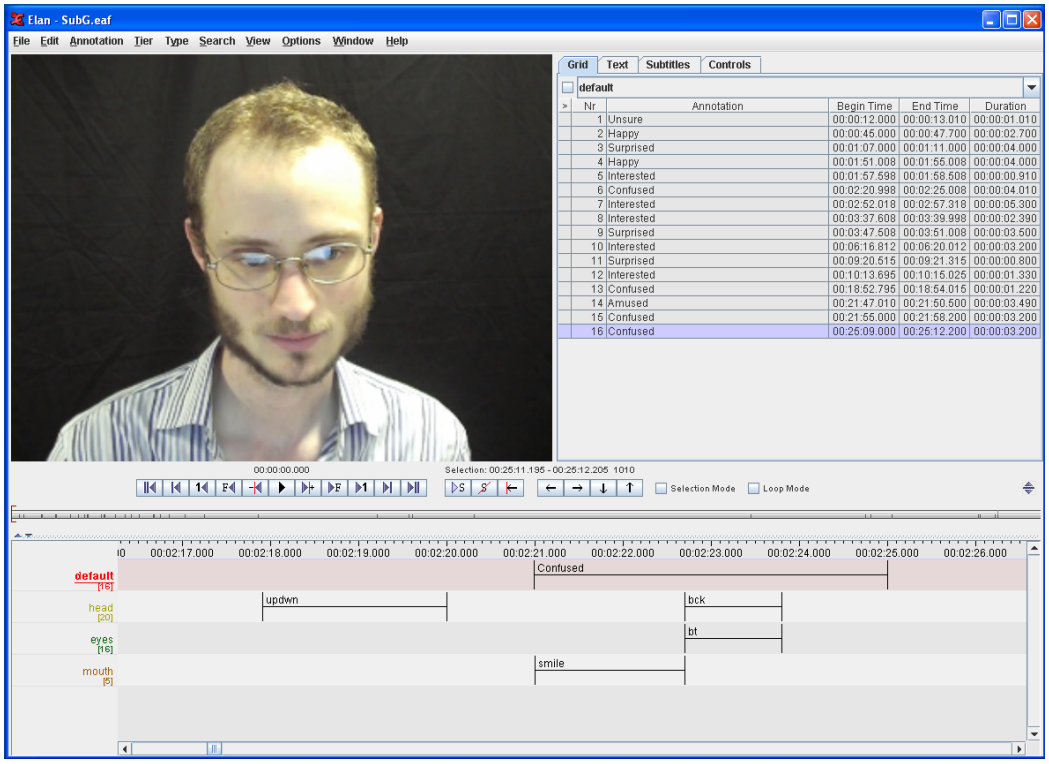


Figure 3.3-2: Snapshot of a test annotation session in ELAN

[2] http://www.virtualdub.org/

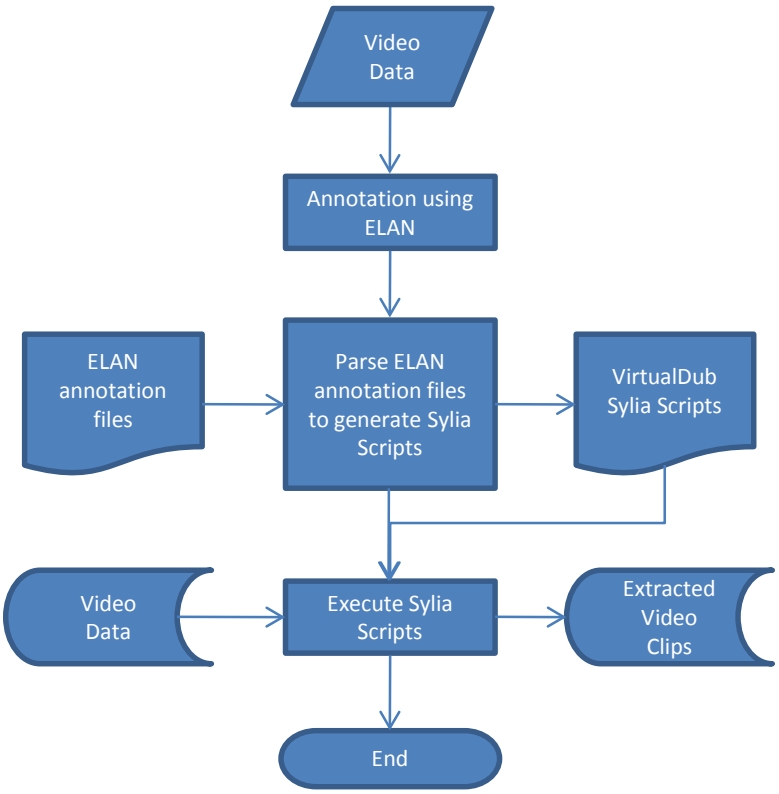


Figure 3.3-3: The video extraction process

Results

Three expert raters labelled the 105 pre-segmented clips independently. Raters could replay a video as many times as they wished. A primary and optional secondary emotion label was allowed for each video clip. Enforcing a simple majority rule resulted in 75% of videos getting classified into one of the pre-selected emotion categories. Table 3.3-1 (column A) summarises the distribution of emotion categories obtained this way when at least two out of the three raters agreed.

Taking primary labels into account, Fleiss' overall kappa was 0.35 indicating fair agreement. Agreement by chance was ruled out, but weakly. Given the low inter-rater reliability, the labelling results remained questionable. Moreover, the expert raters indicated that the video segments often displayed multiple emotions and that a second level of more intensive segmentation would improve judgement accuracy. A finer level of further segmentation was therefore done where segments corresponding to holistic expression changes were extracted. Unlike the first segmentation which was based on distinguishing emotional from non-emotional content, the focus now was to identify occurrences of sufficiently distinct emotional episodes. This meant demarcating the onset and offset of expression changes that provided enough context to be meaningful on their own. This increased the total number of video clips from 105 to 247. A third level annotation on these was designed, as described in the next section.

Amongst the emotion labels, no occurrence from the emotion group a*fraid* and related concepts like anxiety, nervousness, etc was found. Instead, s*urprised* which did not feature in the original selected affect groups was marked with a modest frequency. This was therefore included in the choice of affect descriptors in the next level of annotation. Also, on recommendation from the expert raters, the naming of the emotion categories was changed to more subtle and commonplace terms like replacing *unsure* with *confused* and *angry* with a*nnoyed*.

3.3.3 Third Annotation

Design

The corpus now consisted of 247 video clips with a mean duration of 2.8 seconds (σ = 1.86), ranging from a minimum of 0.4 seconds to a maximum of 16 seconds. An online labelling interface was set-up to facilitate access to a large number of raters. The coding scheme was modified so that for each video clip raters were required to mark the following: the emotion they attributed to the video, their confidence level (from 1-10) and whether they could perceive more than one emotion in the clip. The decision time for emotion judgement was computed as the time difference between the end of a sample video and when annotation was done. A video clip was played only once in order to get the initial reaction and to control the effects of replaying across raters. Initial reaction was preferred in line with evidence of the inverse relationship between accuracy of judgement and the time-taken to make a decision from facial expressions (Edwards, 1998). The focus at this level of annotation was to analyse emotion judgements from a large number of raters and improve annotation results. All raters underwent a training session before the actual labelling during which they were familiarised with the emotions taxonomy as well as the annotation interface. Fig 3.3-4 below shows snapshots from a typical labelling session.

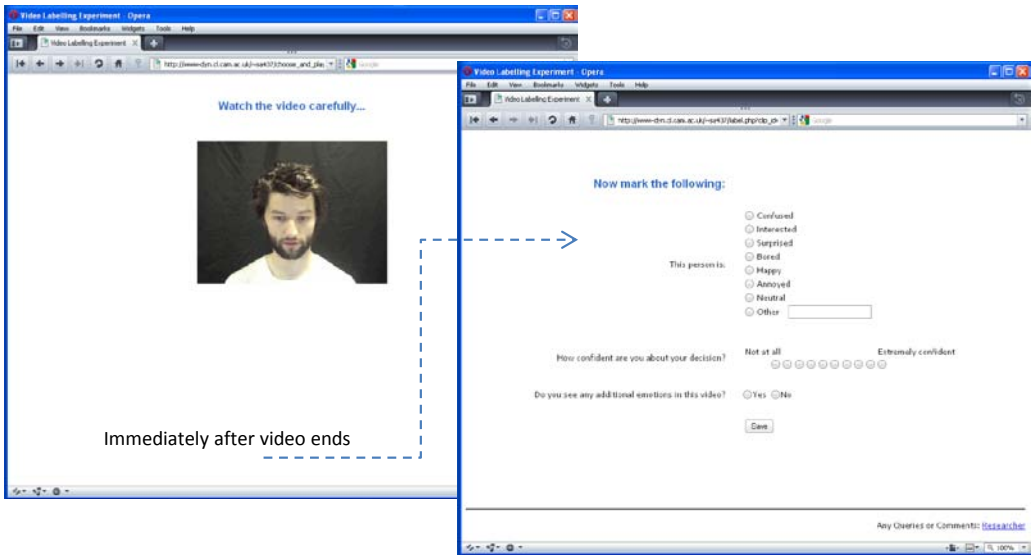


Figure 3.3-4: Snapshot from an online labelling session

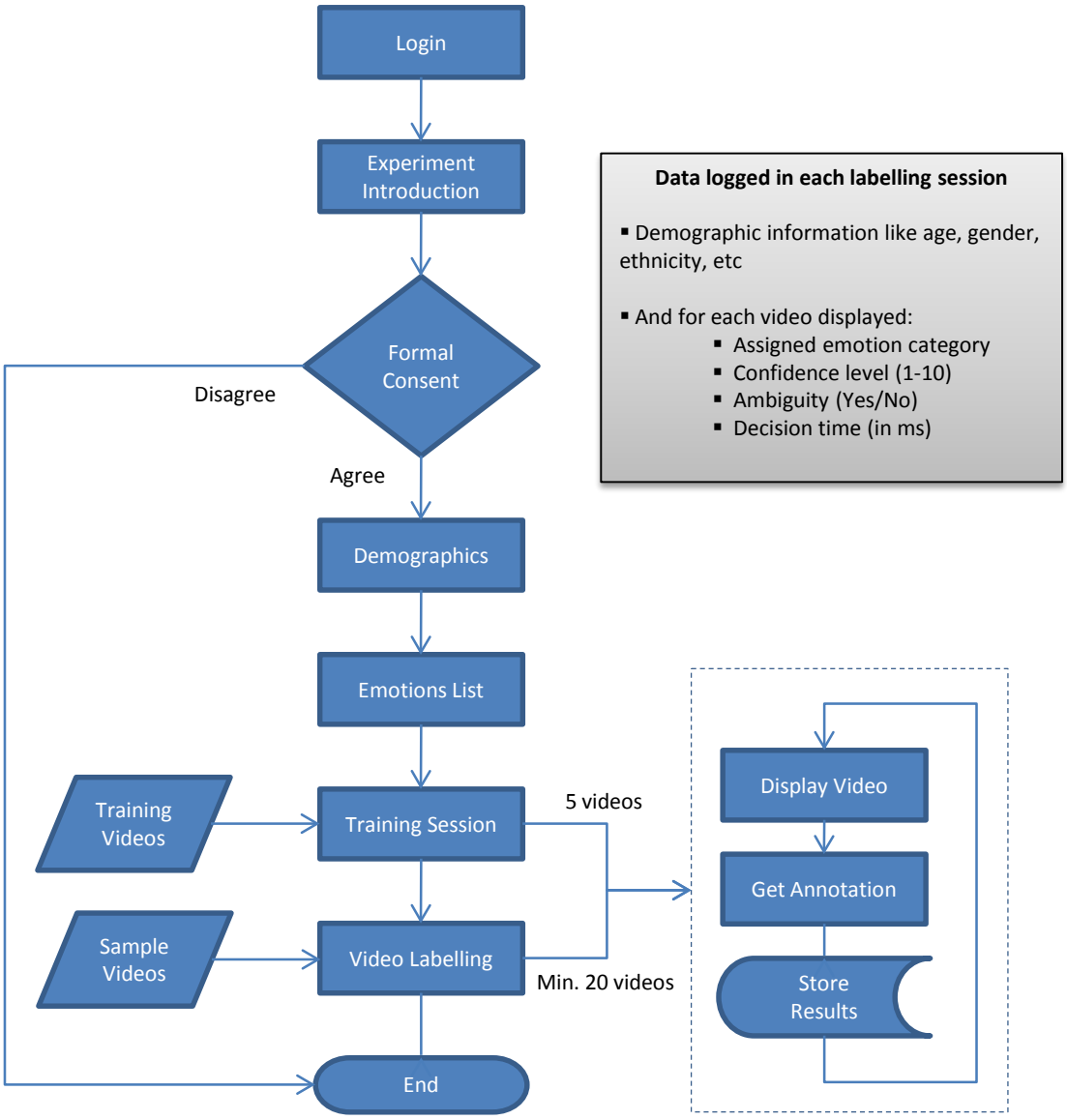


Figure 3.3-5: Outline of a labelling session

A flow-chart of how the online labelling proceeded is depicted further on in Figure 3.3-5. It also illustrates the information logged at the end of each labelling session. The interface was implemented using Perl scripting with MySQL as the back-end data store.

Results

108 raters, 39 male 69 female, signed up for the online study and coded an average of 20 videos each. They were aged between 18 and 56 years (μ = 28.28, σ = 6.20) and were of diverse ethnicities and background. A total of 2221 annotations were obtained so that each video was coded on average 8.99 times (σ = 0.13). Emotion labels present under *other* category were parsed using emotion taxonomies, GALC (Scherer, 2005) and Mind Reading (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) in order to group semantically similar terms into macro-classes. For example, pleased, amused, and enjoying, were grouped together under *happy*.

Table 3.3-1: Distribution of video clips across emotion categories

Annotation(s)	A		B	
	3 experts 105 clips		108 coders 247 clips	
	No.	% age	No.	% age
confused	26	24.8 %	73	29.6 %
interested	18	17.1 %	35	14.2 %
surprised	12	11.4 %	40	16.2 %
bored	5	4.8 %	19	7.7 %
happy	16	15.2 %	35	14.2 %
annoyed	0	0 %	13	5.3 %
neutral	3	2.9 %	29	11.7 %
other	25	23.9 %	3	1.2 %

Inter-rater reliability estimated using Fleiss' weighted kappa for multiple ratings per video with multiple raters (Fleiss, Levin, & Paik, 2003) was 0.20 overall, indicating slight agreement. Individual kappa agreements for the emotion categories are listed in Table 3.3-2. Only *happy* shows a good agreement while others show marginal kappa values. The lowest agreement was for videos labelled as *other* followed closely by *interested*. In fact, if we look at Figure 3.3-6 (left) showing decision times for the videos grouped by their labels, it appears that the videos classified as *happy* are indeed quicker to recognise than others. Videos classified as *other* show relatively longer decision times and as shown in Table 3.3-2, the lowest kappa as well. If decision time is construed as an indicator of difficulty in judgement, this suggests that inter-rater agreement is somewhat inversely related to difficulty in classification. In terms of duration, videos labelled as *surprised* were the shortest while those classified as *other* were the longest as shown in Figure 3.3-6 (right). No clear relationship was apparent between the duration and kappa's of the emotion categories though.

To note, the decision times and duration were both statistically significant across the emotion categories with $F(7, 667.61) = 7.89$, $p < .001$, $\omega = .24$ and $F(7, 29.42) = 2.53$, $p < .01$, $\omega = .21$, respectively, indicating modest effect sizes. As the Levene's Test for homogeneity of variances was significant indicating that group variances were not similar, Welch's F is reported here and used to correct the degrees of freedom (Field, 2009).

Table 3.3-2: Individual Fleiss' kappa scores for emotion categories

Emotion Group	Kappa	Agreement
confused	0.18	Slight
interested	0.09	Slight
surprised	0.20	Slight
bored	0.13	Slight
happy	**0.52**	**Moderate**
annoyed	0.10	Slight
neutral	0.17	Slight
other	0.05	Slight
Fleiss's (overall) kappa = 0.20 , p<.0001		

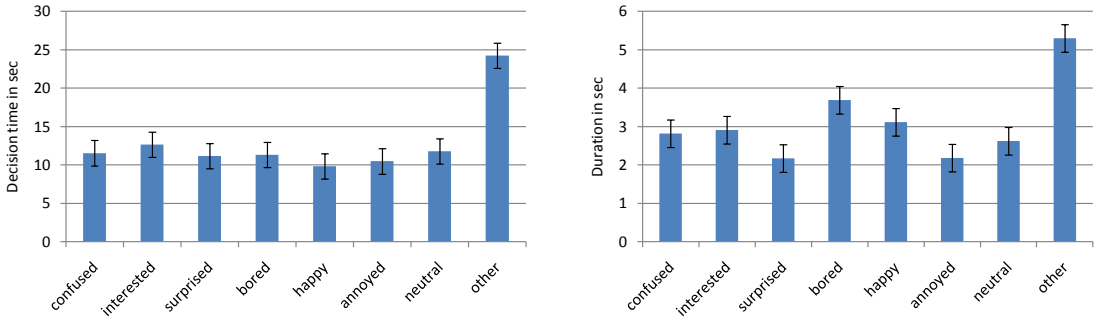


Figure 3.3-6: Decision time (left) and Duration (right) for videos labelled under the different emotion categories

3.3.4 Discussion

Having clearly labelled samples is a pre-requisite for designing automatic classifiers. The annotation process reveals that this is indeed very problematic to obtain from naturalistic data.

Self-rating

The attempted self-annotation was not successful in capturing a subjective emotional account as originally planned. The participants' reflection and meaning-making during the process appeared to influence their ability to record their actual emotional experience. Furthermore, their varied reactions of surprise, discomfort and even boredom suggested strong individual differences and mixed levels of emotional awareness. With respect to annotation, it was therefore difficult to ascertain if these would reflect the true emotional trace without being diluted with a level of rationalisation or cognitive mediation.

While this reiterates concerns about our ability to accurately provide a true account of emotional experience (Larsen & Fredrickson, 1999), it by no means discounts the merit of self-report as a method of assessment entirely. Self-report methodology has its advantages with respect to the ease with which data can be gathered, is cheap to apply and has high face validity. There are scenarios where self-rating can prove to be useful and there exist numerous studies that have effectively used self-report strategies and instruments to reliably capture subjective feelings. For example, Kapoor et al (2007) use a self–report button on the computer screen to allow participants to indicate when they are frustrated during a learning task. They then collect behavioural data leading up to each click and use it as an index to determine 'pre-frustration' in their recognition system. Another relevant example is the work by Whitehall et al. (2008) who use self-reported difficulty scores to predict difficulty in task material using facial expressions. D'Mello et al. (2008, 2007) use an emote-aloud procedure and post-hoc self-rating of emotions to analyse and contrast emotion judgements obtained from self, peers and teachers to understand relationships between their emotional accounts.

In general, self-report measures of emotion rely on the underlying assumption that participants are both able and willing to report on their emotional experience as well as on

their ability to accurately assess and express their emotions in some standardised format. A practical issue to consider is the meagre comprehension of semantic information in certain populations and differences owing to cultural, demographic, and contextual factors (Larsen and Fredrickson, 1999). The prevalence of trait-like deficits like Alexithymia (Nemiah, Freyberger, & Sifneos, 1976), which is marked by an impoverished cognitive processing of emotions resulting in a diminished ability to accurately identify and label emotional states, is a case in point.

In the context of a learning scenario, balancing the timing and quantity of measurement also needs consideration. Spontaneous access methods are likely to disrupt the learning task or influence the emotion(s) itself, while any retrospective emotional accounts will have issues of reliability (Schutz et al., 2006; Conati & Maclaren, 2004; Porayska-Pomsta & Pain, 2004). All self-report assessment methods have their specific limitations in terms of usage and context suitability and ultimately, it is the specific research view and user profile that would determine the best strategy to be incorporated. In an experimental evaluation of five different self-report methods for instance, Isomursu et al. (2007) successfully collected emotional responses evoked by mobile applications but finally proposed methodological triangulation as the best strategy. de Vicente and Pain (1998) arrived at the same conclusion when assessing the feasibility of self-report during assessment of students' motivation levels and recommend against relying exclusively on self-report. An extended discussion on these and related issues for accessing emotions in a learning environment is provided by Wosnitza and Volet (2005).

Segmentation

Segmenting the original videos into emotionally salient clips was the most challenging and time-consuming process. This is because identification of an emotional episode from a continuous video is immensely difficult at both an inferential as well as technical level. The judgement, whether of discriminating an emotionally salient segment from a baseline of what appears to be non-emotional behaviour, or of determining sufficiently distinct emotional episodes, is highly subjective and therefore likely to vary across coder(s) and time. Coders' own personality and disposition can make a significant difference in an interpretively complex task as this. Re-visiting the data, for example, often changes emotional judgements as familiarity habituates a rater to the range of facial signs in encoders. A renewed sense of understanding and perception occurs every time a video is replayed. The more familiar a face becomes, the more meaning you can discern from it. It is important therefore to acknowledge the inherent subjectivity in emotion perception and the likely variance in agreement on annotation labels. The low inter-rater reliability measures obtained during annotations only substantiate this.

To complicate things further, the actual process of determining the exact duration and onset of an emotional episode from the surrounding relatively neutral context to a peak and back to a relatively neutral state was tedious, yet noisy, and therefore, approximate at best. Demarcating the beginning and end of emotional expressions was incredibly hard as they often overlap, co-occur or blend subtly into a background expression. Sometimes it also appeared as if a single emotion persisted throughout the whole video and other expression changes were

noticeable on top of that. As an example, participant 1 appeared predominantly interested and engaged overall but displayed transient expression changes of other emotions in the foreground which made it extremely difficult to delineate an emotional episode. As a result, the extracted video-segments correspond to coders' subjective judgements about both the occurrence as well as the trajectory of emotional episodes and therefore cannot be considered as absolute exemplars. In retrospect, pre-segmentation of videos should ideally be validated by a second and if possible, more raters even though some noise is unavoidable because of the difficulty in marking precise boundaries and judging the exact onset, peak and offset of expressions.

Decoding Ability

Even for a human expert, it is difficult to define what constitutes an emotion. The whole process is unavoidably subjective and therefore dependent on the affect decoding skills and experience of raters. Figure 3.3-7 for example shows gender-wise results for confidence ratings, decision time in making emotion judgements, and marking of more than one emotion. Female raters appear to be more confident, arrive at judgements faster but perceive more than one emotion consistently. On average, female raters were more confident in their ratings (M = 7.54, SD = 0.16) than males (M = 7.42, SD = 0.25) and took less time to take a decision (M = 11.71, SD = 0.46) than males (M = 12.38, SD = 0.77). Although these differences are not statistically significant at $p < .05$, there is extensive evidence in nonverbal behaviour research showing that women are better than men in nonverbal decoding ability (Riggio & Riggio, 2005). As Elfenbein, Marsh and Ambady (2002) summarise from previous studies, "As early as three years of age, and across many cultures, females have a greater ability than males to perceive facial expressions of emotion...Psychologists have linked this finding to a wide range of other gender differences, including women's greater empathy, greater expressiveness, greater practice, greater tendency to accommodate others, greater breadth in using emotional information, and subordinate role in the larger culture" (Elfenbein, Marsh, & Ambady, 2002, p. 41).

Nonverbal Decoding Ability measures the accuracy of nonverbal cue processing and is a subset of interpersonal sensitivity. As such, quantitative affect decoding measures like the Profile of Nonverbal Sensitivity - PONS (Rosenthal, 1979), Communication of Affect Receiving Ability - CARAT (Buck, 1979), Empathic Accuracy (Ickes, 2001) or the Empathy Quotient (Baron-Cohen & Wheelwright, 2004) can be used to pre-screen annotators, indicate when training might be required as well as serve as reliability indicators for labelled data. Given the established individual differences in emotion judgement, inclusion of such measures might help improve and facilitate annotation of behavioural data.

Figure 3.3-7: Gender grouped annotation results

Ambiguity

Another factor that comes to fore from the annotation results is the prevalent ambiguity in emotion judgements. 38.7% of the total videos were perceived as containing more than one emotion. Female raters on average made higher use of this option than males (approx. 27% more) again emphasising heightened gender sensitivity to emotion perception. This is consistent with the findings of Abrilian et al. (2006) whose coding results on natural interview data also revealed that female coders perceived ambiguity in emotions 25% more than male coders. In general, the ambiguity in emotion perception shows that the occurrence of one emotion does not rule out the presence of another and an ideal automatic emotion inference system should be able to track co-occurring emotions.

During annotation itself, people find it difficult to articulate what they perceive in words. This is understandable because in everyday life emotion perception is rarely expressed in explicit terms and is subtly intertwined in social interactions. Consequently, raters often used a combination of labels and even phrases to express their judgements. The *other* category was liberally used during labelling which reveals the dependence of raters' active vocabulary on annotation. A possible alternative would be to balance free-form responses with fixed-choice alternatives in order to maximise accuracy while ensuring a degree of standardisation. Having taxonomies that allow parsing or mapping of free-form lexical emotion labels into different levels or groups of emotions would be of great help to standardise annotation results. Taxonomies like the GALC (Scherer, 2005) and Mind Reading (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) though not entirely comprehensive as yet, are good examples for this.

Finally, using multiple layers of annotation may help to reduce the subjectivity of annotations and get more convergent results. Abrilian et al's (2005) multi-level annotation framework is exemplary in that it combines emotion, context and multimodal annotations to overcome issues related to temporality and abstraction. However, as in any comprehensive coding technique, the coding-time, expertise and cost remain the main constraints.

3.4 The database

In general, databases can be organised into four types based on two underlying characteristics – spontaneity and degree of experimental control (El Kaliouby & Teeters, 2007). Spontaneity refers to the level of constraints imposed in the portrayal of emotions, while experimental control is related to the ecological validity of recorded behaviour. Figure 3.4-1 categorises some widely used corpora along these two continua. Ideally, a corpus should fall within the first quadrant that depicts naturalism with limited or no experimental control.

The database collected in this work (Afzal & Robinson, 2009) - CAL, together with the Belfast Naturalistic Database (Douglas-Cowie, 2004) signifies the recent shift from posed to more naturalistic and context-relevant data. Table 3.4-1 presents a detailed comparison of the collected database (CAL) against some widely used visual databases in terms of emotional range, level of naturalness, and other technical and contextual factors.

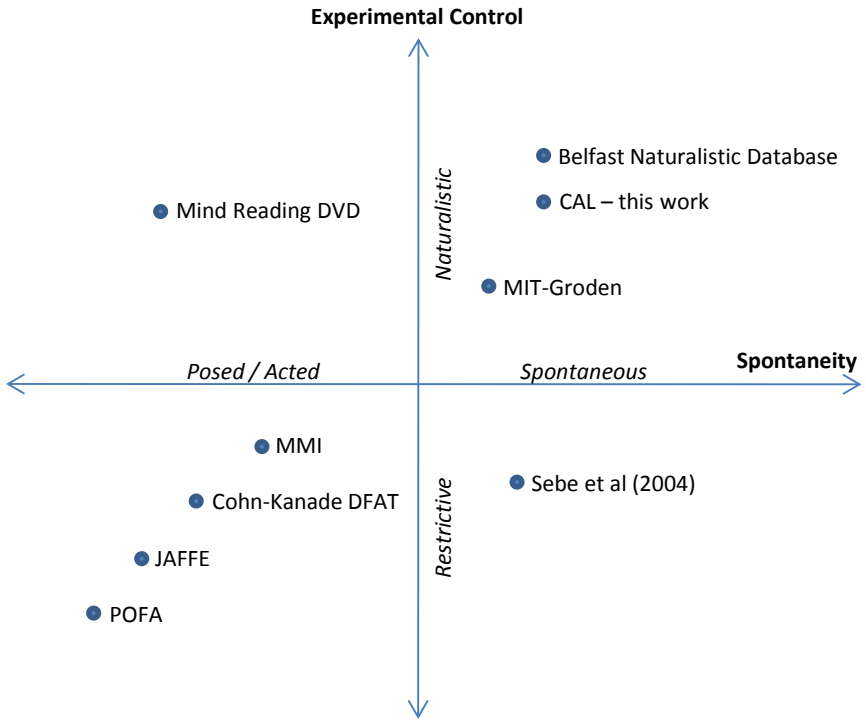


Figure 3.4-1: Categorisation of face corpora along spontaneity and experimental control

Table 3.4-1: Comparison of some common databases

Properties	Cohn-Kanade	MMI	Mind Reading DVD	Belfast Naturalistic	MIT-Groden-Autism	CAL (this work)
General						
Availability	Public / Licensed	Public / Licensed	Nominal fee	Licensed	Protected for privacy reasons	Protected for privacy reasons
Labelled	No	Yes	Yes	Yes	Yes	Yes
FACS-coded	Yes	Yes	No	No	No	No
Format	Downloadable	Web-based, Downloadable	DVD	CDs	-	CDs
Author(s)	(Kanade, Cohn, & Tian, 2000)	(Pantic, Valstar, Rademaker, & Maat, 2005)	(Baron-Cohen, Golan, Wheelwright, & Hill, 2004)	(Cowie, Douglas-Cowie, & Cox, 2005) (2004)	(El Kaliouby & Teeters, 2007)	(Afzal & Robinson, 2009)
Elicitation Method						
Spontaneity	Posed	Posed	Posed	Naturalistic	Induced	Naturalistic
Expt. Control	Directed	Directed	Unconstrained	Unconstrained	Unconstrained	Unconstrained
Scenario	Instructed by expert	Instructed by expert	Example scenarios	Television Interviews	Games & interaction scenarios	Learning environment
Context	Individual	Individual	Individual	Social Interactive	Social (Dyadic)	Standard HCI
Emotional Content						
Modalities	Visual	Visual	Visual	Audio-Visual	Visual	Visual
No. videos	2105	848	2742	239 sequences	2090	4 hrs original; 247 clips
Min-Max duration	0.3 - 2.0 sec	1.66 – 21.6 sec	5.0 - 8.0 sec	10 – 60 sec	? - 10.9 sec	0.4 - 15.9 sec
Resolution	640 x 480	720 x 576	320 x 240	-	-	320 x 240
Frame Rate	-	24 fps	30 fps	-	-	30fps
Lighting	Uniform	Uniform	Uniform	Variable	Variable	Variable
Pose	Frontal	Frontal + Profile	Frontal	Frontal	Frontal	Frontal
Initial frame	Neutral	Neutral	Non-neutral	Non-neutral	Non-neutral	Non-neutral
Rigid head motion	No	No	Yes	Yes	Yes	Yes
Occlusion	No	No	No	Yes	Yes	Yes
Talking	No	No	No	Yes	Yes	Yes
Encoders						
No. subjects	210	19	30	125	8	8
Gender	31 : 69	10 : 9	15 : 15	31 : 94	1 : 7	3 : 5
Age-group	18-50 yrs	19-62 yrs	16-60 yrs	-	18-20 yrs	21-32 yrs
Ethnicity	Diverse	Diverse	Diverse	-	-	Diverse
Glasses	No	Yes	No	-	Yes	Yes
Facial Hair	No	Yes	No	-	-	Yes
Labelling						
Coding Model	FACS	FACS	affect labels	affect labels; dimensions	affect labels	affect labels
No. coders	2 FACS experts	2 FACS experts	10	6	10 (pre-segmentation by an expert)	108 (pre-segmentation by an expert)
Inter-coder reliability	0.86 Cohen's kappa	Consensus	8/10 raw agreement	-	8/10 raw agreement	0.20 Fleiss' kappa
Emotional Content	Basic emotions: joy, surprise, anger, fear, disgust, sad	Basic emotions: joy, surprise, anger, fear, disgust, sad	412 affective-cognitive states	48 emotion words, valence, activation, intensity	agreeing, disagreeing, interested, anger, thinking, etc	confused, interested, surprised, bored, happy, annoyed, other
Context info	No	No	No	Yes	-	Yes

The CAL database is one of the first published naturalistic databases obtained in the target application scenario. It differs from works in the ITS community, for example D'Mello, Craig, Gholson, Franklin, Picard, & Graesser (2005), in terms of the learning context, the emphasis on the visual modality, the nature and temporal resolution of the annotation as well as by the intentional absence of 'agent-based adaptive tutoring'. The nature of emotional behaviour is intricately tied to the nature of interaction so that a non-adaptive, self-regulated learning task as used in this research, is likely to evoke emotional behaviour that is different from that evoked using a learning agent. Addition of any 'active' participation whether from a human or the computer, brings additional complexity while attributing emotions. By limiting the influence of emergent factors in a non-standard human-computer interaction task, my objective was to reduce the complexity in interpretation of emotional behaviour and simplify annotation. As discussed above, this turns out to be a massively challenging by itself.

3.5 Summary and conclusions

Emotion and expressivity have contextual significance so that if we adopt an application-oriented view, reliance on re-usable general databases is perhaps of limited value. For affect recognition technology to reliably operate in target applications, we need context-specific corpora to serve not only as repositories of sample data, but importantly to shape our understanding of the problem itself. This chapter has described one such attempt to capture naturalistic emotional data in a computer-based learning scenario. I have described the data collection process and the annotation framework in detail and have discussed important observations and results arising from these.

A self-regulated learning task was used to collect samples of emotional behaviour in an unconstrained setting. The data obtained went through three levels of annotation each giving a new insight into the nature of the problem. Six pre-selected domain relevant emotion groups were used during the annotation process and to interpret additional labelling results like duration and decision time. It was found that the main problems in annotation are derived from the dynamic nature of emotions, the ambiguity in categorisation and the high subjectivity of emotion perception. Accuracy measures were reported in terms of inter-rater reliability using Fleiss's kappa and were found to be quite low. The low inter-rater reliabilities, rather than being an error of measure as one could interpret, are in fact an acknowledged observation reported for naturalistic data and highlight the difficulty in ascribing emotions in real-life data (Abrilian, Devillers, Buisine, & Martin, 2005; Cowie, Douglas-Cowie, & Cox, 2005; Douglas-Cowie, et al., 2005). What is important is to reflect on how we can decide on an optimal metric of recognition accuracy for evaluating automatic classifiers, when we lack a reliable and objective ground-truth in the first place.

4. Facial Affect Analysis

Emotion perception is an innate skill that most people practice with a fair degree of accuracy in everyday lives. Yet defining what constitutes an emotion, and importantly, how we differentiate between emotions is incredibly hard to formalise. As soon as an objective and quantitative method of study is imposed on the process, several methodological problems arise, as found in Chapter 3. As we lack an explicit mapping from facial expressions to different emotions, automatic classifier design is a challenging task, more so for non-basic emotions like the group of affective-cognitive classes that are the focus of this work. In this chapter I consider the data collected and described in Chapter 3 and explore it statistically and structurally using several data mining techniques and machine learning. The aim is to study the signature of the emotion classes in spontaneous data and examine strategies for building an optimal emotion classification system. Considering that there are very few reported works on the machine level analysis of naturalistic data (Zeng, Pantic, Roisman, & Huang, 2009), the discussion here makes for an interesting and significant contribution to the field. Since automatic emotion inference is essentially a *pattern recognition* problem, I will first introduce some basic terms and methods as relevant to the analyses in this dissertation.

4.1 Pattern recognition

Machine perception of affect can be posed as a pattern recognition problem, typically classification or categorisation, where the classes or categories correspond to the different emotion groups. A pattern can be conceptualised as a quantitative or structural description of the entity of interest which is an emotion class in this case. Each pattern is represented in terms of d features or attributes as:

$$\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_d], \text{ where } x_1, x_2, x_3, \ldots, x_d \text{ are the features}$$

The d-dimensional feature vector thus characterises a pattern which can be discrete or continuous, depending on the problem domain and the type of input data. The feature vector is used to assign a given pattern to one of k specified categories based on a classification measure. Conventional classification measures include distance (Euclidean or Mahalanobis), likelihood and Bayesian *a posteriori* probability (Wang, 2002). An ideal feature set minimises the intra-class variation and maximises the inter-class variation while being invariant to irrelevant transformations of the input. Determining an optimal feature representation thus is crucial to overall classifier design.

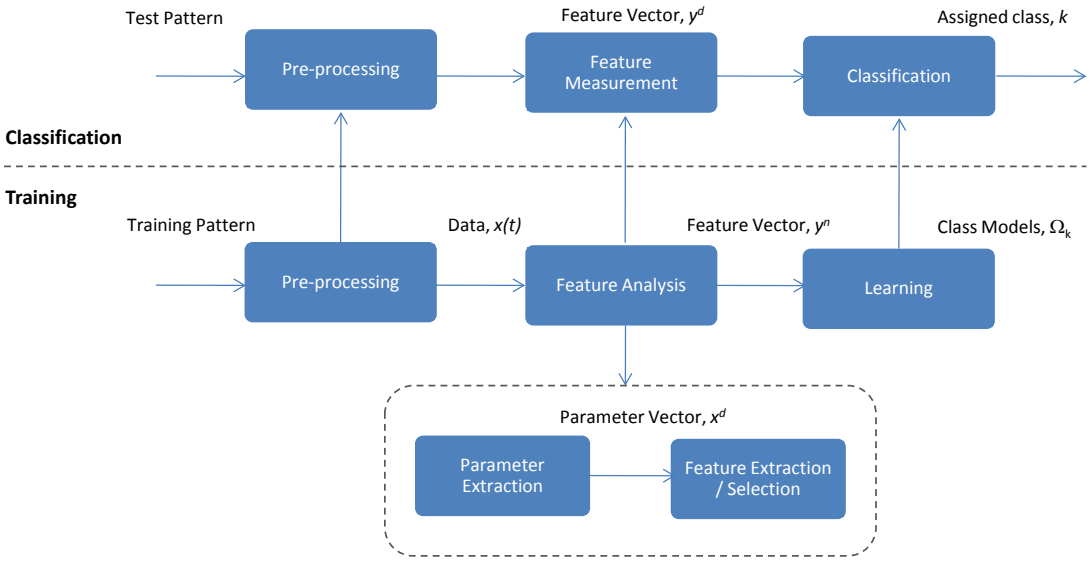


Figure 4.1-1: Overview of a typical pattern recognition system

Figure 4.1-1 summarises the data flow in a typical pattern recognition system which is operated in two modes: training (learning) and classification (testing). Pre-processing involves segmentation, removal of noise, normalisation and any other operation that contributes to defining a compact representation of the pattern. In the training mode, the feature analysis module finds the appropriate features for representing the input patterns and a classifier is trained to partition the feature space. Pre-processing and feature analysis are optimised to achieve the best classification performance. Finally, in classification mode the trained classifier assigns the input pattern to one of the classes based on the measured features and the learned class models (Duda, Hart, & Stork, 2001; Jain, Duin, & Mao, 2000).

The accuracy and complexity of a classifier is highly dependent on how well a set of features represents the input pattern and determining such a salient feature set is referred to as feature analysis. It is achieved in two steps: parameter extraction, and feature extraction or feature selection. In general, the term feature is often used in place of the term parameter, the distinction having more of an illustrative purpose here.

In the parameter extraction step, pattern information that is relevant to classification is extracted from the input data in the form of a d-dimensional parameter vector x. The original dimensionality of the parameter vector is generally high and needs to be reduced for the sake of computational cost and system complexity. In most cases, the parameter vector is not suitable for input to a classifier directly and requires some pre-processing. For example, parameter vectors need to be de-correlated before applying them for further classification (Wang, 2002).

In the feature extraction/selection step, the parameter vector x is transformed to a feature vector y, which has a dimensionality n, where $n \leq d$ and n is a subspace of the original feature set. Feature extraction differs from feature selection in that the former uses methods to create new features based on transformations or combinations of the original feature set as against

selection of the best feature subsets in the latter case. In practice, transformed features generated by feature extraction provide a better discriminative performance than the best subset of features (Jain, Duin, & Mao, 2000). Considering this and the constraint of a limited sample size relative to parameter dimensionality, I have used feature extraction as part of the feature analysis for refining the parameter vector.

The overview in Figure 4.1-1 helps frame the problem of emotion inference and will be used as a general template for the analyses performed and described in the following sections.

4.2 Data

The assumption of an unambiguous reference class is central to pattern classification. However, the inter-coder reliability results obtained during the manual annotation of data, as described in Chapter 3, were not convincing enough to accept the emotion annotations as the true class based simply on raw agreement. In order to enhance the reliability of final annotations, I adopted a weighted system of classification by using the coders' confidence level ratings obtained during the annotation procedure. This way emotion labels are assigned a weight equivalent to the coders' confidence level and the maximum weighted emotion label is then taken as the true label for a video clip. For example, a video clip coded as *happy* with confidence 9 by Coder 1, *confused* with confidence 1 by Coder 2, *happy* with confidence 7 by Coder 3, and *surprised* with confidence 9 by Coder 4; would be classified as *happy* since the total confidence weight for emotion *happy* is the highest (9+7).

Table 4.2-1 shows the final assignment of video clips to the emotion classes by applying the weighting rule over the entire set of 2221 available annotations. A maximum of about 30% of the videos were classified as *confused* as against the least proportion for *other* at 1%. This highlights the occurrence of *confusion* as a dominant emotion associated with learning followed by *surprised*, *interested* and *happy*. Only 8% of the videos were classified as *bored* which is not surprising considering the nature and duration of the experimental task used in data collection.

In order to compare the confidence-based allocation with the original annotations, Table 4.2-2 shows the proportion of raw annotations that agree with the emotion class assigned after using the weighting rule. Row i of the table indicates the proportion of raw annotations for class i while column i lists the proportion of confidence based category assignment for class i. The table shows for example that videos that are weight-assigned to category *confused* received raw annotations of the same class about 50% of the time even without factoring in the confidence levels. Similarly, *surprised* and *happy* show good agreements at approximately 55% and 66%, respectively. The remaining emotions also achieve agreement but at relatively moderate percentages. Overall, the diagonals dominate thus substantiating the emotion distribution obtained using confidence weighting.

Accordingly, the sample set depicted in Table 4.2-1 was taken as the *ground-truth* to be used in subsequent data exploration and machine analysis.

Table 4.2-1: Assignment of emotion labels using weighted confidence level ratings

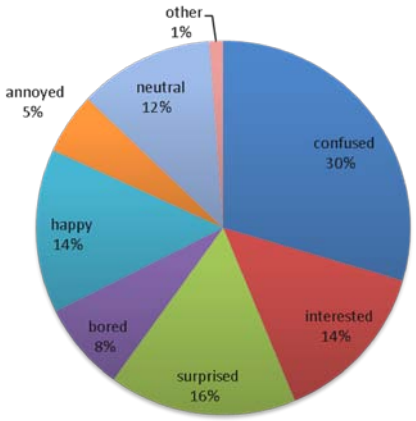
Emotion Class	Total No. of coders = 108 Total No. of videos, N = 247 Total No. of annotations = 2221	
	No. of samples	% age distribution of the original sample set
bored	19	
confused	73	
happy	35	
interested	35	
neutral	29	
surprised	40	
annoyed	13	
other	3	

Table 4.2-2: Proportional agreement of raw annotations with the weight-assigned emotion labels

Weight-based assignments →

← Raw assignments

	bored	confused	happy	interested	neutral	surprised	annoyed	other
bored	**40.94**	11.70	2.92	12.28	15.20	4.68	11.11	1.17
confused	4.41	**49.92**	0.91	9.44	6.24	12.94	10.35	5.78
happy	0.96	3.50	**66.24**	10.51	3.82	11.46	2.23	1.27
interested	6.67	13.97	2.86	**44.13**	12.38	13.97	3.81	2.22
neutral	12.26	7.66	3.45	14.94	**47.51**	7.66	3.07	3.45
surprised	4.74	8.64	5.01	11.14	8.64	**55.43**	3.90	2.51
annoyed	9.40	24.79	0.00	2.56	7.69	13.68	**38.46**	3.42
other	11.11	22.22	0.00	11.11	3.70	7.41	3.70	**40.74**

A basic visual analysis of the sample data was carried out to determine if there were any observable patterns that broadly mapped onto the emotion classes. For each video, a unit increase in activity score was made if a perceptible change occurred in three broad regions: upper-face, lower-face and head actions. The results are depicted in Figure 4.2-1 where the graphs display the proportion of activity observed in each emotion class. The y-axis shows the number of times activity was observed in videos belonging to each emotion class. Classes where the proportion exceeds a hundred percent indicates that a single video contributed to more than one observation. For example, the *other* category clearly shows a lot of overall activity which may in fact explain its membership status. Recall here from Figure 3.3-6 that videos categorised as *other* corresponded to the longest decision times as well as duration.

This confirms that these videos contained complex visual changes making their categorisation difficult.

Although this is only an indicative graph showing noticeable changes in the dominant visual channel, it affirms our understanding of the contribution of vital cues in emotion perception. For example, smiling is associated with happiness and this is reflected clearly with a higher activity in the lower face region.

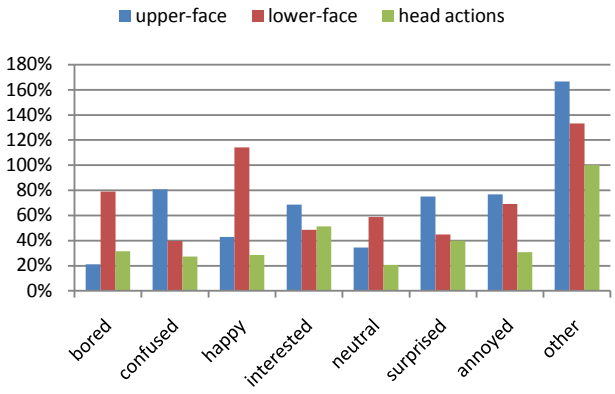


Figure 4.2-1: Perceived visual activity across emotion classes

Surprised and confusion on the other hand show distinguishing activity in the upper-face pointing to familiar cues like raised eyebrows or frowning linked to these states. These observations are consistent with previous findings (Nusseck, Cunningham, Wallraven, & Bulthoff, 2008) and I will revisit the implications of such regional contributions when defining feature descriptors for automatic emotion inference. Note that the emotion groups of *annoyed* and *other* have very few representative samples to merit proper statistical analyses and will therefore not be discussed further. The rationale for this will become clear when dealing with concepts of dimensionality and generalisation.

4.3 Feature analysis

Defining features implies developing a representation of the input pattern that can facilitate classification. Domain knowledge and human instinct play an important role in identifying such descriptors. Although a large body of work dealing with human perception of facial expressions exists, there have been very few attempts to develop objective methods for quantifying facial movements (Essa, 1997). One of the most significant works in this area is that of Ekman & Friesen (Ekman & Friesen, 1978) who have devised a system for objectively coding all visually distinguishable facial movements called the *Facial Action Coding System* (FACS).

FACS associates facial expression changes with the actions of the muscles that produce them and by enumerating 44 action units (AUs) it encodes all anatomically possible facial expressions, singly or in combination. Since the AUs are purely descriptive measures of facial expression changes, they are independent of interpretation and provide a useful grammar for use as feature descriptors in expression studies as this. Although a well-known limitation of FACS is its sheer complexity and disregard for temporal information, it still remains a popular method for measuring facial behaviour and continues to have normative significance in automatic facial expression analysis. It is also the only psychometrically rigorous and comprehensive grammar of facial actions that is available (Cohn, 2006). Tables 4.3-1 and 4.3-2 illustrate some of the upper and lower face AUs encoded in FACS.

4.3.1 Representation and measurement of facial motion

To characterise the facial motion, I used the 2D face model of the Nevenvision FaceTracker[3] depicted in Figure 4.3-3. This FaceTracker is a state-of-art facial feature point tracking technology that requires no manual pre-processing or calibration. It is resilient to limited out-of-plane motion, can deal with a wide range of physiognomies and can also track faces with glasses or facial hair. It enables fully automatic unobtrusive feature point tracking making it attractive for real-time applications.

Feature point tracking estimates facial motion by tracking the movement and deformation of markers or feature points placed on the prominent intransient facial features (Fasel & Luettin, 2003). The FaceTracker uses a generic face template to capture the movement of 22 fiducial points on the face over the sample video sequences. The displacement of these feature points over successive frames encodes the motion pattern of the AUs. Horizontal and vertical distances as shown in Figure 4.3-3 are then computed to obtain a distance vector. To remove the effects of variation in face scale and projection, the distance measurements are normalised with respect to a positional line connecting the inner eyes in the first frame.

Statistically, the representative values of AUs in terms of local concentration (median) and dispersion (standard deviation) are selected as parameters, along with the first temporal derivative corresponding to speed as an additional attribute. The inclusion of speed helps qualify the dynamic information in expression changes and is found to increase the interpretive power and performance of classifiers (Tong, Liao, & Ji, 2007; Pantic & Patras, 2006; Ambadar, Schooler, & Cohn, 2005). The measurements used to construct the final parameter vector are depicted in Table 4.3-1. As the dimensionality of the parameter vector relative to the sample size is unrealistic for designing an efficient classifier, a reduced set of significant features needs to be extracted from the aforementioned parameter set.

4.3.2 Feature extraction

Principal Component Analysis (PCA), also known as the *Karhunen-Loeve Transform* (Pearson, 1901; Hotelling, 1933), is one of the best known linear feature extractor used in dimensionality reduction. Given a data set consisting of a number of interrelated variables, the central idea of PCA is to reduce its dimensionality while retaining as much of the variation present in the original data as possible. With PCA, the data is transformed into a new set of variables that are a linear combination of the original ones. The transformed variables in the new principal component space are less strongly correlated and therefore relatively free of redundant information.

[3] Licensed from Google Inc.

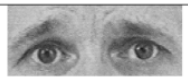
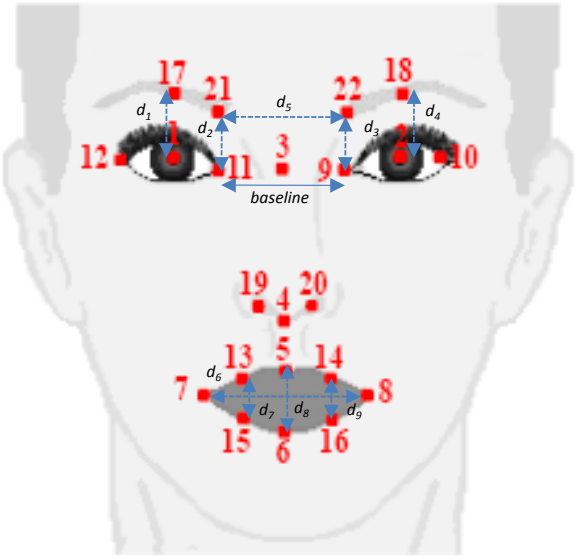
NEUTRAL	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
				
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
				
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 4.3-1: Upper face AUs and some combinations (from Tian, Kanade, & Cohn, 2001)

NEUTRAL	AU 9	AU 10	AU 12	AU 20
				
Lips relaxed and closed.	The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent.	Lip corners are pulled obliquely.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU15	AU 17	AU 25	AU 26	AU 27
				
The corners of the lips are pulled down.	The chin boss is pushed upwards.	Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched open and the mandible pulled downwards.
AU 23+24	AU 9+17	AU9+25	AU9+17+23+24	AU10+17
				
Lips tightened, narrowed, and pressed together.				
AU 10+25	AU 10+15+17	AU 12+25	AU12+26	AU 15+17
				
AU 17+23+24	AU 20+25			
				

Figure 4.3-2: Lower face AUs and some combinations (from Tian, Kanade, & Cohn, 2001)



1. Right Pupil
2. Left Pupil
3. Nose Root
4. Nose Tip
5. Upper Lip Centre
6. Lower Lip Centre
7. Right Mouth Corner
8. Left Mouth Corner
9. Left Inner Eye Corner
10. Left Outer Eye Corner
11. Right Inner Eye Corner
12. Right Outer Eye Corner
13. Right Upper Lip
14. Left Upper Lip
15. Right Lower Lip
16. Left Lower Lip
17. Right Eyebrow Centre
18. Left Eyebrow Centre
19. Right Nostril
20. Left Nostril
21. Right Inner Eyebrow
22. Left Inner Eyebrow

Figure 4.3-3: The 2D face model used by the FaceTracker to track 22 facial feature points

Table 4.3-1: Components of the parameter vector

<table>
<tr><th colspan="2">Non-rigid motion measurements
Face</th><th>AU descriptors $[t-1, t]$</th><th>Statistical measures</th></tr>
<tr><td>Positional</td><td>Dynamic</td><td></td><td></td></tr>
<tr><td>$d_1 = \overline{P_1P_{17}}$</td><td rowspan="9">$\frac{dx}{dt}$ - speed</td><td rowspan="4">1, 2 , 1+2</td><td rowspan="15">Median and standard deviation</td></tr>
<tr><td>$d_2 = \overline{P_{11}P_{21}}$</td></tr>
<tr><td>$d_3 = \overline{P_9P_{22}}$</td></tr>
<tr><td>$d_4 = \overline{P_2P_{18}}$</td></tr>
<tr><td>$d_5 = \overline{P_{21}P_{22}}$</td><td>4</td></tr>
<tr><td>$d_6 = \overline{P_7P_8}$</td><td rowspan="4">12, 18, 20, 23, 24, 25, 26, 27, 28</td></tr>
<tr><td>$d_7 = \overline{P_{13}P_{15}}$</td></tr>
<tr><td>$d_8 = \overline{P_5P_6}$</td></tr>
<tr><td>$d_9 = \overline{P_{14}P_{16}}$</td></tr>
<tr><td colspan="2">Rigid motion measurements
Head</td><td>AU descriptors $[t-1, t]$</td></tr>
<tr><td>Positional</td><td>Dynamic</td><td></td></tr>
<tr><td>Euler x - pitch</td><td rowspan="3">$\frac{d\theta}{dt}$ - speed</td><td>53, 54</td></tr>
<tr><td>Euler y - yaw</td><td>51, 52</td></tr>
<tr><td>Euler z - roll</td><td>55, 56</td></tr>
</table>

In mathematical terms, PCA projects the original d-dimensional parameter vector x_k to a reduced feature space as y_k. The projected n-dimensional feature vector $y_k = A^T x_k$ where $A \in R^{d*n}$ is the transformation matrix consisting of orthonormal eigenvectors of the total scatter matrix S_T defined as $S_T = \sum_{k=1}^{N}(x_k - \mu)(x_k - \mu)^T$, corresponding to the n largest eigenvalues (Chen & Huang, 2003; Turk & Pentland, 1991). PCA effectively approximates the data by a linear subspace using the mean squared error criterion. The generated features thus capture the main scatter directions and form an optimal representation in a reduced space (Jain, Duin, & Mao, 2000).

I explored two groupings of the original parameter space for feature extraction: region-based and collinearity-based. In region-based feature grouping I used PCA on parameters grouped from anthropometrically meaningful regions. This produced a reduced feature space characterising the changes in prominent regions corresponding to the right eye, the left eye, between the eyes, and the vertical and horizontal mouth. In the face model of Figure 4.3-3, this corresponds to combining the positional and dynamic features over d_1, d_2; d_3, d_4; d_5; d_6; and d_7, d_8, d_9, respectively. The motivation for this was to approximate the visual changes in these canonical regions in terms of an expressive feature vector and evaluate its efficiency in prediction of affect states. In the second grouping, I performed feature extraction to remove multi-collinearity by doing a PCA on highly correlated parameters (where $\rho > 0.8$). The aim was to reduce feature dimensionality by combining the variance from highly correlated parameters in a single weighted vector thereby eliminating any significant redundancies. Some correlations of interest were those found between the parameters d_1 and d_4, d_2 and d_3, and, d_7 and d_9, indicating a degree of structural symmetry in facial expressions.

In both groupings I used the percent variability to decide on the number of Principal Components (PCs) to retain. This allows selecting those n PCs that contribute a significant cumulative percentage of total variation in the data, and is calculated as $t_d = 100 \sum_{i=1}^{n} l_i \div \sum_{j=1}^{d} l_j$ (Martinez & Martinez, 2005). In my analyses the first PC is dominant in explaining the total variation ($t_d > 90\%$) and is therefore retained as the representative feature descriptor.

Finally, each feature set derived from the two transformation groupings was sub-divided into two variants depending on whether or not head motion parameters were included. This was done to examine the effect of head motion independently on classifier performance. The result of feature extraction at this stage thus produced four feature sets F1, F2, F3 and F4 as listed in Table 4.3-2 along with their dimensionalities. These were used as feature descriptors of the sample emotion classes in subsequent data exploration and analyses.

Table 4.3-2: Feature sets used for analysis

Feature-Set	Description
F1 (n = 16)	De-correlated parameters
F2 (n = 28)	F1 including rigid motion (head roll, pitch and yaw)
F3 (n = 5)	Region-based parameters
F4 (n = 17)	F3 including rigid motion (head roll, pitch and yaw)

4.4 Visualising the problem space

To examine the structure of the underlying sample data, I used two exploratory data analysis techniques: clustering and discriminant analysis. Clustering was performed to uncover any natural groupings in the sample data while discriminant analysis was applied to view the emotion classes in a reduced dimensional space. The primary objective in both cases was to examine the class separability in order to get an estimate of classification complexity.

4.4.1 Unsupervised clustering

Clustering is the unsupervised classification of patterns into groups or clusters in a way such that observations within a cluster are more similar to each other than they are to observations belonging to a different cluster (Jain, Murty, & Flynn, 1999; Martinez & Martinez, 2005). The concept of s*imilarity* is thus fundamental to the definition of a cluster and is defined in terms of a distance measure which is usually Euclidean for continuous features. Based on how the data is grouped, clustering can be hierarchical or partitional, exclusive or overlapping, and complete or partial.

I used agglomerative hierarchical clustering which is a popular method used to group data in a hierarchy or set of nested partitions according to an optimality criterion. The term agglomerative signifies initialisation by using each data observation as a singleton cluster and repeatedly merging the two closest clusters until a single, all encompassing cluster remains. Closeness is defined by the linkage method used to compute the inter-point distances and can be single, complete, average, centroid or Ward's. The Cophenetic Correlation Coefficient (CPCC) is used to evaluate which of these linkage techniques best fits the data by comparing the fusion level of observations with their proximity. Using the feature sets listed in Table 4.3-2, clustering using average linkage consistently gave the best CPCC value (> 0.85) than those obtained using other linkage methods. Average linkage defines the distance between clusters as the average distance from all observations in one cluster to all points in another cluster. It tends to combine clusters that have small variances, and also tends to produce clusters with approximately equal variance.

The results of hierarchical clustering can be visualised graphically using a *dendrogram.* A dendrogram is a tree-like diagram that displays both the cluster-sub cluster relationships as well as the order in which the clusters are merged. Figure 4.4-1 (left) shows the dendrogram produced by performing agglomerative hierarchical clustering of the sample data compiled in the previous Chapter and summarised in Table 4.2-1. The horizontal axis plots the sample data points which are collapsed here for clarity. Fusion levels are indicated along the vertical axis so that cutting the dendrogram at a given height results in a set of clusters separated by at least the corresponding distance. Thus the numerical value on the vertical scale represents the distance/dissimilarity between clusters at a given level.

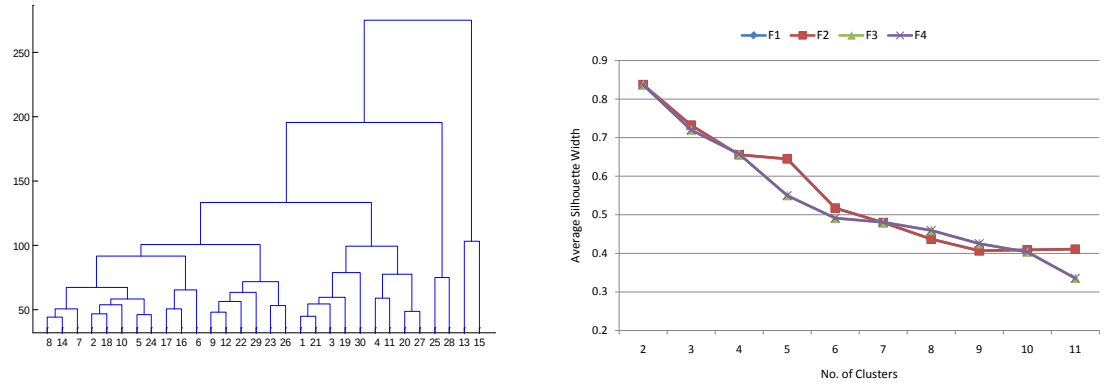


Figure 4.4-1: (left) Dendrogram obtained using agglomerative hierarchical clustering; (right) Corresponding silhouette curve

It appears from the corresponding *silhouette curve* in Figure 4.4-1 (right) that there are about four to six natural clusters that can be identified based on distinctness and compactness. Although these are at low values of similarity, this is promising given that the data is originally derived from six emotion classes and shows that the feature sets are in fact able to uncover groups in the data. The silhouette curve is derived using a cluster evaluation measure called *average silhouette width* in order to get a quantitative estimate of the goodness of clustering for different cluster sizes. This measure will be used further in the analysis and merits a brief description here.

For the ith observation, let a_i denote the average dissimilarity to all other points in its own cluster. For any other cluster c, let $\bar{d}(i, c)$ represent the average dissimilarity of i to all objects in c. Further, let b_i denote the minimum of these average dissimilarities $\bar{d}(i, c)$. Then the *silhouette coefficient or width* for the ith observation is $sw_i = (b_i - a_i)/\max(a_i, b_i)$ with values varying between -1 and 1. Ideally, the coefficient should be positive ($a_i < b_i$), and a_i be as close to 0 as possible since the coefficient assumes its maximum values of 1 when $a_i = 0$. The *average silhouette coefficient* of a cluster is finally computed by averaging the silhouette coefficients sw_i over all observations belonging to the cluster as $\overline{sw} = \frac{1}{n} \sum_{i=1}^{n} sw_i$ (Tan, Steinbach, & Kumar, 2006). As a rule of thumb, an average silhouette width greater than 0.5 indicates a reasonable partition of the data while a value of less than 0.2 exhibits no cluster structure (Kaufman & Rousseeuw, 1990).

By varying the cluster size from 2 through to 11 and determining the corresponding average silhouette widths one can obtain a *silhouette curve* to show how well the data groups under various cluster sizes. It is apparent from Figure 4.4-1 (right) that increasing the number of clusters radically decreases the quality of clustering in a way that the number of optimal clusters in data fades out quickly after cluster size 6. Amongst the feature sets, F2 appears to show the best overall silhouette width until cluster size 6. Of course, only a comparison with the original class memberships can judge how well these data groupings actually correspond to the original six emotion classes. This entails using a supervised classification method like discriminant analysis.

4.4.2 Multiple discriminant analysis

Multiple Discriminant Analysis (MDA) is a supervised classification method that makes explicit use of class information to extract the most discriminatory features while ensuring that the classes are maximally separated in a transformed representation. Mathematically, the objective is to find the transformation matrix that maximises the ratio between the between-class and within-class scatter matrix of the projected data. This corresponds to emphasising the interclass separation by finding the eigenvectors of $S_{\mathrm{w}}{}^{-1}S_{\mathrm{b}}$ where S_{w} is the within-class scatter matrix and S_{b} is the between-class scatter matrix.

I used MDA to assess the class separability in the original sample data by projecting the multi-dimensional feature representation onto a 2-dimensional space. Figure 4.4-2 shows the results of MDA using the first two retained dimensions. The colours specify the original emotion class memberships and in the 2D projection space allow a comparison of the class discrimination. The MDA visualisation indicated a great degree of overlap in the classes suggesting that discriminating emotion groups in the sample data was going to be hard. Feature set F2 allowed the best cluster cohesion and a minimum error of 0.37 but this is still not an ideal separation between the emotion classes. The MDA plots give an overall picture of classification complexity but ought to be interpreted with care as the reduced 2D view is likely to have suppressed potentially important discriminant information. To evaluate this formally, I will apply different classification methods as described in the following sections.

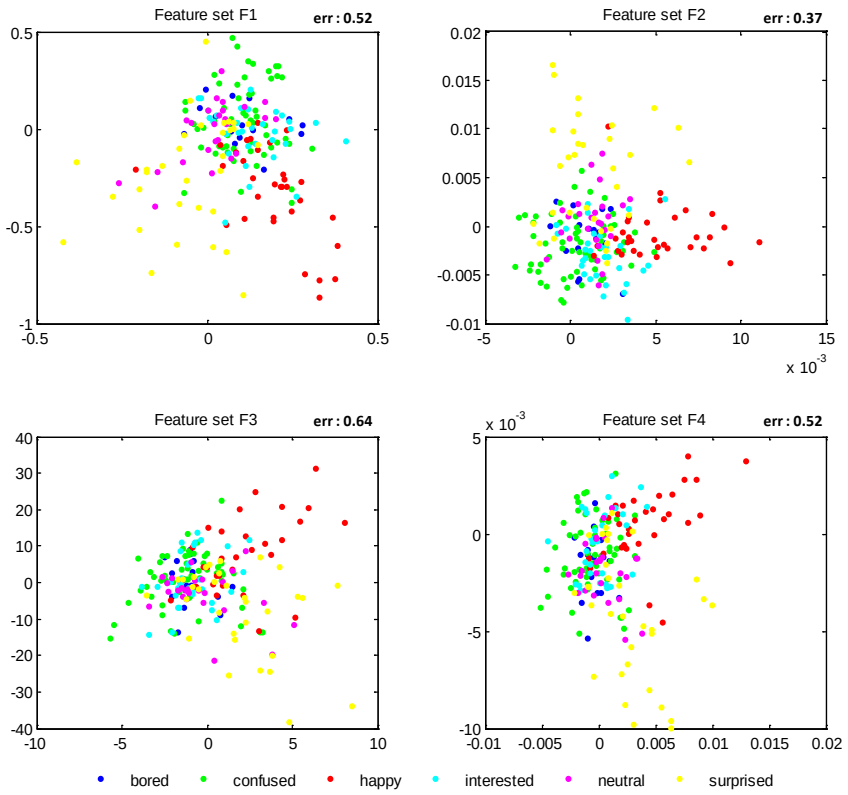


Figure 4.4-2: MDA plots using the four feature-sets

4.5 Learning and classification

I used the Waikato Environment for Knowledge Analysis (WEKA) (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) to compare the performance of algorithms from some standard classification schemes, as in other works (D'Mello, Picard, & Graesser, 2007). These are: Naïve Bayes for Bayesian classification, Multilayer Perceptron for Neural Networks, Nearest Neighbour for lazy classification, C4.5 for decision trees and additive logistic boosting for meta-classification schemes. Since this is essentially a six way classification choice the base chance rate is ~16.6%. For each method, the sample data characterised by the four feature sets described in Table 4.3-2 was used to learn and test classifiers for the affect class models. Classification accuracies were computed using stratified k-fold cross-validation $(k = 10)$ wherein the available samples N are divided into k disjoint subsets using random sampling. The classifier is then trained on the $(k - 1)$ subsets and tested on the remaining. Stratification ensures that each class is represented in approximately equal proportions in the training and test sets. Overall accuracy is finally determined by averaging the true positive rate over k process iterations.

Table 4.5-1 shows the classification results obtained. The best accuracies for each feature set are highlighted and further detailed out with class-level classification results in Table 4.5-2. About 35 sample videos had to be eliminated from the dataset as the FaceTracker failed to initialise and track the faces in them. This reduced the sample set to $N = 196$.

Table 4.5-1: Classification accuracies in percentage for each of the feature sets

Classification Scheme	Algorithm	F1	F2	F3	F4
Bayesian	Naïve Bayes	30.6	34.9	35.2	34.9
Neural Network	Multilayer Perceptron	35.2	35.7	**41.3**	35.2
Lazy	Nearest Neighbour	29.1	34.7	26.5	32.1
Decision Tree	J48 /C4.5	36.7	33.7	28.6	36.2
Meta-classification	Additive Logistic Regression	**42.9**	**38.3**	37.8	**36.2**

The results indicated that classification by regression and Multilayer Perceptron showed the best and yet, moderate accuracies relative to the base chance rate of 16.6%. This was somewhat expected given that the within-class vs. between-class differentiation was not very prominent as visualised with MDA. Furthermore, the performance of classifiers did not appear to be affected drastically by the feature set used. This is interesting because the dimensionality of the feature sets is different and suggests that an equivalent accuracy can be achieved by a smaller but salient feature set. For example, feature set F3 comprising just 5 region-based features gave the second best success rate of 41.3% as compared to the best classification of 42.9% achieved by feature set F1 of dimensionality 16.

Table 4.5-2 shows some detailed results of the best performing classifier and feature set combinations. The kappa coefficients indicate the agreement of prediction with the true class here and are comparable to the inter-rater agreement scores obtained during manual annotation of ~0.2. To recall, kappa quantifies the chance-corrected inter-rater agreement

with possible values ranging from +1 indicating perfect agreement to -1 indicating complete disagreement.

Table 4.5-2 also lists the F-measures for each emotion class. An *F-measure* or *F-score* combines the *precision* and *recall* (or the true positive rate) in a single metric of performance and is used here to assess the classification accuracy individually for the emotion classes. It is computed as the harmonic mean of precision and recall as:

$$F = 2 * \frac{precision * recall}{precision + recall}$$

where,

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall\ (true\ positive\ rate) = \frac{true\ positives}{true\ positives + false\ negatives}$$

A clear trend is evident where *confused*, *happy* and *surprised* show higher F-scores consistently across all feature sets indicating their overall reliable detection. Recall that *happy* and *surprised* are in fact a part of the basic-emotion set even though their definition here has a much broader scope. The emotion classes of *bored*, *neutral* and *interested* on the other hand show very low F-measures implying difficulty in discrimination. In fact, removing samples of these classes and performing classification as a three-way choice between *confused*, *happy* and *surprised* boosts the overall accuracies by *at least 25%* across all feature-sets reaching a maximum of 76% with F2. From the perspective of classifier design this reveals how the selection of emotion classes in itself can radically influence the overall recognition figures. In general, it appears that the information captured through facial and head gestural cues for bored, interested and neutral is insufficient to properly distinguish them from other emotion groups under consideration.

Table 4.5-2: Expanded results for the best classifiers

Feature-set	F1	F2	F3	F4
Best accuracy (%)	42.86 %	38.27	41.33 %	36.22 %
Classifier	Additive Logistic Regression	Additive Logistic Regression	Nearest Neighbour	J48 and Additive Logistic Regression
Kappa	**0.28**	**0.22**	**0.23**	**0.20**
F-measures:				
bored	0.17	0.00	0.11	0.07
confused	**0.52**	**0.50**	**0.52**	**0.44**
happy	**0.62**	**0.55**	**0.58**	**0.51**
interested	0.23	0.17	0.19	0.24
neutral	0.32	0.24	0.16	0.09
surprised	**0.42**	**0.44**	**0.40**	**0.50**

On the whole, the comparative results of the classification schemes indicated a limited range of performance. This suggested studying the complexity of the problem space by exploring alternative classifier designs. As such, I used *class binarisation* to re-frame the classification task by decomposing it using two well-known partition-based strategies. Class binarisation reduces the complexity of multi-class discrimination by transforming the original multi-class learning problem $\gamma = \{1,2, \dots, k\}$ into a series of binary problems and evaluates the overall performance by combining the multiple outputs (Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006). The advantages are faster training times, less computational cost as well availability of some efficient binary classification algorithms like Support Vector Machines (SVM). Not only does this facilitate the design of simple and robust classifiers but it also maximises the potential of a limited training set.

To implement binarisation, I examined two popular partitioning strategies: the one-versus-all approach, and the pairwise or round robin approach. I also used a popular partition-based clustering technique called *k-means* and the resulting silhouette curves to visualise the separability of the partitional groupings. Silhouette curves were introduced in Section 4.4 and will be used here again for cluster validation by k-means.

By definition, *k-means clustering* is an optimisation method which partitions the observations into a predetermined number of non-overlapping groups such that the within-group sum-of-squares is minimised. For a given number of clusters k, the basic algorithm begins with determining k cluster centroids which can be randomly initialised or specified by a user. Each observation is then assigned to its closest group, usually using the Euclidean distance between the observation and the cluster centroids. The centroids are then updated using the assigned observations. The assignment and updating of centroids is repeated until there are no changes in cluster membership, or equivalently, until the centroids remain the same.

4.5.1 One vs. All classification

One-versus-all (OvA) is the most common binary classification approach based on the assumption that there exists a single (simple) separator between a class and all others. Learning proceeds by learning k independent binary classifiers, one for each class, where the positive training examples are those belonging to the class while the negative examples are formed by the union of all other classes (Park & Furnkranz, 2007; Har-Peled, Roth, & Zimak, 2003). OvA classifiers operate by a winner-takes-all strategy so that a new example is assigned to the class corresponding to the maximum output value from the k binary classifiers. The OvA scheme is powerful because of its conceptual simplicity and a comparative performance relative to other binarisation methods but at lower computational costs (Rifkin & Klautau, 2004).

Applying OvA strategy therefore creates six binary classifiers, each differentiating a class from all others. Positive and negative samples of relevant emotion classes are randomly sub-sampled to learn each binary classifier. Sampling is repeated over ten iterations in order to get an even but inclusive representation. Figure 4.5-1 shows the silhouette curves obtained using

k-means clustering on the generated sample sets. The maximum silhouette values correspond to cluster size two for all pairs across the all the four feature-sets. This indicates that positive and negative samples are well distinguished in a two cluster solution making the two cluster grouping optimal.

Having a better confidence in this design, the same set of classification algorithms mentioned earlier on in Table 4.5-1 were applied, but with the addition of linear SVMs. Table 4.5-3 shows the best classification performance achieved by averaging results over ten-fold stratified cross-validation. The corresponding feature-set and classifier combinations are also listed next to the percentage accuracy. A significant jump in accuracy can be observed across all emotion classifiers with *surprised* and *bored* achieving close to 80% recognition accuracy. This is in fact reflected in the silhouette curves in Figure 4.5-1 where *surprised* and *bored* showed the highest average silhouette widths amongst all pairs.

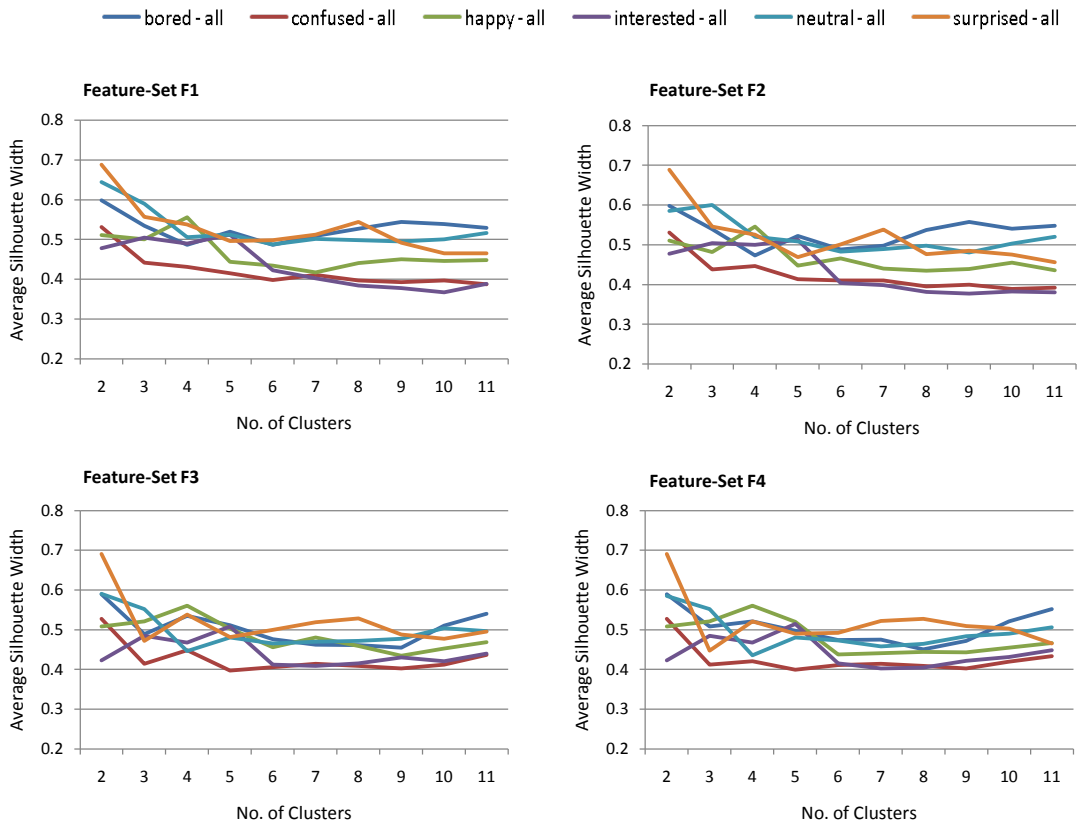


Figure 4.5-1: Silhouette plots for One vs. All binary classification using k-means

Table 4.5-3: Best classification results using One vs. All partitioning

Vs.	all other emotions		
bored	79.33 %	F1	SVM
confused	70.35 %	F2	Lazy
happy	77.1 %	F1	Logitboost
interested	61.83 %	F4	J48
neutral	64.07 %	F4	SVM
surprised	80.26 %	F3	Logitboost

4.5.2 All-vs-All classification

All-versus-all (AvA) is an alternative binary partitioning strategy which assumes the existence of a separator between any two classes. Also known as pairwise or round-robin classification, the basic idea here is to transform a k-class problem into $k(k-1)/2$ binary problems, one for each pair of classes. Each binary classifier $K_{\mathrm{i,j}}$ is trained on the subset of training examples that belong to the classes k_{i} and k_{j} only. Examples from all other classes are ignored for the training of $K_{\mathrm{i,j}}$. At classification a simple voting method can be used to combine the predictions from each of the $K_{\mathrm{i,j}}$ binary classifiers. Although AvA is a more expressive formulation and can give a boost in accuracy, its complexity is quadratic in the number of classes and is therefore more computationally expensive (Har-Peled, Roth, & Zimak, 2003; Rifkin & Klautau, 2004).

Implementing AvA for six emotion classes thus implies training 15 binary classifiers from the positive and negative samples for each pair of classes. Training data is compiled by sampling positive and negative examples for each pair repeated over ten iterations. Figure 4.5-2 displays the silhouette curves obtained by applying k-means clustering. The graphs show that the two-class grouping obtained using k-means is indeed favourable. Accordingly, classification results computed by averaging the results over ten-fold stratified cross-validation substantiate this. Table 4.5-4 lists the best performing classifier and feature-set combinations over the 15 binary classifiers.

4.5.3 Discussion

It is evident from the analysis presented in this section that classifiers perform differently on different feature-sets and show variable performance across emotion classes. Indeed no single classifier was able to demonstrate a consistent discriminative ability over the emotion classes. Moreover, the feature sets appeared to vary their predictive power based on the classification strategy. Even a simple region-based feature set of dimensionality five was able to achieve comparable performance to more expressive feature-sets in some cases. All this complicates the design of a single unified classifier for automatic emotion inference.

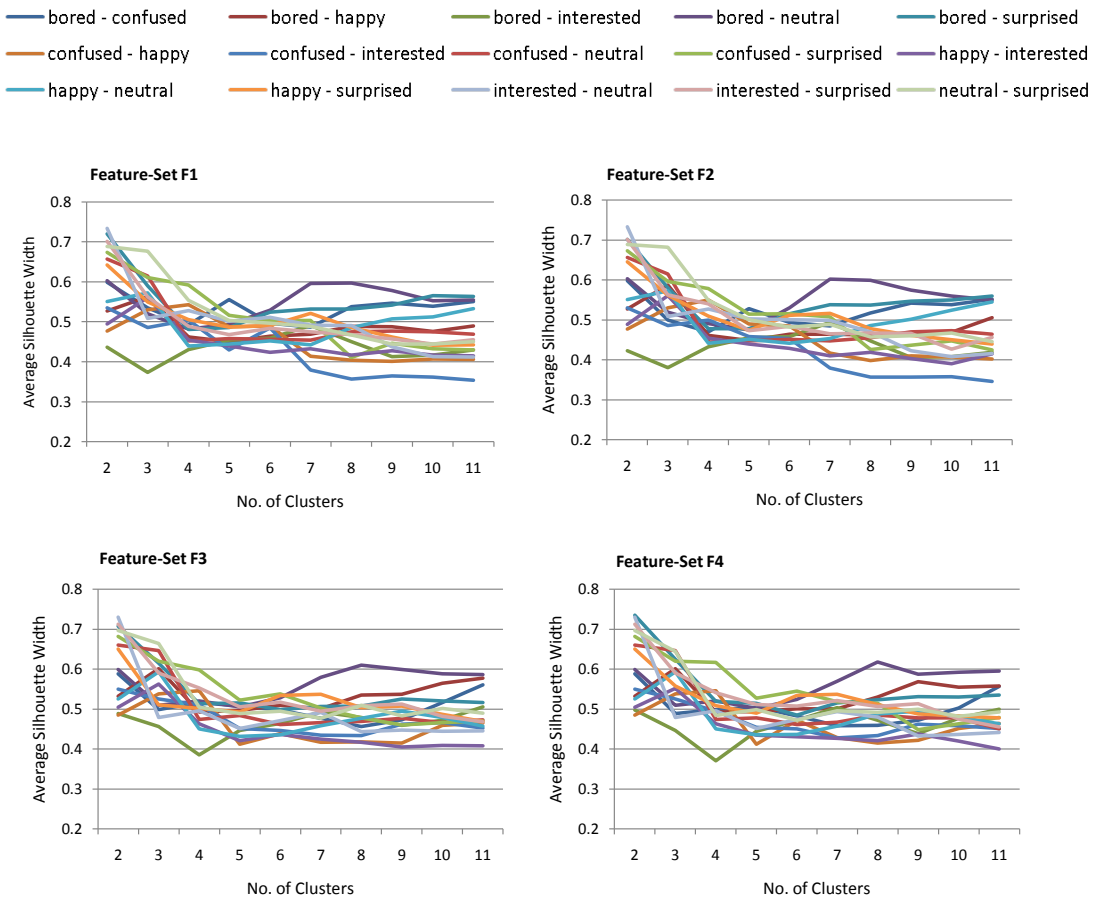


Figure 4.5-2: Silhouette plots for pairwise binary classification using k-means

Table 4.5-4: Best classification results using pairwise All vs. All partitioning

Vs.	bored	confused	happy	interested	neutral	surprised
bored		78.50 % F2 SVM	85.50 % F1 Lazy	58.67 % F4 J48	69.92 % F3 SVM	93.67 % F1 naïve-bayes / j48
confused			83.90 % F1 Logitboost	74.21 % F2 Lazy	71.03 % F2 NN	83.29 % F2 Lazy
happy				75.67 % F3 NN	79.80 % F2 Lazy	80.43 % F2 Lazy
interested					62.47 % F4 SVM	80.26 % F2 SVM
neutral						78.70 % F1 j48
surprised						

Perhaps using an ensemble of classifiers, each optimised for an individual emotion class, is a more practical alternative. It was seen for example how the removal of certain emotion classes boosted the overall recognition accuracy by a significant proportion. Accordingly, decomposing the multi-class classification problem using simple partitioning strategies enhanced the predictive power of classifiers and allowed a better discrimination. Although there was a relative improvement in results using the predictions from 15 binary classifiers in AvA, the difference was marginal when compared to just 6 comparisons using OvA. As Rifkin and Kalutau (2004) argue, OvA is preferable for its conceptual simplicity and computational power when the results are not substantially different.

In general, developing independent classifiers can facilitate modular learning and can be helpful in designing more robust and generalisable systems. It will also enable tracking of multiple or co-occurring emotions which has more viable applications. Until we know exactly how information from multiple visual cues in fused during emotion perception, exploring parallel recognition systems seems like a pragmatic option. I will examine such a parallel inference system when implementing temporal modelling as described in the next section.

4.6 Temporal Modelling

So far I have looked at general classification schemes without explicitly making use of the temporal characteristics in facial expressions. The statistical parameters used to construct the feature vectors summarised an entire time-series of sample sequences in a single representative value and are likely to have suppressed crucial temporal information. From an application perspective as well, a classifier should be able to deal with real-time data input and be able to model the temporal evolution of facial expressions. To address this, I now describe a classification system that uses a class of dynamic probabilistic network to model the temporal signatures of the six emotion classes under study.

4.6.1 Hidden Markov Models

HMMs are a popular statistical tool for modelling and recognition of sequential data and have been successfully used in applications like speech recognition, handwriting recognition, gesture recognition and even automatic facial expression classification. Based on whether the observations being modelled are discrete or continuous, HMMs can be constructed as having discrete or continuous output probability densities. Within a discrete HMM framework, *vector quantisation* is used to map the continuous space to a quantised discrete space. The process can however induce distortions in the original signal information for hidden Markov modelling thereby affecting system performance (Takahashi, Aikawa, & Sagayama, 1997; Rabiner, 1989). Although a number of methods have been proposed to accommodate for quantization error, it is intuitively more advantageous to use *continuous HMMs* (CHMM) to model continuous observations. As such I use CHMMs for modelling the temporal patterns of the six emotion classes.

Based on the progression of state sequence, HMMs can have two broad variants: the ergodic and the left-to-right model (Rabiner, 1989). In a left-to-right model, also known as the Bakis model, the probability of going back to a previous state is zero, so that the only states accessible from a state s_i are the states s_i and s_{i+1}. An ergodic model in contrast is fully connected so that every state can be reached from any other state in a finite number of time steps. Figure 4.6-1 illustrates the difference between the two. The arrows indicate permitted state transitions. To model the HMM topologies I use a left-to-right structure over an ergodic one. Apart from the obvious advantage of modelling a sequential event in a left-to-right structure, it is also considered more efficient in terms of generalisation as it involves training of fewer parameters and is thus less susceptible to over-fitting (Rabiner, 1989; Abou-Moustafa, Cheriet, & Suen, 2004).

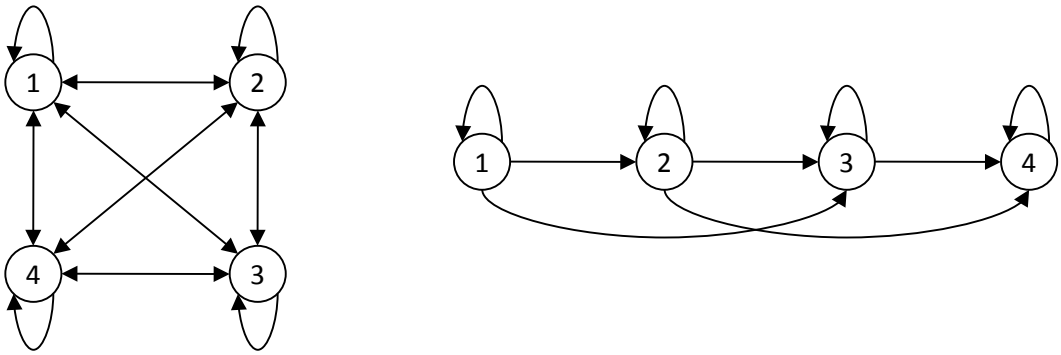


Figure 4.6-1: Types of HMM models (left) 4-state ergodic HMM; (right) 4-state left-to-right HMM;

4.6.2 Representation

An HMM models a stochastic process producing a sequence of observations vectors at discrete times according to an underlying Markov chain. In other words, an HMM defines a probability distribution over the observation sequence by invoking a sequence of hidden states. The model imposes Markov dynamics on the sequence of hidden states implying that future states depend on the present state but are independent of the past states. At each observation time, the Markov chain can be in one of N_s states $s_1, \dots, s_{Ns}$, each associated with a set of state transition probabilities governing their propensity to either stay in the current state or move to another one.

Formally, an HMM is characterised with three sets of probability density functions: the transition probabilities (A), the state probability density functions (B) , and the initial or prior probabilities (π). The compact notation $\lambda = (A, B, \pi)$ is used to indicate the complete parameter set of the HMM model in which:

- $A = [a_{ij}]$ is the state transition probability matrix where $a_{ij} = \Pr(q_t = s_j | q_{t-1} = s_i)$ for $i, j = 1, \dots, N_s$
- $B = \{b_i(o_t),\ i = 1, \dots, N_s\}$, is the set of observation probability distributions in state i where $b_i(o_t) = Pr(o_t | q_t = i)$; and
- $\pi = [\pi_j]$ are the initial state probabilities where $\pi_i = \Pr(q_1 = s_i)$.

For CHMMs, $b_i(o_t)$ represents the feature distributions using some parametric probability functions. When the training data is small, using a parametric distribution is effective in estimating suitable distributions because the constraint on the distribution shape interpolates the unseen data (Takahashi, Aikawa, & Sagayama, 1997). I use the most common parametric probability density function used in CHMM, which is the finite mixture Gaussian density so that

$$b_i(o_t) = \sum_{j=1}^{M_i} c_{ij}\ b_{ij}(o_t), i = 1, \dots, N_s$$

M_i is the number of components in state i, c_{ij} is the mixture coefficient for the j^{th} mixture component in state i and satisfies the constraints $c_{ij} \geq 0$ and $\sum_{j=1}^{M_i} c_{ij} = 1$ for $i = 1, \dots, N_s$.. $b_{ij}(o_t)$ is a d-dimensional multivariate Gaussian density, $N_d(\mu_j, \Sigma_j)$ with μ_{ij} and Σ_{ij} as the mean vector and covariance matrix respectively (Missaoui & Frigui, 2008).

4.6.3 Training and classification

The training process attempts to learn the HMMs from the sample data. The approach is maximum likelihood (ML) and the objective is to estimate the model $\lambda = (A, B, \pi)$ that maximises the likelihood of the sample training sequences, $O = [o_1, \dots, o_t]$. Typically the *Baum-Welch* algorithm is used to estimate the model parameters (A, B, π). Although there is no known analytical method to determine an optimal set of parameters, an iterative procedure like the Baum-Welch can be used to estimate the model λ such that the probability of the observation sequence $\Pr(O|\lambda)$ is locally maximised (Rabiner, 1989). The algorithm uses *Expectation Maximisation* (EM) to iterate over the observation sequences until convergence to a locally optimal set of parameters.

Given a learned model $\lambda = (A, B, \pi)$ and a sequence of observations, classification or evaluation, computes the probability that the observed sequence was produced by the model. The *forward-backward* procedure is used to estimate this probability $\Pr(O|\lambda)$ to evaluate how well the model λ matches a given observation sequence $O = [o_1, \dots, o_t]$.

4.6.4 Discriminative HMMs

I use HMMs in a discriminatory manner which implies learning one HMM per class, running all HMMs in parallel and choosing the model with the highest likelihood as the most likely classification for a sequence. This way an HMM models the temporal signature of each emotion class so that the likelihood that an unseen sequence is emitted by each of the models can be estimated and be classified as belonging to the model most likely to have produced it (Oliver & Horvitz, 2005; Cohen, Sebe, Garg, Chen, & Huang, 2003).

Thus, a bank of HMMs, λ_c where $1 \leq c \leq 6$, were trained using the Baum-Welch algorithm over the sample sequences. During training, the Gaussian mixtures with diagonal covariance are used and the initial estimates of state means and covariance matrices are found by k-means clustering. For classification, all HMMs are run in parallel and the model with the

highest likelihood, computed using the forward backward procedure, is selected as the true class. Figure 4.6-2 shows the system design.

Mathematically, given an observation sequence O_t, the model likelihoods for all the HMM models, $\Pr(O_t|\lambda_c)$, is computed, followed by the classification of the sequence to the emotion class corresponding with the highest model likelihood, i.e.,

$$c^* = \underset{1\leq c\leq 6}{\operatorname{argmax}}[\Pr(O|\lambda_c)]$$

The observation vector for the HMMs consists of the position and speed parameters as defined in Table 4.3-1. These are sampled over a sliding-window of five frames sequentially. This results in a multi-dimensional feature vector characterising a filtered pattern sequence of the temporally evolving facial and head motions. To reduce feature dimensionality and remove multi-collinearity I used PCA to extract salient features as described in Section 4.3.2.

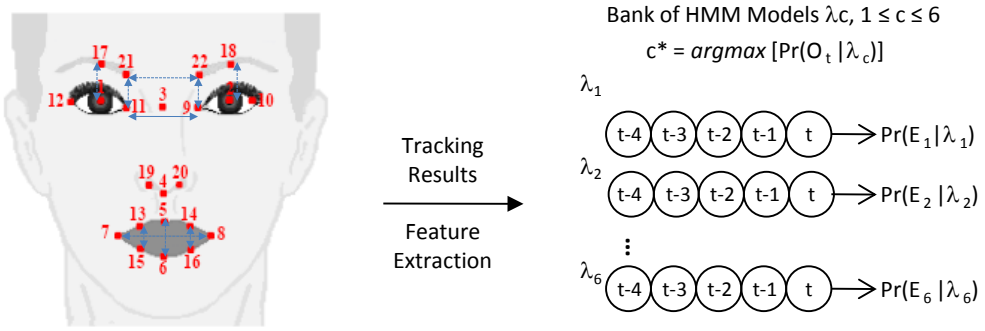


Figure 4.6-2: Classification using discriminative HMMs

An average estimate of the true positive rate by means of tenfold cross-validation is used to report the classification accuracy. To determine the best performance empirically, recognition accuracies are computed by varying the free parameters - the number of states and the number of Gaussian mixtures from two to eleven and from one to ten, respectively. Classification results computed over all the trials can be visualised in Figure 4.6-3.

In the Figure 4.6-3, the vertical axis plots the classification accuracy corresponding to the number of states which are indicated along the horizontal axis. The number of Gaussian mixtures are plotted as separate curves to emphasise the individual trends. It is evident that the performance of HMMs is highly affected by the free parameters. The graphs indicate that a simpler topology - in terms of number of states, needs complex feature distributions to achieve optimal performance while simpler feature distributions compensate with a relatively more complex topology. In other words, it appears that the temporal signatures of the example sequences are better modelled when the number of states is low but the feature distributions are complex Gaussian and vice versa.

Figure 4.6-4 further illustrates the performance of CHMMs across the individual emotion classes. The plots show how the recognition accuracy varies as a function of the number of states and the number of mixtures for each emotion classes. It appears from the figure that individual emotion classes differ in their temporal patterns and require distinct topologies and feature distributions to model them accurately. Amongst the classes, *happy* and *surprised* show best overall accuracies and reach near perfect classification at relatively modest values of topology and mixture complexities.

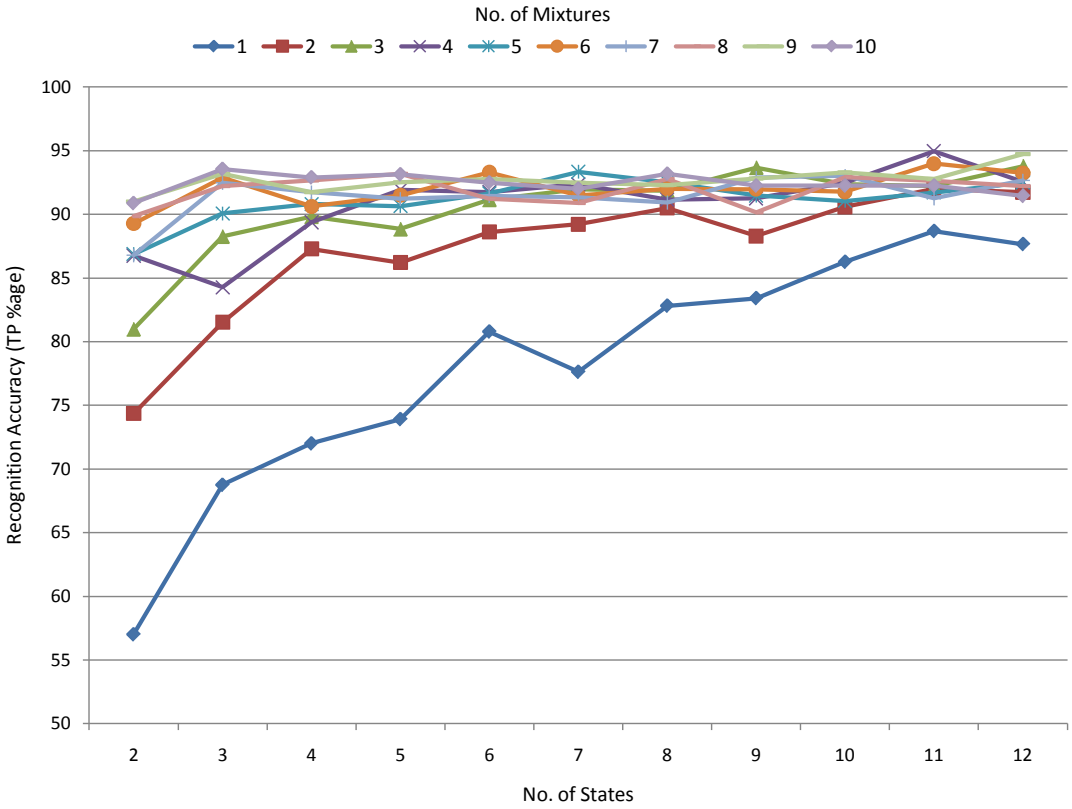


Figure 4.6-3: Performance of the discriminative HMMs over the experimental trials

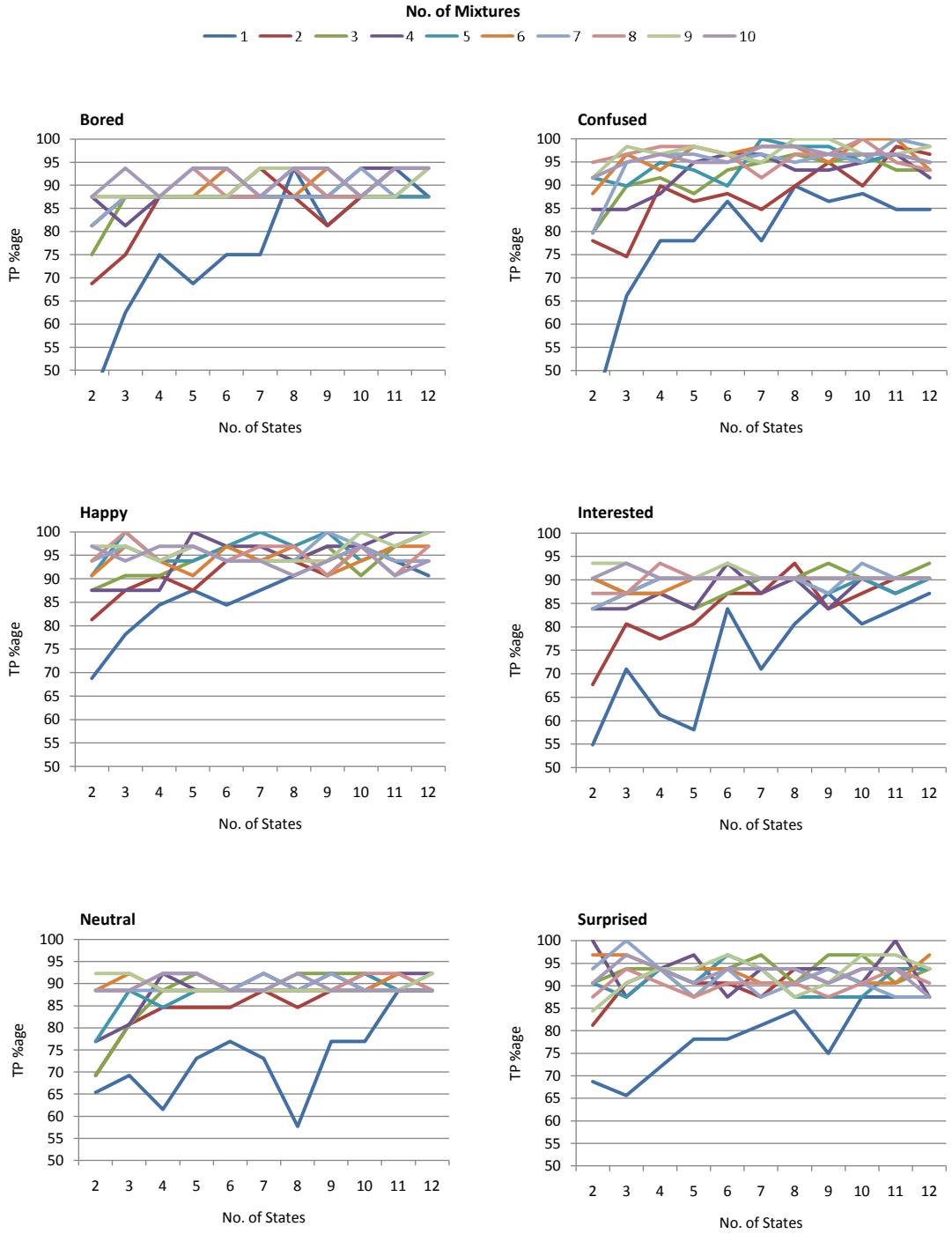


Figure 4.6-4: Performance of HMMs for each emotion class

To give an overall figure for the efficiency of modelling the temporal dynamics, Table 4.6-1 shows the detailed confusion matrix for the best classification achieved over all the conducted trails. The false positive rate is highest for confusion and negligible for others. Overall, for a mean false positive rate of just 1% the best average accuracy of 95% is obtained with eleven states and four Gaussian mixtures. *Happy* and *surprised* attain perfect true positive rates while others follow with satisfactory recognition. *Interested* shows the least classification accuracy at 87.1%.

Table 4.6-1: Best performance of discriminative HMMs

Predicted →

← Actual	bored	confused	happy	interested	neutral	surprised	total	**TP %**
bored	15	1	0	0	0	0	16	93.8
confused	0	57	1	0	0	1	59	96.6
happy	0	0	32	0	0	0	32	100.0
interested	0	4	0	27	0	0	31	87.1
neutral	0	0	1	0	24	1	26	92.3
surprised	0	0	0	0	0	32	32	100.0
total	15	62	34	27	24	34	**196**	**95.0**
FP %	0.0	3.6	1.2	0.0	0.0	1.2	**1.0**	

Although a more explicit and modular framework like the MindReader (introduced in Section 2.4.1) could have been used to implement a classification system, this would require an additional level of data preparation to first identify specific AU occurrences, delineate them in order to define precise onset and offset boundaries and then, crucially, to get these labelled by trained FACS experts. Given these constraints, I instead used a simple one step classifier as against the two-level classification in the MindReader framework in order to demonstrate the performance benefit of modelling the temporal dynamics of facial expressions using *Occam's razor* as the design heuristic. Occam's razor is a scientific and philosophical principle that postulates a preference for the simplest explanation and in this case choosing of a simpler model over a complex one (Dictionary, 2008; Alpaydin, 2004).

4.7 Summary and conclusions

This chapter detailed the various analyses conducted on the corpus described in Chapter 3. The primary objective was to develop an understanding of the complexity of automatic emotion classification from naturalistic data. Accordingly, feature definitions were constructed based on the standard grammar of facial actions, namely FACS. Feature analysis was followed by exploratory analyses in order to get an insight into the problem space. Several classification schemes were then applied to compare the performance of different statistical pattern recognition methods. Consequently, two popular partitioning strategies OvA and AvA were

implemented to evaluate performance enhancement. Finally, a dynamic classification system using a bank of discriminative HMMs was described to show the efficiency of modelling temporal information.

Overall, it was observed that different feature sets give varying classification accuracies depending on the particular classification method applied as well as the design adopted. In practice, one can apply various optimisation strategies and keep tuning parameters until an acceptable level of performance is achieved over a preferred feature set. However, each of the emotion classes has a different temporal and structural pattern which ought to be addressed in classifier design. This suggests a more modular approach for designing individual emotion classifiers based on their own distinctive feature descriptors. Aligning relevant feature combinations with emotion classifiers can also optimise the training process and reduce redundancy.

Designing an emotion classification system is in principle completely reliant on the ground-truth used to test and evaluate it. Although its generalisation can be assessed by testing performance on an entirely unseen corpus, this is unlikely to be of great merit unless the datasets are comparable at least in terms of context and recording conditions. There are several reported systems that already achieve a close to perfect classification performance but are not as yet optimised for real-world application. To accept classification accuracy as the ultimate measure then, is perhaps misleading as it essentially masks the underlying problem of the nature of emotions and the way they are conceptualised in these systems. This argument will be explored further in Chapter 6 where an alternative design approach is proposed.

In the next chapter, the sufficiency of automatically tracked facial feature points in encoding the facial patterns relevant for affect inference will be experimentally analysed.

5. Emotional information in facial feature points

Facial expression recognition is an enabling technology for affective computing and many existing facial expression analysis systems rely on automatically tracked facial feature points. Although psychologists have studied emotion perception from manually specified or marker-based point-light displays, no formal study exists on the amount of emotional information conveyed through automatically tracked feature points. To assess the utility of automatically extracted feature points in conveying emotions, this chapter presents results from an experiment that compared human raters' judgements of emotional expressions between actual video clips and three automatically generated representations of them. Specifically, the recognition accuracy for five emotions - *interest*, *confusion*, *boredom*, *happiness* and *surprise*, was analysed using samples obtained from posed and naturalistic databases of facial expressions, in different representations of varying information detail - point-light displays, stick-figure models and 3D animations. The implications of these results for optimal face representation and creation of realistic animations are also discussed.

This experiment was conceived and designed jointly with Tevfik-Metin Sezgin; additionally, Yujian Gao helped in the preparation of experimental stimuli (Afzal, Sezgin, Gao, & Robinson, 2009).

5.1 Motivation

The face, as a modality for emotion recognition, has occupied a dominant position in the study of affect in human-machine interaction. This follows from the significance of facial signs in human perception of emotion (Darwin, 1872; Ekman, 1982) as well as the relative advantages it offers over other modalities like speech and physiology. Facial information can be detected and analysed unobtrusively and automatically in real-time, without requiring any specialised equipment except a simple video camera. Even though issues such as occlusion, lighting and pose variation still remain problematic, the field has seen some increasingly good results (Zeng, Pantic, Roisman, & Huang, 2009).

Mapping of facial expressions to affective states is still however a challenging problem. Facial expressions are not simple read-outs of affective states and their interpretation is largely context-driven. To reduce this complexity for automatic affect inference, measurement and

interpretation of facial expressions has traditionally been separated. However, in order to move from expression recognition to expression interpretation, it is necessary to discriminate between facial configurations that have a psychological significance from those that have morphological value (Cohn & Schmidt, 2004). The success of this transition depends to a large degree on how much of the information relevant for affect perception is actually captured - or missed, by the techniques employed in facial affect analysis. This chapter investigates the properties of one such method, namely facial feature point tracking, to explore the information value of automatically tracked facial landmarks in conveying emotions.

5.2 Background

The typical sequence of steps in an automatic facial expression recognition system is face acquisition, followed by facial feature extraction and finally facial expression classification (Ekman, 1982; Fasel & Luettin, 2003; Zeng, Pantic, Roisman, & Huang, 2009). These were illustrated previously with the help of Figure 2.4-1. Facial feature extraction can be classified as either deformation-based or motion-based. Deformation extraction includes appearance-based techniques, while motion extraction is feature-based and includes methods such as facial feature point tracking and geometric face models. Although appearance-based feature extraction methods yield better recognition results, they require extensive pre-processing (e.g. manual alignment and scaling), tend not to generalise, and are more sensitive to variation in pose, occlusion and lighting. Facial feature tracking on the other hand, is more robust to pose variation and can deal with partial occlusion. It is therefore considered more suitable for real-time automatic emotion classification, and has been used extensively in emotion recognition systems (Zeng, Pantic, Roisman, & Huang, 2009; El Kaliouby, 2005).

The motivation for using facial feature point tracking is based on psychological studies that emphasise the role of facial motion in the perception of emotional expressions [c.f. Bassilli, 1978]. These employ an adaptation of Johansson's (1973) point-light display technique to analyse the contribution of facial movement in the discrimination of emotions. With this technique, Johansson portrayed the activity of a human solely by the relative motions of a small number of markers positioned on the head and joints of the body. To study how pure kinematics facilitates the perception of emotional states, point-light displays are constructed by recording blackened faces with numerous white spots or reflective markers while displaying emotional expressions. The white spots or markers are the only visible source of information and serve as the "carriers of motion", independent of any form or appearance information. Although an increase in the number of markers and exposure duration improves the accuracy of perception, the point-light technique has been found to be remarkably robust under impoverished or potentially ambiguous conditions. Importantly however, any disturbance along the temporal dimension seriously impairs the quality and consequently, the sensitivity in perception (Blake & Shiffrar, 2007).

The point-light technique has proven to be useful because it preserves the visual experience in a simplified representation and reduces the complexity of perceptual input (Thomas & Jordan,

2001), while retaining the temporal structure of facial expressions which is crucial for emotion interpretation (Bassilli, 1978). Feature-point tracking closely models this technique making it attractive for computational modelling and communication. As feature-point based representations are not affected by the idiosyncrasies of individuals' facial appearance, they also enable development of more generalisable computational techniques.

However, the actual utility of a feature-point-based representation in affect recognition depends on the degree to which affective information can be conveyed through the specific set of features used. Therefore it is important to understand how well a set of feature points can convey affective content, especially if the feature points are to be used for automatic facial expression recognition. The experiment described in this chapter is an attempt in this direction. In particular, it measures the amount of affective information that can be conveyed by a set of 22 facial feature points extensively used in the automated affect recognition literature.

This is done by asking human raters to identify emotions in sequences that are generated from automatically tracked feature points of videos displaying facial affect. In order to account for the effects of different representations on raters' judgements, three representation formats ranging from elementary point-light representations to intermediate stick-figure models, to complete and finer 3D ones are used. A state-of-art automatic facial feature point tracking technology is used to generate video sequences of different descriptive detail. Human raters' performance on emotion perception is then compared across the conditions. It should be noted that the sufficiency of facial motion information for the discrimination of emotions has previously been analysed only for the six basic emotions (Pollick, Hill, Calder, & Paterson, 2003; Bassilli, 1978) which makes this study a novel and interesting one.

5.3 Data preparation

The dataset for this experiment was compiled using samples taken from four different databases. These were selected to represent a range of posed and naturalistic experimental control conditions. For posed data samples, the Cohn-Kanade DFAT database (Kanade, Cohn, & Tian, 2000) and the Mind Reading DVD (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) were selected. The DFAT database consists of image sequences of facial displays acted out by different encoders under explicit instructions from an experimenter. The Mind Reading DVD on the other hand consists of emotion samples from actors given example scenarios rather than specific instructions on facial displays. As such, these are acted but unconstrained. For naturalistic data, samples collected from a simulated driving scenario (Sezgin & Robinson, 2007) and from the database described in Chapter 3 were used. The former falls into the category of induced emotion while the latter is completely naturalistic. Figure 5.3-1 shows examples of the emotion *happy* taken from each of the databases.

Three expert coders labelled the data samples from the different databases to create a final corpus of emotion samples. Five examples for each of *Interested*, *Bored*, *Confused*, *Happy* and

Surprised were taken from each database based on perfect agreement by all three coders. The DFAT database lacked examples of interest and boredom giving us a dataset of 65 samples. Mean duration of the selected video clips was 3.46 seconds (σ = 1.99). Table 5.3-1 shows the distribution of samples and their average duration for each emotion category per database.

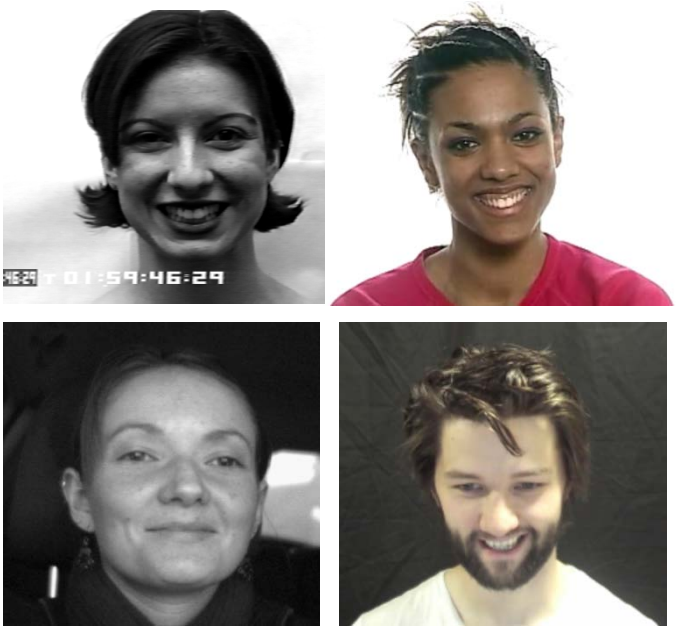


Figure 5.3-1: Examples of *Happy* taken from the four databases used; (clockwise from top left) DFAT, MindReading DVD, CAL database and the driving simulator videos

Table 5.3-1: Sample distribution of the emotion categories

Database \ Emotion	***DFAT***	***MR-DVD***	***Natural***	***No.***	***Duration (in sec)***
interested	-	5	5	10	μ = 4.1, σ = 2.1
bored	-	5	5	10	μ = 5.2, σ = 0.9
confused	5	5	5	15	μ = 3.1, σ = 2.2
happy	5	5	5	15	μ = 3.0, σ = 1.9
surprised	5	5	5	15	μ = 2.7, σ = 1.8
No.	15	25	25	**65**	
Duration (sec)	μ = 1.0, σ = 0	μ = 4.8, σ = 1.3	μ = 3.6, σ = 1.8		**μ = 3.46, σ = 1.99**

For each of the 65 video samples, three representations at varying levels of information detail were generated – point-based, stick-figures, and 3D animations. These representation formats were chosen because of their perceptual significance as well as their relevance in animation techniques. The renderings for each were generated using automatically tracked facial feature points on the original video clips. This controls for variation across displays and enables true comparison of human perceptual performance across displays (Thomas & Jordan, 2001). In all, the final corpus containing the original 65 emotion samples and their three levels of representation totalled 260 video clips at 25fps.

Figure 5.3-2 shows examples of the different representations generated from an original emotion sequence showing *surprise* from the Mind Reading DVD.

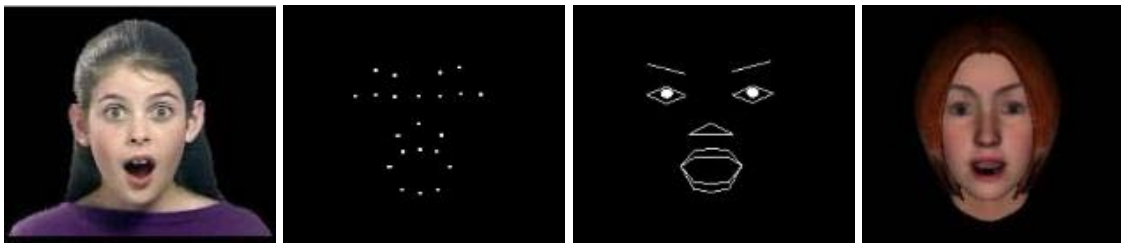


Figure 5.3-2: Example of the three representations generated from the original video of *Surprised* in the MR-DVD using automatically tracked facial feature points

5.3.1 Point-based displays

The point-based representation was created from the output of an automatic face-tracker on a black background to resemble the point-light experimental stimuli introduced before in Section 5.2. The face-tracker used to generate the point-based displays was selected after a careful review of available facial feature-point trackers, both research and commercial. The Nevenvision FaceTracker[4] requires no manual pre-processing or calibration. It is resilient to limited out-of-plane motion, can deal with a wide range of physiognomies and can also track faces with glasses or facial hair.

5.3.2 Stick-figure models

The stick-figure displays formed the next level of representation. A stick figure is an elementary drawing made of lines and dots, and was created by adding minimal detail to the landmarks, i.e., connecting the automatically tracked feature-points using straight lines and sketching eyes using typical shape. Eye height was empirically computed as half of its width. Compared to point-based displays, the stick-figure representation presents the rough outline of facial features, and is therefore more face-like and familiar to people. The stick-figure models were also rendered on black background consistent with the point-based displays.

5.3.3 3D XFace animations

XFace is an open source toolkit used for the creation of 3D animated facial expressions and displays. It implements an MPEG-4-based facial animation mechanism (Pandzic & Forchheimer, 2002) and can generate 3D facial animation by simply inputting the facial animation parameters (FAPs). FAPs are the basis of MPEG-4 Animation of synthetic face models. The automatically tracked feature points were directly converted into a set of FAPs for driving the animations. XFace animation was chosen as the third representation because of its simplicity in usage and feature support for rendering animations using FAPs.

[4] Licensed from Google Inc.

5.4 Experiment design

The aim of this study was to ascertain the information value of automatically tracked feature points in conveying emotions. Sample videos of selected emotions were used to generate three different facial representations. The objective was to analyse the perceptual differences in emotion recognition using varying levels of information detail driven by automatically tracked feature points and to know whether elementary representations like point-based and stick-figure models made emotion perception easier and more accurate, or whether a complex 3D representation allowed for finer distinction. More specifically, the purpose was to compare:

- How affect recognition accuracy differed across the three generated representations used in displaying facial information
- How affect recognition accuracy differed across databases
- How affect recognition accuracy differed across emotions
- How inter-rater agreement varied across these experimental conditions, and
- How affect recognition accuracy compared with the emotional sensitivity / affect decoding ability of participants.

The experiment was designed as a within-subjects repeated measures study where each participant labelled all the sample video clips. To minimise order and practice effects, the presentation of clips was randomised across and within each participant.

5.4.1 Participants

14 participants (8 male, 6 female) in the age-group of 20 to 34 volunteered to take part in this study. All had normal or corrected-to-normal vision and were fluent in English. They were of diverse ethnicities and were reimbursed for their participation.

5.4.2 Stimulus materials

The dataset compiled from selected original samples and their representations formed the stimulus material for the experiment. For better visual fidelity all video sequences were presented on a black background at 320 x 240 pixels. See illustration in Figure 5.3-2.

5.4.3 Labelling interface

The labelling interface was programmed as a stand-alone application using Visual Basic.NET. It allowed participants to watch randomly presented video clips and label them for emotions. Labelling was disabled while a video was playing. With the exception of a replay button, no media controls like pause, rewind or forward were made available. The replay button was however disabled while a video was playing. This was to ensure that all participants watched a video clip in its entirety without selective play and then marked an emotion label. A cross-hair was displayed between consecutive video clips to fixate attention and clear the mind from previous visualisation. A snapshot of the labelling interface is shown in Figure 5.4-1.

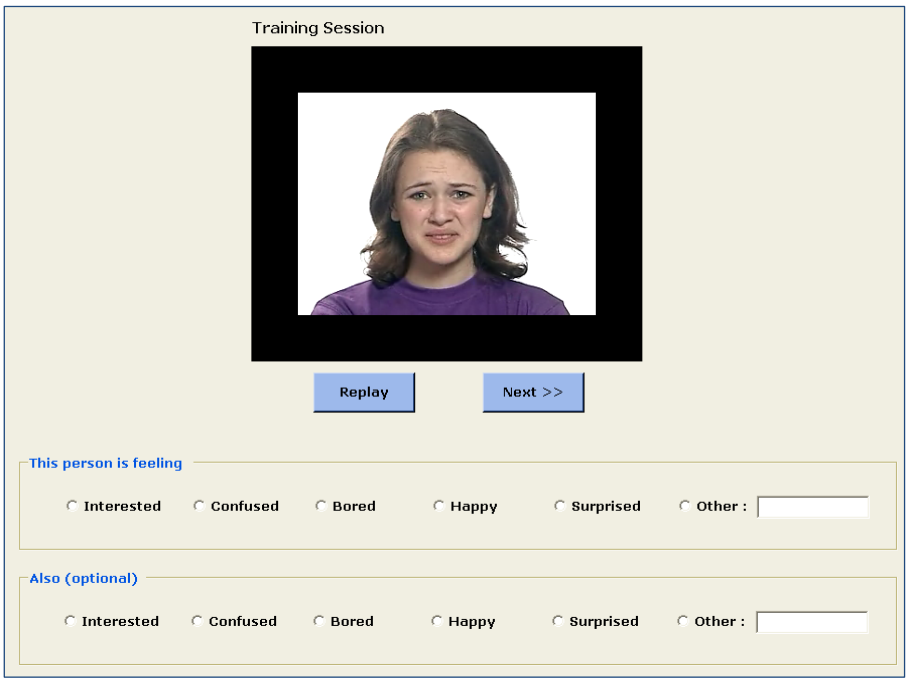


Figure 5.4-1: Snapshot of the labelling interface from a training session

5.4.4 Procedure

Participants completed the experiment individually in a usability lab. Written instructions were provided about the nature of the task. They were informed that they would be shown different facial displays which they were required to judge for emotional content. Participants also read through an emotion word list before starting the experiment. This was to acquaint them with some emotion terms used in everyday language. After signing a consent form and providing demographic information the participants underwent a brief training session.

The training session was the same for all participants and consisted of eight carefully selected videos, two from each of the different representation formats. The videos used as stimuli for training were not included in the experimental set and were sampled from both posed and naturalistic databases. The purpose of the training session was to ensure that participants understood the task and were confident in using the interface.

After the training session, participants began the labelling task, in which the stimuli were presented to them in a randomised order. They were given the option to replay a video as many times as they wanted but were instructed to follow their initial reaction as much as they could. For each video a maximum of two labels – primary and secondary, were allowed. The secondary label was optional and participants were asked to make use of this sparingly. The option of labelling more than one way was provided in order to incorporate some level of flexibility in emotion labelling. Further, in addition to the five emotions, an *Other* option was provided to allow judgements not captured through the pre-selected categories. To avoid fatigue, three short-breaks were scheduled during the labelling session for each participant.

After the labelling session, participants were prompted to fill out an Emotional Quotient (EQ) Test (Baron-Cohen & Wheelwright, 2004). This is a 40-item self-administered questionnaire used to assess emotional intelligence. The EQ Test was used as a measure of the affect decoding ability with the aim of studying if it had a bearing on the participants' judgement ability in the task. The experiment ended with participants filling out a post-experiment questionnaire and providing informal feedback to the experimenter.

5.4.5 Measures

The following measures were defined and computed:

- The primary label given to the video was considered as the true response emotional label for the presented video
- The secondary label, when present, was used as an indicator of ambiguity and co-occurrence of emotions
- The EQ test scores were used as supplementary information to interpret the effect of emotion decoding ability on this task of emotion perception
- The difficulty level in labelling a video was computed using the replay counts. Decision time was also logged and used as an indicator of difficulty

5.5 Results

As lexical emotion terms often overlap in their meaning, the responses for the *other* category were post-processed using an emotions taxonomy (Baron-Cohen, Golan, Wheelwright, & Hill, 2004). For example, emotion terms like puzzled, unsure or baffled were considered to refer to the same emotion as they all belong to the emotion group *confused*. After parsing the *other* categories where present, the primary ratings obtained from all the participants were compared with the ground-truth labels of the videos. Recognition accuracy of emotion categorisation in each of the representations and databases is compared in Table 5.5-1.

Table 5.5-1: Mean percent recognition accuracy for each emotion under each of the representations and databases used

Key: D – DFAT, M – Mind Reading DVD, N – Natural database

Accuracy (%)	Original			Point-Based			Stick-Figure			XFace			*Overall per emotion*
	D	***M***	***N***	***D***	***M***	***N***	***D***	***M***	***N***	***D***	***M***	***N***	
bored	-	92.9	88.6	-	35.7	42.9	-	34.3	50.0	-	15.7	8.6	*46.2*
confused	42.9	72.9	60.0	17.1	27.1	24.3	28.6	35.7	34.3	21.4	51.4	30.0	*37.1*
happy	100	92.9	72.9	92.9	48.6	32.9	95.1	65.7	35.7	74.3	5.7	17.1	*61.2*
interested	-	77.1	51.4	-	42.9	28.6	-	51.4	41.4	-	35.7	44.3	*46.6*
surprised	97.1	91.4	45.7	72.9	77.1	15.7	87.1	75.7	22.9	62.9	45.7	18.6	*59.4*
Overall per representation			*75.8*			*43.0*			*50.7*			*33.2*	

In terms of representation, the original videos show the highest recognition accuracy of approximately 75.82%, followed interestingly by the stick-figure models at 50.66%, point-based displays at 42.97% and the XFace animations at the least with 33.19%. Across the emotion categories, *happy* and *surprised* show higher overall recognition rates at 61.19% and 59.40% respectively. A clear trend of decreasing accuracy can be observed as we move from posed to natural data with DFAT getting an average classification accuracy of 66.07% followed by MindReading database at 53.79% and least for the natural database at 38.29%.

5.5.1 Categorisation performance

Taking representation, database and emotion as the three independent variables, a 3-way repeated-measures ANOVA was performed to test the statistical significance individually for accuracy in classification, difficulty in assigning labels and the ambiguity or occurrence of secondary emotions. Separate ANOVAs were conducted instead of a single multivariate test as the main dependent variable was the accuracy of judgement - the difficulty and ambiguity were additional data collected, but not controlled for in the experiment. Since the DFAT database did not include examples of *bored* and *interested*, these two levels of emotion were excluded from the statistical analysis. The resulting experiment design was thus 4 x 3 x 3 corresponding to representation (4 levels), database (3 levels) and emotion (3 levels). Mauchly's test was used to check whether the assumptions of sphericity were violated. In such cases, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (Field, 2009). All effects are reported as significant at $p < .05$, unless otherwise stated. The experimental analyses were conducted using SPSS v. 17.0.

Accuracy

The primary emotion labels assigned by the participants were compared with the original videos labels to compute the accuracy of classification. Significant main effects of representation, $F(3, 39) = 72.29$, $p < .001$, $\eta^2 = .85$; database, $F(2, 26) = 64.12$, $p < .001$, $\eta^2 = .83$; and emotion, $F(2, 26) = 27.29$, $p < .001$, $\eta^2 = .68$, were observed. Further, all interaction effects were found to be significant at $p < .001$, but they yielded small estimates of effect size with the exception of the interaction term database x emotion, $F(2.29, 29.74) = 48.32$, $p < .001$, $\eta^2 = .79$, Greenhouse-Geisser corrected.

Figure 5.5-1 below plots the effect sizes in decreasing order to signify the proportion of total variability attributable to the factors and their interactions. The partial eta-squared, η^2, reported is interpreted as the percent of variance in the dependent variable (accuracy) uniquely attributable to a given effect variable. So for example, a $\eta^2 = .85$, as in the case of representation, means that this factor by itself accounts for 85% of the overall (effect + error) variance in performance. It is evident that representation, database, database x emotion and emotion show very large effects while other interactions yield relatively small to medium effects.

To follow-up the significant main effects, pairwise comparisons using Bonferroni adjustment ($\alpha = .05$) were conducted. The marginal mean estimates for the significant main effects are shown

in Figure 5.5-2. A significant difference ($p < .001$) in accuracy across all representation levels except the point-light and stick-figure schemes ($p > .05$) was revealed. Accuracy was highest for the original representation (M = .75, SD = .02) followed by the stick-figure (M = .54, SD = .02) and point-light (M = .45, SD = .02) schemes. XFace videos were the least likely to be identified correctly (M = .36, SD = .02).

Pairwise comparisons for the database factor revealed significant difference in accuracy across all types (all $p < .001$). Accuracy was highest for the DFAT database (M = .66, SD = .02) followed by the MR-DVD (M = .58, SD = .02) and least for the Natural database (M = .34, SD = .03). This shows that the classification accuracy is strongly linked to the type of database and that it reduces significantly as we move from posed to natural data. Finally, the significant main effect of emotion was explained by significant differences between *confused* and *happy*, and, *confused* and *surprised* while as the mean difference in accuracy between *happy* and *surprised* was not significant. This indicates that the accuracy estimates for *happy* (M = .61, SD = .03) and *surprised* (M = .59, SD = .06) are higher than that for *confused* (M = .37, SD = .03), but not significant between themselves.

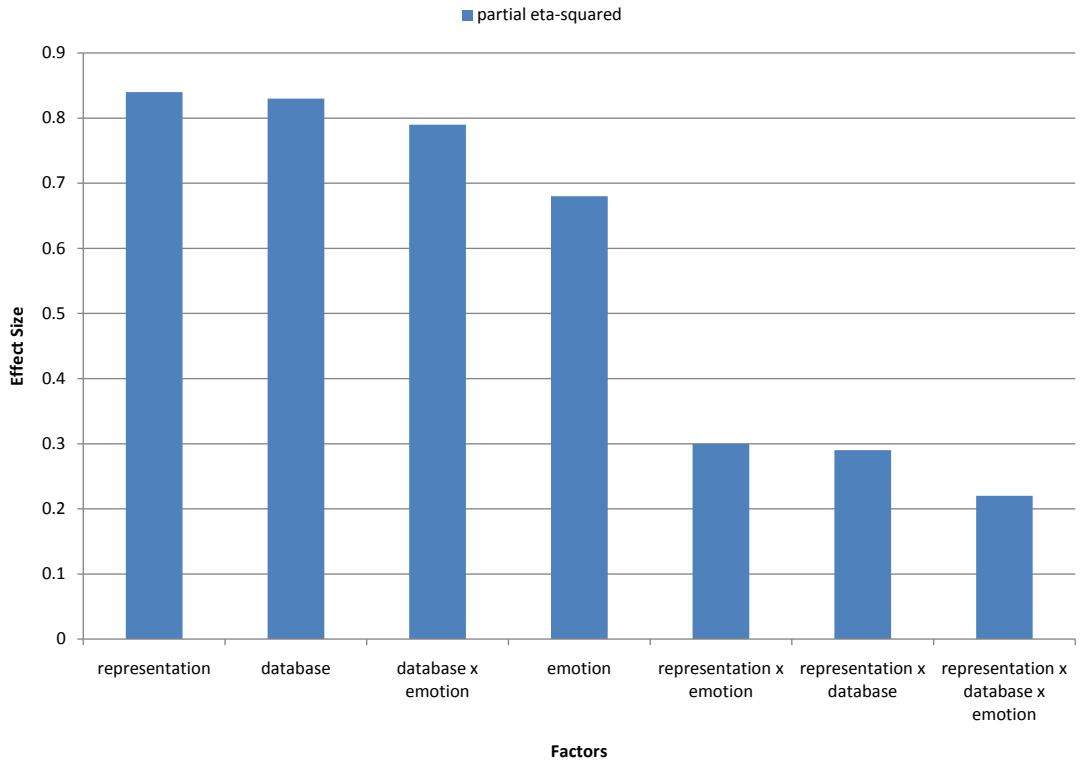
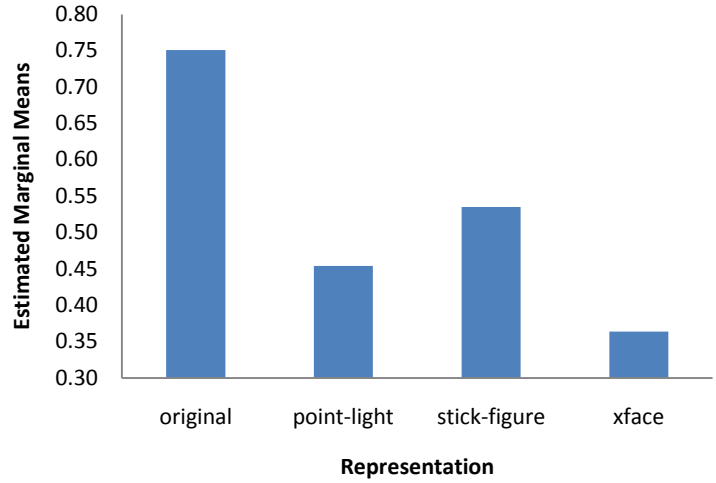
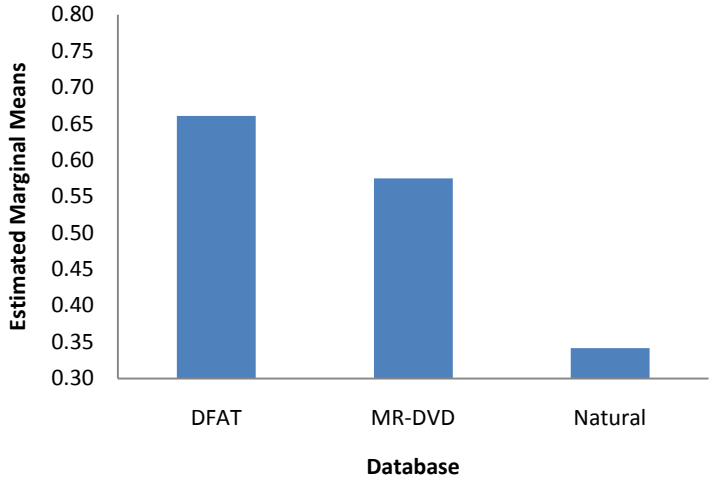


Figure 5.5-1: Effect size estimates for the main factors and their interactions of accuracy





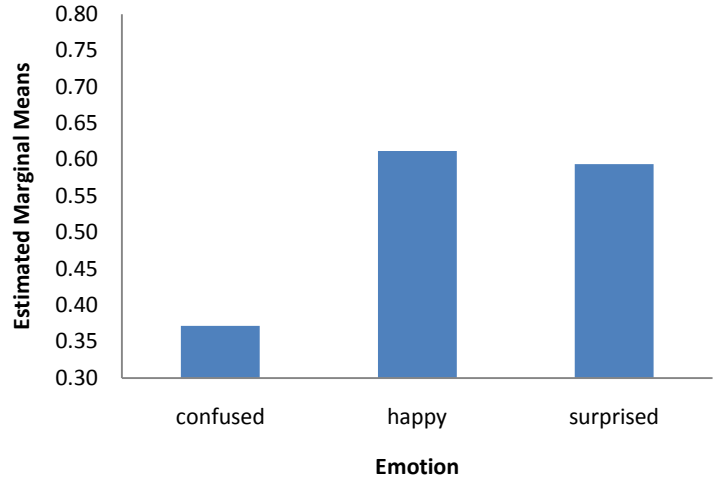


Figure 5.5-2: Estimated marginal means for the significant main effects of accuracy

Difficulty

The difficulty of identifying the emotion associated with a video was estimated by the number of times it was replayed. Significant main effects of representation, F (1.57, 20.36) = 5.00, $p < .005$, $\eta^2 = .28$ (Greenhouse-Geisser corrected), and database, F (1.16, 15.07) = 10.72, $p < .005$, $\eta^2 = .45$ (Greenhouse-Geisser corrected), were observed. No significant interaction effects were observed.

The marginal mean estimates for the main effects are shown in Figure 5.5-3. Pairwise comparisons using Bonferroni adjustment ($\alpha = .05$) were conducted to analyse the nature of these effects. In case of representation schemes, the replay counts for XFace appear as the highest with a decreasing trend as we move towards the original representation. A significant difference in the representation levels of original and point-light, and, original and XFace schemes was revealed. Specifically, the XFace (M = .51, SD = .15) and point-light (M = .42, SD = .06) schemes were replayed significantly more times than in the original representation (M = .22, SD = .08).

In case of the database factor, significant differences were found between the DFAT and DVD as well as the DVD and Natural database. This implies that the average replays for DFAT (M = .56, SD = .15) and Natural database (M = .42, SD = .09) were both significantly higher than the DVD (M = .20, SD = .05). The difference between the DFAT and Natural database replays was however not significant. While one would have expected the difficulty in terms of replays to decrease from posed to natural data, the high replay count for DFAT videos could be explained by the short duration (about 1 sec) of videos in this database which necessitated an increased number of viewings before deciding on a label. This increased number of replays did not however, affect the accuracy in identifying videos from this database as found earlier.

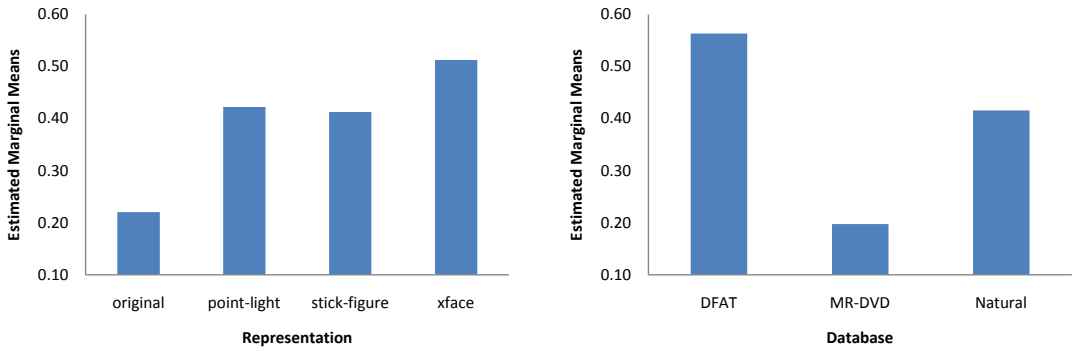


Figure 5.5-3: Estimated marginal means for the significant main effects of difficulty

To check if the difficulty in associating an emotional label was also accompanied by a similar pattern in the time taken to arrive at a decision, a repeated-measures 3-way ANOVA was conducted with decision time as the dependent variable. The decision time was calculated as the time taken to specify the emotion label for a video. The duration of the video multiplied by the number of replays, was deducted from the decision time in case a video was viewed more

than once. However, no significant effects were observed and all effect sizes were found to be quite low, $\eta^2 \leq 0.2$.

Ambiguity

Ambiguity was estimated from the number of times the secondary emotion option was marked during emotion ratings. A significant main effect of representation, F (3, 39) = 5.60, $p < .005$, $\eta^2 = .30$, and database, F (1.35, 17.53) = 13.82, $p < .001$, $\eta^2 = .52$ (Greenhouse-Geisser corrected), was observed. This was qualified by a significant interaction between representation x emotion, F (3.01, 39.08) = 3.08, $p < .01$, $\eta^2 = .19$ (Greenhouse-Geisser corrected) although with a small effect size.

Bonferroni corrected ($\alpha = .05$) pairwise comparisons revealed no significant effects in the type of representation while differences across all the database types were found to be significant. The marginal mean estimates for the significant main effects are illustrated in Figure 5.5-4. This shows that ambiguity was significantly related to the type of database and was highest for the MR-DVD (M = .15, SD = .04) followed by Natural (M = .07, SD = .02) and least for the DFAT (M = .01, SD = .01). This is not surprising considering that the DFAT database has been recorded with actors given strict instructions conforming to the universal expressions view and is therefore less susceptible to ambiguity. In contrast the MR-DVD depicts unconstrained display of emotions by actors resulting in a higher number of secondary emotion labels.

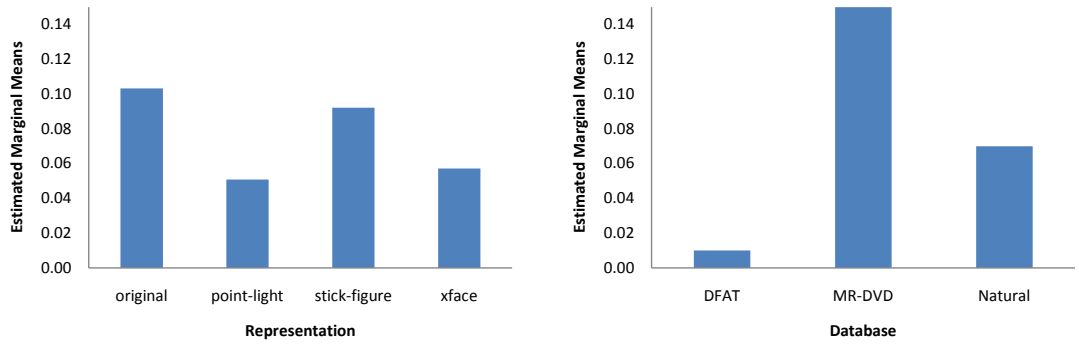


Figure 5.5-4: Estimated marginal means for the significant main effects of ambiguity

5.5.2 Emotion quotient

The EQ scores of the participants ranged between a minimum of 34 to a maximum of 67 ($\mu = 47.29$, $\sigma = 10.83$). The non-parametric correlation coefficient Kendall's Tau, τ, was computed to test the relationship between EQ scores of participants with their recognition accuracy across the representation schemes, databases and emotion categories. The EQ scores were found to be significantly related to the overall accuracy in classifying the videos, $\tau = -.41$, $p < .05$. Across representation types, there was a significant correlation between EQ scores and accuracy over stick-figure models, $\tau = -.43$, $p < .05$. Amongst the databases categories, the DFAT and the MR-DVD databases were found to be significantly correlated with $\tau = -.44$, $p < .05$ and

τ = -0.46, p < .05, respectively. No significant correlation was observed between the EQ scores and accuracy over the five emotion categories at p < .05.

The significant but negative correlation observed is peculiar but interesting as it seems to suggest that a higher emotional quotient is inversely related to the performance in emotion identification on the stimuli. One of the possible explanations could be that EQ is a measure of overall emotional acumen which in fact includes sensitivity to not only the facial cues but to the whole body and context cues and as such may not be a useful measure for a restricted modality perception task as this. In fact, if we look at the coefficient of determination for overall accuracy, R^2 - obtained by squaring the correlation coefficient and multiplying by 100; it appears that EQ actually explains only about 16% of the variability in accuracy thus leaving about 84% variance unexplained.

5.5.3 Inter-rater reliability

The inter-rater reliability was computed in order to estimate the agreement in participants' emotion ratings across the video stimuli. Fleiss's kappa (Fleiss, 1971) was calculated as the measure of agreement since the experiment involved a choice between multiple emotion categories by multiple participants as raters. The overall kappa was 0.31 indicating a fair agreement and ruling out agreement by chance. Table 5.5-2 shows the kappa values for each of the emotion categories. It is evident that *happy* and *surprised* show good agreement followed by *bored*, *interested* and then *confused*. The residual *other* category shows very slight agreement signifying somewhat random assignment. This matches with similar trends in agreement observed previously in Chapter 3 during the annotation of natural data.

Table 5.5-2: Fleiss's kappa values on the categorisation task

Category	Kappa	Agreement
bored	**0.27**	Fair
confused	0.20	Fair
interested	0.21	Moderate
happy	**0.52**	Fair
surprised	**0.45**	Moderate
other	0.03	Slight
Fleiss's (overall) kappa = 0.31 , p<.0001		

5.6 Discussion

This study was designed to investigate how the accuracy of emotion recognition is affected by the nature and representation format of stimuli generated from automatically tracked facial feature points. The type of representation and database appeared consistently as the main influencing factors for accuracy, difficulty as well as ambiguity in classification performance. Moreover, type of emotion was found to be related to the source database in determining the accuracy. As expected, original videos showed higher recognition rates consistently across representations and databases. Surprisingly however, the stick-figure models show relatively

higher levels of recognition accuracy compared to both the point-light and 3D XFace animations. The difference between stick-figure and point-light performance though was not found to be significant. Although the observed accuracies in the point-based and stick-figure models are not very high, the results support the power of temporal cues in aiding emotion perception even in the absence of explicit configurational information (Bassilli, 1978). A possible explanation is that stick-figure models and point-light displays provide the necessary cues that may then trigger emotion judgements from instinctive mental representations.

The recognition rates for emotions like *happy* and *surprised* were found to be consistently higher irrespective of the representation scheme used or the database sampled. This implies that the facial feature points commonly employed for emotion recognition using facial expression analysis may be suitable for inferring only some emotions and may not be sufficient to discriminate patterns for all emotions. States like *confused* and *bored* for instance, are accompanied by subtle changes in the face which are not adequately captured by the set of facial feature points. Except for a few selected works these emotions are in fact rarely addressed in facial expression based emotion recognition (Zeng, Pantic, Roisman, & Huang, 2009) and where done, show low classification rates. This does raise an interesting debate on whether there is a limit to emotion recognition using facial feature point tracking even if it is perfected or if certain emotions are better recognised using feature tracking, while as for others, a hybrid or alternative method (e.g. appearance-based method) may be more effective. In the MindReader, for example, image processing is used over the mouth area to supplement the tracking of lip and mouth actions (El Kaliouby, 2005).

Overall, these results provide new insights into perception of emotion from automatically generated facial displays. While the results indicate that automatic facial feature point tracking does in fact retain the underlying emotion dynamics, the efficiency of this largely depends on the type of data source as well as the emotion type (see Table 5.5-1). This becomes challenging specifically when handling naturalistic data.

The results also suggest that an intermediate-level of representation, where only an outline of facial expressions is provided, affords better perception of emotion in automatically generated displays. Any more or less detail in such artificial renderings, as presented using XFace animations, may be counter-intuitive. If stick-figure models are perceived as better encoders of emotions, then this has implications for synthesis of emotions using computer animations. It is possible that the abstraction level of a stick-figure model allows rendering flaws to be ignored and to focus attention on emotionally salient movements. In contrast, complex models like the 3D XFace animations may enhance flaws in renderings thereby diverting attention to non-significant areas or artefacts. The low recognition accuracy obtained for the 3D animation videos could in effect be attributed to the quality of the animations. Moreover, 3D models require specific attention to eye-gaze which if uncoordinated with the facial and head gestures can result in inaccurate and importantly, unnatural portrayals of emotion. Participants did indeed mention that their judgement of a facial expression was often confounded by an uncoordinated eye-gaze. Aside from the technical quality, there is some

evidence supporting the Uncanny Valley Theory (Mori, 1970) and users' discomfort with highly realistic portrayals of embodied agent behaviour (Groom, Nass, Chena, Nielsen, Scarborough, & Robles, 2009). Whether or not such a negative preference would explain the increase in difficulty and reduced accuracy in recognising XFace animations is a subject requiring further exploration.

5.6.1 Limitations

The automatic face tracker used in this study tracks only 22 facial landmarks. Acknowledging that this could indeed have affected the recognition accuracy in the judgement task, it should be emphasised that one of the objectives of this study was in fact to assess how effectively affect-related information was encoded in the facial feature points currently used in facial expression recognition technology.

The number of samples for each category of emotion was small. This was a design constraint as the number of samples to be viewed per participant was already too big (260) and adding more emotion categories would have complicated the experiment. In future, the results and observations from this study can be used to repeat the experiment perhaps with a larger sample size, but limiting the emotion categories to analyse specific affects. The ground-truth for videos, against which accuracy was computed, was based on perfect agreement between three experts which constrained the choice of emotion categories to those available in the naturalistic database. Consequently, examples of bored and interested could not be included in the statistical analyses as the DFAT database lacked samples from these.

Finally, systematic confusions in emotion ratings were not analysed during this experiment but would be an interesting direction for future experimental work.

5.7 Summary and conclusions

This chapter described details and results from an experiment examining how effectively automatically extracted facial feature points encode emotional expressions. Results comparing judgements on five emotions - interest, confusion, boredom, happiness and surprise, in three different representations – point-light displays, stick-figure models and XFace animations, generated from original emotional clips taken from posed and naturalistic databases were reported. Using state-of-art facial feature point tracking, the utility of automatically extracted feature points in conveying emotions for posed and naturalistic data was assessed. Results indicated a strong effect of the representation type and database on the accuracy of perception with a decreasing trend from original to XFace animations and from posed to naturalistic data, respectively. It was found that the emotions of *happy* and *surprised* were better discriminable than others irrespective of the representation or nature of stimuli. Inter-rater agreements for these emotions were also higher signifying a consistency in their identification. The results have interesting implications in terms of optimal representation and interplay of facial displays in emotion judgements as well as in analysing the perceptual quality and realism of computer animations.

6. Conclusions

In this Chapter I will summarise the work described in the previous chapters to draw out the principal contributions of my research and suggest directions for future work. I will also re-visit the data and in the light of qualitative observations, discuss the ramifications of using intentional affect communication as a more achievable alternative to automatic affect inference.

6.1 Summary and contributions

This dissertation takes an application-oriented stance on affective computing and addresses the problem of automatic affect inference within learning technologies. Based on previous studies that highlight the importance of affective diagnoses in learning (see Chapter 2), a relevant set of affect states was chosen using Baron-Cohen's (Baron-Cohen, Golan, Wheelwright, & Hill, 2004) lexical emotions taxonomy. A context-based corpus was compiled and a detailed annotation process undertaken. State of art facial feature point tracking was then used to encode the corresponding patterns of visual cues from facial expressions and head gestures for the pre-selected affect states.

These were analysed for their inferential ability at both machine and perceptual levels using multiple classification approaches as well as human judgements. The whole pipeline - of identifying the requirements, to collection of data, to the development of an annotation protocol, to labelling of data, and the final analyses, was completed in this dissertation. In effect, a framework for conducting research using natural data was set out and the challenges encountered at each stage identified. This makes this work significant as one of the first thorough analysis of visual behaviour based entirely on naturalistic data in a target scenario. More specifically, the main contributions arising from this work are:

- A set of affect descriptors relevant in learning scenarios.

  This was initially chosen based on a review of former studies but was refined during the successive annotation stages. Baron-Cohen's hierarchical lexical taxonomy was used to organise and conflate the numerous but semantically similar affect terms. This set of affect states was validated during the annotation process except for the addition of *surprise* which did not feature in the originally selected list of relevant emotions. The annotation method subsequently used during the labelling of collected video data was not purely fixed-response and allowed the use of a free form *other*

category to give coders some flexibility in articulating their emotion judgements. The contents of such *other* annotations were parsed using Baron-Cohen's taxonomy and the GALC (Scherer, 2005). With the exception of *surprised*, this resulted in a negligible number of videos that could not be assigned to any of the pre-selected emotion categories. *Surprised* was thus added to the list of domain relevant categories because of its frequent occurrence in the data as noted by the coders. The set of affect states thus represents the range of emotions observed in the collected video data. Furthermore, the proportion of labelled instances showed the predominance of *confusion* followed by *surprised*, *interested*, *happy*, *bored* and *annoyed*.

- Compilation of a naturalistic context-based corpus.

  As the focus of this research was to undertake an application oriented feasibility study, it was necessary to make use of naturalistic and spontaneous displays of affect rather than posed or acted ones. This involved a data collection exercise followed by an exhaustive and detailed annotation process documented in Chapter 3. While this was the most time-consuming part of the research, it was nevertheless the most vital as it gave a first hand experience of the basic problems in conceptualising affect and the complexity in its measurement. While the eventual outcome was a database of naturally occurring emotional behaviour, there were several important contributions that arose as a direct consequence of this:

  - A systematic and thorough investigation of the issues involved in both collection and labelling of naturally occurring behaviour. This included a formal introduction of methods and instruments from nonverbal behaviour research showing their utility in understanding and handling issues faced in affective computing studies.

  - Assessment of dispositional emotional expressivity through standard psychological instruments like ACT, EES and BEQ; and their comparison to corresponding measures for HCI. Except for a subscale of BEQ, namely BEQ-STR denoting strength of emotional expressions, none of the traditional psychometrics were found to correlate with the observed emotional behaviour in HCI pointing to apparent differences in the emotional behaviour that occurs in HCI as against social interactions. In general, the relevance of individual differences in emotional expressivity can be useful not only in understanding the problem space but also in conceiving more personalised models as in speech recognition, in which a canonical model adapts to individual verbal styles.

  - Development of an annotation protocol which included subjective as well as objective emotional experience accounts. A video preparation framework was also defined to finally extract the annotations into corresponding sample video clips for training.

- Results related to observed inter-rater agreements and their variance across the emotion categories. Associated with the annotation results were the gender differences in labelling wherein the female raters were found to be better than male raters in making emotion judgements. Following from this, suggestions were made to pre-screen annotators based on nonverbal decoding ability estimated using standard psychometrics like PONS, EQ, etc.

In all, the preparation of data and the annotation procedures were introduced and described in sufficient detail to serve as an important knowledge base for researchers planning to work with naturalistic data. Further, by exploring issues like dispositional expressivity and gender differences in emotion judgement, I have broadened the understanding of the problem and hopefully made the community more aware of basic issues involved in formalising a phenomenon as complex as affect.

- Comparison of multiple classification approaches and implementation of an automatic emotion inference system.

Considering that there are very few works on naturalistic data, the machine level analysis presented in Chapter 4 gives an important insight into the complexity associated with naturally occurring emotional displays. Using AU descriptors from FACS, two different feature groupings – anthropomorphically meaningful regions and de-correlated feature space, were selected with/without inclusion of rigid head motion in order to derive four different feature sets. Being of different dimensionalities, their relative discriminative ability was assessed. Data exploration methods like unsupervised classification and MDA were then used to uncover the structure of the underlying sample data and to visualise the class discrimination complexity.

The feature sets were used to evaluate the performance ability of several standard classifiers which was found to be quite low. As a result, two class binarisation strategies, OvA and AvA, were implemented to boost the classification accuracy. Reducing the multi-class classification problem into a set of binary problems resulted in a significant jump in recognition accuracy highlighting the value of exploring alternative classifier designs to reduce classification complexity.

This was finally reflected in the design of a parallel emotion inference system wherein discriminative HMMs were used to model the temporal signatures of the individual emotion classes. For classification, all trained HMMs corresponding to the individual emotion classes were run in parallel and the model with the highest likelihood was selected as the predicted class. The performance of this simple system was evaluated over several experimental trials and shown to give a best performance of approximately 95%. The underlying differences in the temporal signatures of the individual affect states were also highlighted.

- Measurement of the amount of emotional information encoded by automatically tracked facial feature points.

  Automatic feature point tracking is one of best suited methods for real-time emotion inference and has consequently enjoyed considerable attention in automatic facial affect analysis. However, no formal study exists on the amount of emotional information actually captured or conveyed through the automatically tracked feature points. This becomes more significant when dealing with naturalistic data which is relatively unconstrained and therefore more challenging to track as compared to the posed or acted expressions of emotion.

  Chapter 5 presented the first experimental evidence on the performance and ability of automatic facial feature point tracking across databases as well as emotions. By comparing human raters' judgements on emotional expressions generated from the tracked feature points, a clear evidence of decreasing performance in human emotion judgement was observed with the shift from posed to naturalistic experimental stimuli. Moreover, it was found that this ability differed based on the emotion type with *happy* and *surprised* showing overall better recognition rates irrespective of the database sampled. This indicated that facial feature point tracking might be suitable for encoding the discriminative patterns of only some emotions and may be inadequate for others.

6.2 Reflections

Advances in affective computing have indeed opened the possibility of modelling the expertise and social dynamics of expert human mentoring. Using computer vision techniques and statistical inference it is now possible to conceive of automatic affect recognition from nonverbal behavioural cues. But automatic prediction using machine learning relies on an extensive training corpus which requires preparation of labelled representative data. This serves as a baseline for training and testing different techniques and is therefore crucial for development and evaluation of computational models of emotion. The data drawn upon in this dissertation was aggregated to train such an automated facial affect recogniser. The objective was to collect naturalistic data in the target scenario - an increasingly emphasised stance in the field in order to ensure that systems generalise to real-world scenarios (Batliner, et al., 2003; Cowie, Douglas-Cowie, & Cox, 2005). The process of designing and carrying out the data collection exercise and subsequent analysis however, produced more questions than answers, with each stage raising questions about what was actually meant by emotion.

Conceptual and methodological issues for example kept recurring in different forms. Identifying the appropriate affect descriptors was difficult and even though a domain-relevant set of categories were identified, their inclusiveness and perceived meaning remained questionable. The intricate relationship between the recording context and the

resultant behaviour posed questions with regard to ecological validity. Yet the conflict between the recording setup and the eventual video quality suitable for video processing made for some compromises like shifting to a usability lab. The choice of the learning environment also required careful consideration. To minimise any potential confounds in the assessment and interpretation of emotional expressions, a one to one self-regulated learning scenario was chosen. Further, two different learning tasks were selected to get variety in emotional behaviour. Then, when selecting an index of emotional expressivity, the variety of possible measures for the same concept made it uncertain whether these measures were context-specific and whether they would in fact apply to human-machine interaction. Indeed, with the exception of the subscale BEQ-STR, the expressivity measures did not correlate, positively or negatively.

After collecting the video data, its segmentation and annotation highlighted the complexity and variance in emotion judgment across a range of human raters. During self-annotation, participants' reflection and meaning-making seemed to conflict with the required emotional account. Their surprise, amusement, as well as boredom, in watching their videos affected their judgement ability. While this subjective interpretation is a known shortcoming of self-report (Larsen & Fredrickson, 1999), it does make it difficult for use as a ground-truth for training a classifier. Subsequently, identifying the emotional episodes from the continuous video records was difficult and demarcating more specific boundaries into expressions indicative of different affect states was extremely challenging. Not surprisingly then, the inter-rater reliabilities were low indicating the lack of consensus in emotion judgement between the human judges themselves. So even though a number of classification methods were explored with good success, a more fundamental problem was getting concealed amidst the focus on objective measures of recognition accuracy and error rates.

Such inconsistencies and practical issues encountered during the research prompted me to reflect on the practicality of incorporating automatic affect inference, as currently understood, in the target application. With the help of an ethnographer, Cecily Morrison, I took an exploratory approach to re-examine the data. The objective was to develop a more qualitative understanding of emotion expression in this setting and reflect upon the role of emotions in human-machine interaction. The outcome, although unexpected, was a re-evaluation of the importance of context in human expressivity and the re-thinking of assumptions inherent in the methods and goals of affective computing.

6.3 Revisiting the data

6.3.1 Automatic Affect Inference

The aim of affect-sensitive technology is to interpret a user's affect state from nonverbal behavioural cues. Applications such as a computer tutorial can adapt intelligently based on its understanding of user behaviour and without need for the user to express explicit intent. The goal is to get an insight into the emotional state of a user from observable signs like facial

expressions or gesture. In the following two sub-sections, I examine the practicality of this desire by taking some examples from the data.

Task difference

Consistently across the participant group, regardless of the level of overall expressivity, more emotional reactions were observed during the card game than during the tutorial. In the card matching game, sharp but frequent expression changes were observed. During the tutorial, the faces became slack and an emotional expression occurred only every few minutes. The expression changes were infrequent, sustained, and slow. This is perhaps not surprising as the card game changed rapidly, giving many 'events' to which to respond to while the tutorial required periods of concentration and understanding during reading, and thinking during recall.

The card matching task set the pace and thus evoked reactive behaviour. It might be likened to a conversation between two people in which there is a constant stream of both non-verbal and verbal 'events,' to which to react. In contrast, the tutorial put the learner in control, involving application of individual learning style and deeper cognitive engagement. The human-machine interaction that takes place during a tutorial then is likely to include less intense expressions of emotion and rather contain periods of low expressivity during concentration. It isn't that the learning process was devoid of an emotional experience but rather that the expression of it was more muted. Machine recognition of emotion would prove to be technically difficult in situations like this.

One of the possible ways to induce more expressivity would be by making the learning interaction more eventful (and two-way) as in done in many learning systems with the addition of a pedagogical agent. However, there is no clear evidence regarding the merit of using embodied affective agents in learning environments. In a comprehensive review of studies using embodied agents, Beale and Creed (2009) demonstrate the inconsistent findings regarding the effectiveness of such agents with contradictory results of enhanced engagement as well as distraction and negative influence in learning and retention. Zakharov, Mitrovic, and Johnston (2007) for example show that such interaction aides often prove distracting and are perceived as unnatural. This suggests that the connection between concentration and low expressivity in the data is not a matter of chance; concentration and engagement during learning likely lowers the threshold of emotions and affects both the quantity and quality (dynamics) of expressive behaviour.

In general, it seems that the type of activity with a machine has a substantial impact on the nature and expressivity of an individual with deeper levels of concentration and engagement associated with subdued manifestations of affect. It is also known that the perceived merit of an activity or task influences the degree of control on behaviour as the appraisal of a situation and the specific strategies used can influence and change the emotions experienced (Schutz, Hong, Cross, & Obson, 2006). This in turn implies that the applicability of affect-

sensitive interfaces may be constrained to certain types of tasks and perhaps to more emotionally evocative interactions.

Situational reactions

The reactions of the participants to the three triggers/events presented during the card activity (Section 3.2.2) were interesting. For example, each participant indicated that they were confused when the screen went blank for five seconds. Nonetheless, their reactions to experienced confusion were quite different. Participant 7, for instance, laughed, while Participant 3 appeared to be extremely concerned, almost alarmed. Although both indicated that they felt the same emotion – confused, their reaction to the situation engendered very different facial expressions. That is, the manifestation of the same subjective experience was quite different. Participants 4 and 6 on the other hand, did not find anything untoward in the screen blanking out while their reactions could be interpreted as mild confusion indicated by eye-brows drawn together and eye-gaze scanning the screen as if waiting for the task to resume. This suggests that there is a distinction between felt emotions and situational reactions and it is rather the latter that is observable. In other words, the nonverbal behaviour that one would interpret as signifying a specific emotional state could actually be misleading.

In a related example, Participant 3 expressed anxiety throughout the tutorial, while Participant 7 never expressed anxiety. The former, as discovered from the interview transcripts, was overly concerned about performing well in the learning task and feared getting something wrong, while the latter did not worry even when things seemed to appear wrong. The two had very different attitudes towards the learning tasks which influenced their behaviour. Similarly, the change in feedback got diverse reactions with some subjects not noticing it all to others confused or distracted and a few pleased with it. When presented with a card with no match and therefore no answer possible, reactions ranged from being confused and surprised, to considering it to be a part of the game itself. Specifically, Participant 4 went for the closest possible match construing it as a trick cue card while Participants 3 and 6 were confused and selected a random answer, interestingly with a display indicative of amusement.

Although this highlights the inaccuracy of mapping emotions directly from observable signs, it also draws out the problem of how an application should interpret user behaviour in such instances when it is the users' attitude that is manifest and not necessarily the underlying emotion. From an application perspective, the dominance of individual learning style in managing attitude rules out the simplistic notion of adapting content and pace of learning based on the learner's affective state. Moreover, determining the type of adaptation, whether empathetic or reactionary, would involve a further complication that could influence subsequent emotional behaviour. The findings of Brave, Nass and Hutchinson (2005), on the varied psychological effects that the orientation of emotion exhibited by an embodied agent has upon users, are a relevant example.

There are latent factors then that govern responses to interaction and these can have a direct effect on apparent user behaviour. Deciding how and when an emotion recognition system should give credence to behavioural cues is indeed going to be difficult. Even if a behavioural reaction can be established reliably, its interpretation in terms of users' internal mental processes, attitudes and experiences presents serious challenges.

6.3.2 Intentional communication of affect

The technical challenge in acquiring relevant information from naturalistic displays together with the complexity in interpreting it amidst factors of personality, attitude and situations shows that spontaneous emotion recognition from user behaviour may not be so practical. On the one hand, concentration and interest seem to reduce expressivity causing technical difficulties; and on the other hand, the distinction between felt emotions and situational reactions creates design issues. The higher level affect state interpretation from observable signs in effect requires that we equip computers with the knowledge, experience, observations and learning that we as humans acquire over our lifetime.

Perhaps a more pragmatic solution could be to actively involve the user and shift the onus of interpretation from the computer to the user. As Ward and Marsden (2004) highlight, it is the intentional affect that is easier to recognise and in fact more important than reactive affect in human-human interaction. Acknowledging that although human perception of emotions is not always accurate, they give examples from strategies that counsellors and therapists teach us to share and seek verification of our judgements through dialogue. Affect, they emphasise, has an intentional communicative function and is used to negotiate meaning in our interactions. This echoes Fridlund's (1994) functionalist view of emotions as being strategic acts that serve to control social interactions. In other words, facial expressions of emotion can be seen as social messages dependent on motive and context and therefore profoundly influenced by the nature and trajectory of ongoing social interactions. Consequently, in terms of incorporating nonverbal information in technology, Fridlund suggests "that it may be better to consider facial expressions to be declarations whose referents are external than as eruptions whose referents are internal" (Fridlund, p. 3).

An example of active user participation in communicating emotion is the Subtle Stone Project (Alsmeyer, Luckin, & Good, 2008) wherein students in a classroom use handheld, squeezable instruments with seven colours signifying seven emotions to communicate their subjective emotional experiences privately to the teacher. They find that encouraging engagement in the process of self-reporting emotional experience was positive for both students as well as teachers. However, teachers did admit to being overwhelmed at times with the volume of information they received.

For a more meaningful adaptation and interaction then, an alternative approach of intentional affective interaction could be suggested wherein users - understanding the consequences of their non-verbal behaviour, make an active effort to be understood. On one hand, this is likely to encourage more expressivity in users making emotion detection easier

and perhaps more robust. On the other hand, it eases the design problem of not knowing if and how a facial expression should be interpreted as the user will purposefully and pro-actively engage in the communication. This would address the concerns of Wosnitza and Volet (2005) who emphasise the importance of knowing the orientation or directedness of emotions in determining effective intervention strategies during learning.

Such a system would also conform to the accepted ethical stance that users remain aware and in control of what and how information is being used by the computer (Picard & Klein, 2002). Transparency about the machine's role and functionality is in fact one of the fundamental principles of Human Centered Design (Norman, 1988). The study by Axelrod and Hone (2005) for example shows how users adapt their behavioural response depending on their belief and expectation of a system's affective response. They simulate an affect sensitive application using a Wizard of Oz scenario wherein a hidden human observer provides affective interventions in a simple word ladder game. They observe more positive and intense expressions when users are told that they are interacting with an affect-sensitive program as compared to when they are not. Although the interactions lasted only about 10 minutes, their study shows that user expectation and intentionality does have a significant bearing on observable emotional behaviour. Apart from the obvious concerns of associated fatigue and distraction from the task at hand my data suggests that designing an intentional affect responsive interface is not simply a design problem but may in fact involve more complex issues.

Seven out of the eight participants indicated on their questionnaires that they would interact differently if they knew the computer could respond to their affective state. From this one can hypothesize that this 'difference' would be a magnification or conscious regulation of behaviour as happens when one tries consciously to communicate an emotion, such as pleasure. However, two incidents in the data indicate that this may not occur since interacting with a computer is devoid of the usual social consequences that stimulate non-verbal behaviour in everyday life.

One participant, Participant 2, was quite flamboyant when speaking with the researchers, liberal in her use of body gestures and facial expressions. It was clear that she intended these non-verbal behaviours to draw attention and reinforce her opinion and personality. It is unlikely that she would use a similar strategy with a computer. Or even if she did, interpreting her expressions would be problematic as their meaning would be highly ambiguous. Indeed, she was one of the least expressive individuals in the study.

Data from another participant, Participant 1, suggests a further complication. Participant 1 had some of the lowest expressivity test scores, but was very expressive while using the computer, so much so that she even surprised herself. Not only did she evidence quick and dramatic changes of emotion, but engaged in other expressive behaviours such as gasping and 'giving the finger' to the machine. In discussion with this subject, she revealed that she worked in a male-dominated technical environment where emotional displays encouraged a gender stereotyped image that she wanted to avoid. Being with a computer gave her an

outlet for expressing herself. Her apparent comfort with a computer was partly due to its being a machine and thus non-judgemental. If the computer could understand her, she would be less likely to use the same expressions or 'abusive' gestures as an emotional outlet.

It appears that the social context plays a significant role in the use of non-verbal expressions during interaction. So far as the behaviour is concerned, the data corroborates the *Computers as Social Actors* (Nass, Steuer, & Tauber, 1994) paradigm which asserts that people mindlessly apply social rules in their interaction with computers. But as Nass and Moon (2000) emphasise, what characteristics of the interaction brings out this behaviour in users and whether or not it is similar to that obtained in a human-human setting is yet to be determined. So while we know that people ascribe persona and social behaviour to devices, we do not know the exact nature of this behaviour. While reviewing evidence on the facilitatory and inhibitory influence of personal affiliation (real or imagined) on emotional expressiveness, Parkinson (2005) highlights the importance of studying the contribution of social motives in expression studies and endorses Fridlund's (1991) views on the communicative nature of emotion as being more accommodating of previous experimental findings.

My analysis indicates that awareness of the machine's passivity and lack of social and interpersonal context will affect the user's expressivity and behaviour. Research into intentional affective interaction then cannot rely on data obtained in a setting without an emotionally sensitive interface as currently done. New methods of data collection are needed to explore this idea fruitfully.

6.3.3 Discussion

Affective computing envisages truly effective human-machine interactions as being affect-sensitive. Yet the field is both motivated, and influenced by an understanding of emotion in an environment, that of person to person, that differs from its eventual application, person to machine. It builds on the premise that adapting applications based on the emotional state of users leads to compelling and effective interaction with machines. This has often been interpreted to produce scenarios of use like the following: if a computer tutorial senses frustration, than it can adapt the content that the user receives to mollify that negative emotion, much like a human teacher would do. Such scenarios however, have an implicit assumption that people 'interact' with machines in the same way that they do with humans – that is, they suppose that users follow the same protocols of emotional behaviour. They expect that: (1) nonverbal behaviour associated with emotional state will be similar to that observed in human-human interaction; and (2) users will accept the same type of adaptive intelligence from a machine as from a person.

Although not as an explicit theoretical stance, the assumption that humans interact with machines as they do with humans is inherent in the methods and practices of affective computing. This can be observed in the way affect is conceptualised and subsequently modelled, as well as in how representative data is prepared for the development and training

of potential affective computing technologies. For example, most of the computational techniques for recognising emotions are developed using databases that are oriented to prototypical representations of a few basic emotional expressions used with humans rather than collected from interaction with machines. Another example is the assumption that people appreciate having their environments changed by a machine, as they would from a well-meaning person. As Ward and Marsden (2004) caution, this is falling into the same mistake as earlier intelligent tutoring systems that put the computer in control and assumed a reactive user.

I have presented an analysis that leads to question both the accuracy and lack of nuance in the assumption that people 'interact' with machines in the same way as they do with other humans. Analysing the data obtained in a potential application environment - computer-assisted learning, I highlight the limitations of such an understanding. Apart from the methodological issues associated with the perception and measurement of affect, there are other issues that need to be considered for viable application of affect-sensitive technology. The qualitative analysis indicates that people express themselves less during a cognitively engaging task like a tutorial than during the faster-paced activity of a card-matching game.

It seems that there is a likely conflict between emotion expression and concentration, indicating that emotion recognition for environments demanding concentration may prove to be difficult and of limited application. Furthermore, the problem in distinguishing between a felt emotion and a situational response makes it difficult to utilize the recognized expression for the purposes of an adaptive application. Intentional affective interaction between humans and machines is proposed in which the user knows that the machine is reacting to its expressions and actively utilizes them. The data suggests that in order to pursue this design idea further, we must gain a better understanding of how intentionality influences interaction and how expression is related to social context - something the computer will never have.

Finally, my analyses and experience while working with natural data have only questioned the robustness and the feasibility of the standard approach. Nonverbal behaviour outside controlled experimental conditions is subtle and varies significantly across individuals and contexts. My contention is that we need to explore the viability of our assumptions and the resulting implications, specifically within the context of real-world applications like learning environments. The notion of developing automatic emotion recognition systems followed by appropriate intervention strategies, as if disparate stages, is based on a reductionist conceptualisation of affect as a measurable and discrete entity independent of the interaction it actually emerges from and continually influences.

The computational modelling of 'context' illustrates a similar representational problem invoking Dourish's (2004) proposition of reconsidering it as an interactional and situated problem instead. Asserting this stance, Boehner et al argue against the general practice of using the information processing model of emotion in affective computing whereby emotion is construed as an internal, individual and private phenomenon that can be delineated and formalised using well-defined constructs (Sengers, Boehner, Mateas, & Gay, 2008; Boehner,

DePaula, Dourish, & Sengers, 2007). The interaction model of emotion thus provides a theoretical framework for exploring the emergent nature of affect and to probe creative ways of modelling affect while explicitly accounting for intentionality.

Different perspectives

At this stage the discussion can be extended to reflect on the differences in methodologies adopted in affective computing and how these influence the way emotion-sensitive systems are conceived to function. To analyse this, I consider two seemingly conflicting approaches namely, the design vs. the engineering approach. These embody a difference in perspective in that a distinction can be drawn in terms of the approach they use as well as the evaluation strategy they employ.

The engineering approach seeks to formalise affect in terms of precise computational models or rules. It presupposes a well-defined problem and seeks to find an optimal solution focusing on the right combination of features to give the right level of recognition accuracy. With the underlying focus on representation, be it categorical, dimensional or appraisal-based, the definition of emotion is assumed to be universal, standardised and therefore portable across contexts. The complex task of emotion perception is thus reduced to determining a mapping between patterns observed in one or more of the nonverbal channels to the affect construct(s) set out to be relevant for a system. Affect intelligence is then built in sets of algorithms using pattern recognition and machine learning techniques on datasets, mostly posed/acted, that conform to the required representational stance.

Embedded within this notion is the underlying concept of emotion as an absolute and unchanging entity that has its own meaning separate across people and interaction contexts. Boehner et al (2005) call this the information-processing model of emotion in which affect is considered to be a kind of information that can be transmitted in a loss-free manner between computational systems and users. Emotion is abstracted in terms of units of information and plugged into an underlying system architecture as yet another component or module. This separation of emotion from its overall context results in an impoverished representation and is unlikely to be of any practical benefit in real-world applications. By codifying rich emotional behaviour into arbitrary categorisations, the engineering approach implicitly tries to fit the problem of affect inference into a preconceived framework based on assumptions that may not even hold true in human-machine interactions. Consequently, this account fails to incorporate the social and cultural context that is necessary to give emotional behaviour its true meaning.

From a computational perspective, there is no clear consensus regarding the notion of context and as to what it means, what it includes and what role it plays in HCI (Dourish, 2004). In affective computing, the term has been used in an equally uncertain manner to refer to everything outside of the behavioural pattern being studied; and therefore includes an infinite number of personal, cultural, historical, environmental and situational factors. So while there is acknowledgement of the significance of context in the interpretation of

emotion, the task of practically dealing with it has been sidelined by reductionist attempts to optimise the sensing and measurement technology first. The use of experimentally controlled and technically easier to deal with posed or acted data has traditionally served to justify precisely this purpose. Now when the field seems to have reached a saturation point in terms of performance evaluated through recognition accuracies, the issue can no longer be avoided. The fact that the developed techniques have not generalised to real-world settings has rightly shifted the focus from posed/acted data to naturalistic data.

However, it is when dealing with naturalistic emotional data that one actually understands the limitations of the engineering stance in arriving at an appropriate conceptualisation of affect. This research has, for example, uncovered the problems associated with affect measurement and interpretation and has discussed the subtlety and ambiguity of emotions, their co-occurrence, as well as individual differences in their expression, regulation and perception. Issues related to the nature of affective interventions and their reliability, the emergent ethical issues and the influence on subsequent user emotional behaviour, have further been highlighted. The mapping from physical signs to affect constructs is an approximation by definition and the utility of such an approach has not been evaluated in sufficient samples, let alone across cultures. The very fact that current systems cannot be compared amongst themselves shows the limits of generalisation.

Despite these concerns, the engineering stance continues to dominate the field and efforts to build standardised datasets, evaluation metrics and emotion representation languages continue. While it may be argued that computers, as information systems, need to treat emotions as information at some level of abstraction, it may be more productive to determine that right form and level of abstraction by considering alternative and importantly, interdisciplinary approaches, that draw on both technology design methods as well as social and cultural analysis (Sengers, 2005). Arguing for a convergence of multiple investigative paradigms in affective computing, Muller (2004) cautions that lack of knowledge about part-whole relationships of components of user experience becomes particularly important with regard to affective aspects of user experience. He proposes taking insights from ethnographic observation and analysis, and to experiment with them conceptually in design explorations. He further suggests engaging with actual/potential users to explore the diversity of their concepts and attitudes about relating emotionally with computers. The knowledge acquired through these experiences may then inform ideas for subsequent formal hypothesis testing.

Such an inter-disciplinary approach is endorsed by Sengers (2005) when highlighting the futility of trying to engineer experiences and has been instantiated in several projects through the concept of technology probes. Technology probes are simple, flexible and adaptable technologies serving the social science goal of collecting information about the use and users of the technology in a real-world setting, the engineering goal of field-testing the technology, and the design goal of inspiring users and researchers to think of about new technologies to support their needs and desires (Hutchinson, et al., 2003). Examples of practical deployment include eMoto (Sundstrom, Stahl, & Hook, 2007) and the Affective Diary

(Stahl, Hook, Svensson, Taylor, & Combetto, 2009) that make use of technology probes in a user-centered design process followed with exploratory end-user studies in an attempt to define, refine and explore the boundaries between user and system roles in emotional communication.

The Affective Diary improvises on conventional diary keeping by recording bodily memorabilia via a combination of sensors and mobile media alongside memorable notes in digital form. Crucially, the output is represented in an open-ended manner through ambiguous visualisation of colours and abstract forms to allow reflection and recollection of emotional events in a personally meaningful manner. eMoto on the other hand is an emotional messaging system that uses emotional-signalling gestures as input to render a message background of colours, shapes and animations to express the emotional content (Sundstrom, et al., 2007). It relies on active user participation in the interpretation of emotional experiences by provoking users to reflect upon their subjective experiences. The technology here seeks to augment the emotional experiences and functions as a medium to channel the emotional communication in mobile messaging. The concept evolved from what the authors term as in situ informants methodology leading them to uncover the diffused nature of emotions in an interaction and the inseparability of their meaning from overall context.

Similarly, Leahu et al. (2008) describe a design study exploring the possibilities for affective technology beyond simplistic notions of affect mapping by understanding how humans negotiate meaning given their own objective signals and their subjective emotions. They show that even when relying on the same input channels as in the information processing model of affect, it is not these objective signals that narrow down the space of interpretation and determine a singular meaning, but rather the place where emotion is 'recovered', whether in the person's mind or in a machine. They explore the relationships between objective measurable signals and their subjective meaning using reflective analysis of a map-based artwork. By overlaying users' galvanic skin response onto city maps as they wander around they are able to create a compelling account of physiological arousal along their routes. The mapped information is used as a mnemonic to trigger subjective events as well as to build collective readings of mapped space to indicate locations corresponding to general interest, excitement or security.

Thus, in affective computing, design-based approaches symbolise a shift in purpose from modelling affect to supporting affect interpretation instead. They are based on the notion that not only is emotion mediated by social and cultural situations, but it is also used to enact and sustain those settings. In other words, emotions are shaped not only by their expression but also by their reception. This forms the basis of adopting an interactional approach to emotion as an alternative to the information-processing one (Boehner, et al., 2007). The interactional approach discounts the objective view of emotions and embraces its ambiguity and subjectivity by actively involving users in meaning making. Strategies for evaluation are then informed from phenomenological approaches and interpretive inquiry using personal

accounts and reflection probes, amongst other methods, to understand the usage of technology and its relationship to practice (Sengers et al. 2008; Sengers & Gaver, 2006). Evaluation thus shifts from optimising quantitative criterion like recognition accuracy or probability, to instead assessing the design goals of how a system is received in practice.

In effect, the vision of emotionally intelligent machines often falls short of what it attempts to accomplish partly because of the complexity of the phenomenon itself, but mostly because of the assumptions about the nature of emotion communication in human-machine interactions. In engineering affect we run the risk of imposing a definition and form that might alter the way affective communication ordinarily takes place. As discussed, simplistic abstractions of emotion into objective and well-defined technical specifications face numerous methodological and practical challenges. In contrast, the design-oriented stance relies on a more holistic study of technology use to account for affect in all its complexity and offers a reasonable alternative to explore the concept of emotionality in computer systems. The field of affective computing can hugely benefit from comprehensive accounts of design-based studies, ideally longitudinal, to guide engineering efforts towards more practical and user-friendly emotion technologies.

6.4 Future Work

The preceding discussion sets the stage for an array of exciting possibilities to extend this research. In general, several directions for future work can be identified following the results and observations from this research.

Alternative conceptualisation

Even though emotional communication is an important aspect of our everyday social interactions it seems that our ability to verbalise or articulate emotion perception in words is extremely impoverished and highly dependent on active vocabulary. This was consistently observed during the annotation process during which raters could identify ‘something’ but found it quite difficult to express this explicitly in words. This highlighted the difficulty in formalising emotional experience into discrete categories. The categorisation approach also presented the difficulty of identifying precise boundaries of expressions to map onto distinct affect states.

Keeping these constraints in view, perhaps a more pragmatic approach would be to map behavioural signals onto broader learning related concepts signifying conducive or obtrusive behaviour along with an intensity component. This would mean adopting a dimensional approach to model emotions with the dimensions assuming a domain relevant character. Making the measured construct a little more abstract can facilitate a more flexible judgement procedure while also reducing the scope of terminological confusion and any cross-cultural incompatibilities. The broad categorisation of emotions on the lines of as activating/deactivating (Pekrun, et al., 2002), or of pertinent behaviour as on/off task (Baker, 2007; Kapoor & Picard, 2005), are relevant examples of such conceptualisations.

Such a representation would not only eliminate the need to accurately identify boundaries of emotional episodes, but would also facilitate labelling in a continuous manner as part of the ongoing interaction. Akin to what Cowie (2009) describes as 'trace-like representations', this would enable capturing of a time-varying record of perceived emotional content or overall affective quality. One could, for example, visualise a labelling session wherein a coder, debriefed about noticing a behaviour of interest, say level of engagement, watches the whole interaction video and is supposed to click a single button whenever he/she perceives anything significant. The density of such markings across multiple coders, preferably pre-screened for their nonverbal decoding ability, would then highlight areas of relevant changes to focus on. The coder would no longer be burdened with assigning a specific category to a portion of video which in most cases is stripped out of context. The efficacy of such a judgement strategy is also supported by experimental findings on the suitability of a dimensional decoding strategy in the case of partial or ambiguous emotional expressions (Mendolia, 2007), as are likely to occur in naturalistic data.

Compromise on functionalities: separation of measurement and meaning

Explicitly assigning emotional meaning onto nonverbal signs is hard even for human raters especially when provided with limited context information. Emotional judgments are extremely subjective and ambiguous, therefore difficult to formalise in rules and procedures. Factors like gender, culture, mood, emotional intelligence, or even disorders like autism, affect emotion judgements (Jack, Blais, Scheepers, Schyns, & Caldara, 2009; Chakrabarti & Baron-Cohen, 2006; Elfenbein, Marsh, & Ambady, 2002). People who are more empathetic, for example, would interpret nonverbal signs differently than those who are not and may not even be consistent about their judgements when in a different state of mind. To actually formulate rules for computers to be able to do this is going to be difficult.

A compromise can be reached by pursuing two parallel methodological approaches focusing on: (1) measurement of behavioural signals by the computers; and (2) interpretation of these signals by actively involving the user(s). This would conform to Sengers and Gaver's (2006) design proposition of 'downplaying the systems authority' when dealing with an interpretively flexible concept like emotional interaction. This distinction can reduce the complexity of emotion judgment in human-computer interaction by aligning functionality with respective abilities – computers with continuous, objective measurement, and user(s) with meaning making and high-level interpretation.

As discussed in the previous section, a potential way to proceed in terms of measurement is to recover a more global level picture of the affective quality by tracking unusual/interesting patterns from the continuously monitored behavioural signs. Further research could then pursue in finding associations between the relevant affect construct(s) and the selected feature descriptors. Efforts towards real-time automatic FACS coding in spontaneous and naturally evoked data (Valstar, Gunes, & Pantic, 2007; Bartlett, et al., 2006) for example, can be instrumental in carrying out objective measurement of visual signs for real-world applications.

Having pre-processed the behavioural cues in terms of relevant constructs, the challenge then is to make use of this information in an effective manner. This is where shifting the onus of interpretation, as well as action, onto the user can help over ride the complexity of letting a computer do the same. By presenting the measured emotional behaviour in creative visualisations, the user can be actively involved in meaning attribution, personal discovery and reflection (Leahu, Schwenk, & Sengers, 2008). Affect modelling would thereby assume a more personal and subjective character allowing the possibility to reflect on individual emotional experiences. In the perceived application context, this would in fact serve as a form of feedback prompting learners to revisit or consider interesting episodes during their interaction, possibly reinforcing learning in the process. The example of the Affective Diary which represents affective body memorabilia in abstract visual representations using shapes and colours is inspiring in this regard (Stahl, et al., 2009). A similar example is that of Emotional Flowers (Bernhaupt, Boldt, Mirlacher, Wilfinger, & Tscheligi, 2007), which uses the concept of an ambient display to represent the emotional states of game participants using flowers that shrink or grow, depending on the emotions measured through facial expressions.

In summary, one can achieve a realistic notion of recognising emotionally salient events by balancing a lower-level understanding of behavioural signals with the human ability to make sense of this information and interpret it in a personally meaningful way. When and how during the interaction this information should be provided would depend on the design of the learning environment and the individual user profile.

Exploring intentionality

Affective computing envisages truly effective human-machine interactions as being affect-sensitive. Yet the field is both motivated, and influenced by an understanding of emotion in an environment, that of person to person, that differs from its eventual application, person to machine. In light of the inconsistencies of expressivity observed in my data – the importance of task type on emotion expression and the distinction between felt emotion and situational response, intentional affective interaction with a machine was proposed as a promising solution. The use of technology probes (Hutchinson, et al., 2003) along the lines of Gaver et al.'s (1999) cultural probes can be helpful to pursue this idea for a more inclusive design approach. While cultural probes use materials like diaries and cameras to encourage reporting of subjective experiences and reflections, technology probes deploy provocative artefacts in real use contexts to stimulate design ideas and to understand user experience in an open-ended manner. This would allow for a shift from the psychological and subjective definition of emotions to a more phenomenological and shared one.

Sections 6.3.2 and 6.3.3 discussed this concept at length.

Using eye-tracking during labelling

An interesting avenue of future research is to incorporate eye-tracking during the labelling process in order to identify the regions of interest as well to help understand how humans

perceive emotions from visual clues. When used in a continuous emotion judgment task, the fixation times and traces of the focused regions can give vital clues for defining appropriate features as well as providing estimates of the temporal window of evidence required for segmentation. The utility of such an approach was demonstrated by the results of Jack et al. (2009) who used eye tracking during an emotional labelling task to uncover significant cultural differences in the decoding of even the so called universal facial expressions of emotion.

Additional modalities

While I have focussed on the visual modality in my dissertation, additional modalities can also be incorporated once the traditional notion of affect as information is discarded and the more interactional aspect is included in the design. In the former approach, multi-modal affect detection would make the task of annotation even more complex and perhaps even more time-consuming. Analysing the annotations obtained on natural data, Douglas-Cowie et al. (2005) show how the audio and visual modalities interact in complex way to make varying contributions along the activation and evaluation dimensions. In general, issues related to the fusion, temporal structure and temporal correlation between multimodal cues remain a virtually unexplored area especially using naturalistic databases (Pantic, 2009), but is nevertheless an important line of enquiry.

The data as a resource

The database compiled during this thesis is an important resource with potential merit for further qualitative as well as quantitative analysis. Considering that there are limited repositories of naturalistic data currently available (Pantic, 2009), it can serve to not only evaluate the robustness of facial expression analysis methods, but to also understand the differences between posed and naturalistic data. The latter is growing as an interesting area of study with several researchers trying to draw out the distinction between acted and natural nonverbal displays in terms of dynamics and configuration (Valstar, Gunes, & Pantic, 2007; Cohn & Schmidt, 2004) with potential applications including automatic detection of deception and pain (Littlewort, Bartlett, & Lee, 2007).

6.5 Final remarks

The objective set out at the onset of this research was to evaluate the feasibility of using facial affect analysis to model the emotional state of a learner. Apart from the challenges associated with the perception and measurement of affect, this chapter has discussed additional issues that require due consideration by virtue of the application context. Our own inability to describe and achieve consensus on emotional behaviour, as well as the individual differences in encoding/decoding nonverbal behaviour, makes it is unlikely that computers will be able to perform the high-level interpretation necessary for emotion inference, at least in the foreseeable future. Nevertheless, there is substantial motivation to pursue this aim and my proposition is that the most pragmatic way this can be achieved is to think beyond

function approximation of specific patterns of behavioural signals to actively engaging the user in meaning making. As Picard and Klein observe,

“Just because humans are the best example we know, when it comes to emotional interaction it does not mean that we have to duplicate their emotional abilities in machines, which may not even be possible.” (Picard & Klein, 2002, p. 154)

This research concludes that we need to understand the nature and expression of emotion in the context of technology use and this may mean exploring alternative ways of what is perhaps a qualitatively different form of emotion expression and communication. In the introductory chapter I categorised the main issues in affective computing research along conceptual, methodological, technical and ethical constraints. The conclusions and proposed directions for future work address each of these issues to advance the problem definition to a more practical re-definition.

Bibliography

Abou-Moustafa, K. T., Cheriet, M., & Suen, C. Y. (2004). On the Structure of Hidden Markov Models. *Pattern Recognition Letters , 25*, 923-931.

Abrilian, S., Devillers, L., & Martin, J.-C. (2006). Annotation of Emotions in Real-Life Video Interviews: Variability between Coders. *Int. Conf. Language Resources and Evaluation.* Genoa, Italy.

Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *HCI International.*

Adolphs, R. (2002). Recognising Emotion from Facial Expressions: Psychological and Neurological Mechanisms. *Behavioral and Cognitive Neuroscience Reviews , 1* (1), 21-62.

Afzal, S., & Robinson, P. (2009). Natural Affect Data - Collection and Annotation in a Learning Context. *Affective Computing & Intelligent Interaction.* Amsterdam.

Afzal, S., Morrison, C., & Robinson, P. (2009). Intentional affect: An alternative notion of affective interaction with a machine. *Human-Computer Interaction.* Cambridge, UK.

Afzal, S., Sezgin, T. M., Gao, Y., & Robinson, P. (2009). Perception of Emotional Expressions in Different Representations Using Facial Feature Points. *Affective Computing & Intelligent Interaction.*

Aist, G., & B. Kort, R. R. Experimentally Augmenting an Intelligent Tutoring System with Human-Supplied Capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens.

Alexander, S. T., Hill, S., & Sarrafzadeh, A. (2005). How do Human Tutors Adapt to Affective State? *Proceedings of User Modelling.* Edinburgh, Scotland.

Alpaydin, E. (2004). *Introduction to Machine Learning.* Cambridge, MA : The MIT Press.

Alsmeyer, M., Luckin, R., & Good, J. (2008). Developing a Novel Interface for Capturing Self Reports of Affect. *CHI* (pp. 2883-2888). Florence, Italy: ACM.

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science , 16* (5), 403-410.

Amershi, S., Conati, C., & Maclaren, H. (2006). Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games. Workshop on Motivational and Affective Issues in ITS. *Intelligent Tutoring Systems*, (pp. 21-28).

Axelrod, L., & Hone, K. (2005). Uncharted Passions: User Displays of Positive Affect with an Adaptive Affective System. *Affective Computing and Intelligent Interaction (ACII)* (pp. 890-897). Springer-Verlag.

Bachorowski, J., & Braaten, E. B. (1994). Emotional intensity: Measurement and theoretical implications. *Personality and Individual Differences , 17*, 191-199.

Bakeman, R., & Gothman, J. M. (1997). *Observing interaction: An introduction to sequential analysis.* UK: Cambridge University Press.

Baker, R. S. (2007). Modeling and Understanding Students' Off-Task Behaviour in Intelligent Tutoring Systems. *CHI* (pp. 1059-1068). San Jose, USA: ACM.

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of Adults with Asperger Syndrome or High Functioning Autism, & Normal Sex Differences. *Journal of Autism & Developmental Disorders , 34* (2), 163-175.

Baron-Cohen, S., Golan, O., Wheelwright, S., & Hill, J. (2004). Mind Reading: The Interactive Guide to Emotions. London: Jessica Kingsley Publishers.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behavior. *7th International Conference on Automatic Face and Gesture Recognition (FGR 2006)*, (pp. 223-230).

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, (pp. 568-573).

Bassilli, J. N. (1978). Facial Motion in the Perception of Faces & of Emotional Expression. *Journal of Experimental Psychology: Human Perception & Performance , 4* (3), 373-379.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Noth, E. (2003). How to Find Trouble in Communication. *Speech Communication , 40* (1-2), 117-143.

Beale, R., & Creed, C. (2009). Affective Interaction: How Emotional Agents Affect Users. *Intl. J. Human-Computer Studies* , 755-776.

Bernhardt, D., & Robinson, P. (2007). Detecting Affect from Non-Stylised Body Motions. *International Conference on Affective Computing and Intelligent Interaction.* Lisbon, Portugal.

Bernhaupt, R., Boldt, A., Mirlacher, T., Wilfinger, D., & Tscheligi, M. (2007). Using Emotion in Games: Emotional Flowers. *Iinternational Conference on Advances in Computer Entertainment Technology* (p. 48). ACM.

Blake, R., & Shiffrar, M. (2007). Perception of Human Motion. *Annual Review Psychology , 58*, 47-73.

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researche , 13* (6), 4-16.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2005). Affect: from Information to Interaction. *4th Decennial Conference on Critical Computing: between Sense and Sensibility* (p. 68). ACM.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How Emotion is Made and Measured. *International Journal of Human-Computer Studies , 65*, 275-291.

Boekaerts, M. (1992). The adaptable learning process: Initiating and maintaining behavioural change. *Journal of Applied Psychology: An International Review , 41*, 377-397.

Boekaerts, M. (2002). The online motivation questionnaire: A self-report instrument to assess students' context sensitivity. *New Directions in Measures and Methods* , 77-120.

Boekaerts, M. (2003). Towards a model that integrates motivation, affect and learning. British Journal of Educational Psychology Monograph. *BJEP Monograph Series II, Number 2 - Development and Motivation , 1* (1), 173-189.

Branco, P., Firth, P., Encarnacao, L., & Bonato, P. (2005). Faces of Emotion in Human-Computer Interaction. *Conference on Human Factors in Computing Systems (CHI)* (pp. 1236-1239). Portland, Oregon, USA: ACM New York, NY, USA.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How People Learn: Brain, Mind, Experience and School.* Washington, DC: National Academy Press.

Brave, S., & Nass, C. (2002). Emotion in Human-Computer Interaction. In J. Jacko, & A. Sears (Eds.), *Handbook of Human-Computer Interaction* (pp. 251-271). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that Care: Investigating the Effects of Orientation of Emotion Exhibited by an Embodied Computer Agent. *International Journal of Human-Computer Studies , 62*, 161-178.

Buck, R. (1979). Measuring Individual Differences in the Nonverbal Communication of Affect: The Slide-Viewing Paradigm. *Human Communication Research , 6* (1), 41-57.

Camtasia Studio. (2006). *Version 3.1* . TechSmith Software.

Chakrabarti, B., & Baron-Cohen, S. (2006). Empathizing: neurocognitive developmental mechanisms and individual differences. *Understanding Emotions , 156*, 403-17.

Chen, X.-w., & Huang, T. (2003). Facial Expression Recognition: A clustering-based approach. *Pattern Recognition Letters , 24*, 1295-1302.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial Expression Recognition from Video Sequences: Temporal and Static Modelling. *Computer Vision and Image Understanding , 91*, 160-187.

Cohn, J. F. (2006). Foundations of Human Computing: Facial Expression and Emotion. *Intl. Conf. on Multimodal Interfaces (ICMI).* Banff, Canada: ACM.

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *Wavelets, Multiresolution and Information Processing , 2*, 1-12.

Conati, C. (2002). Probabilistic Assessment of User's Emotions in Educational Games. *Journal of Applied Artificial Intelligence , 16*, 555-575.

Conati, C., & Maclaren, H. (2004). Evaluating a Probabilistic Model of Student Affect. *7th International Conference on Intelligent Tutoring Systems (ITS).* Maceio, Brazil.

Conati, C., & Zhou, X. (2002). Modelling Students' Emotions from Cognitive Appraisal in Educational Games. *Intelligent Tutoring Systems.*

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences , 364* (1535), 3515.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication , 40*, 5-32.

Cowie, R., & McKeown, G. (2009). The challenges of dealing with distributed signs of emotion: theory and empirical evidence. *Affective Computing and Intelligent Interaction (ACII)* (pp. 351-356). Amsterdam: IEEE.

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks , 18*, 371-388.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine , 18* (1), 32-80.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media , 29*, 241-250.

D'Mello, S. K., Craig, S. D., Gholson, B., Franklin, S., Picard, R. W., & Graesser, A. C. (2005). Integrating Affect Sensors in an Intelligent Tutoring System. *Workshop on Affective Interactions: The Computer in the Affective Loop Workshop, IUI.*

D'Mello, S., & Graesser, A. (2007). Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments. *International Conference on Artificial Intelligence in Education.* Los Angeles.

Darwin, C. (1872). *The Expression of The Emotions in Man and Animals.* London: Murray.

de Vicente, A., & Pain, H. (2002). Informing the Detection of the Students' Motivational State: An Empirical Study. *Intelligent Tutoring Systems.*

de Vicente, A., & Pain, H. (1998). Motivation Diagnosis in Intelligent Tutoring Systems. In B. P. Goettl, C. Halff, C. L. Redfield, & V. J. Shute (Ed.), *Intelligent Tutoring Systems*, (pp. 86-95). Texas.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks , 18*, 407-422.

Dictionary, O. E. (2008, Dec). *OED online*. Retrieved 2009, from Oxford English Dictionary: http://dictionary.oed.com/cgi/entry/00329705/

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Towards An Affect-Sensitive Auto-Tutor. *IEEE Intelligent Systems , 22* (4), 53.

D'Mello, S., Taylor, R., Davidson, K., & Graesser, A. (2008). Self Versus Teacher Judgements of Learner Emotions During a Tutoring Session with AutoTutor. *Intelligent Tutoring Systems (ITS)*, (pp. 9-18).

Douglas-Cowie, E., & Members, W. (2004). *Deliverable D5c-Preliminary plans for exemplars: Databases.* Project Deliverable of Humaine Network of Excellence.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., et al. (2007). The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. *Affective Computing and Intelligent Interaction* (pp. 488-500). Springer Verlag.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, R., Abrillian, S., et al. (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. *Interspeech*, (pp. 813-816).

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing , 8*, 19-30.

du Boulay, B., & Luckin, R. (2001). Modelling Human Teaching Tactics and Strategies for Tutoring Systems. *International Journal of Artificial Intelligence in Education , 12*, 235-256.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). US: Wiley-Interscience.

Edwards, K. (1998). The Face of Time: Temporal Cues in Facial Expressions of Emotion. *Psychological Science , 9* (4), 270-276.

Ekman, P. (1982). *Emotion in the Human Face.* Cambridge: Cambridge University Press.

Ekman, P., & Freisen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology , 17*, 124-129.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto, CA: Consulting Psychologists Press.

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

El Kaliouby, R. (2005). *Mind-Reading Machines: Automated Inference of Complex Mental States.* PhD Dissertation, University of Cambridge, Computer Laboratory, Cambridge, UK.

El Kaliouby, R., & Robinson, P. (2005). Generalization of a Vision-Based Computational Model of Mind-Reading. *International Conference on Affective Computing and Intelligent Interaction.* Beijing, China.

El Kaliouby, R., & Robinson, P. (2004). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. *IEEE Workshop on Real-Time Vision for Human-Computer Interaction, Computer Vision and Pattern Recognition.*

El Kaliouby, R., & Teeters, A. (2007). Eliciting, Capturing and Tagging Spontaneous Facial Affect in Autism Spectrum Disorder. *International Conference on Multimodal Interfaces.* Aichi, Japan.

ELAN. (n.d.). Retrieved 2007-08, from http://www.lat-mpi.eu/tools/elan/

Elfenbein, H. A., Marsh, A. A., & Ambady, N. (2002). Emotional Intelligence and the Recognition of Emotion from Facial Expressions. In L. F. Barrett, & P. Salovey, *The Wisdom of Feelings: Processes Underlying Emotional Intelligence* (pp. 37-59). New York: Guilford Press.

Essa, I. A. (1997). Coding, Analysis, Interpretation, and Recognition of Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 19* (7).

Fasel, B., & Luettin, J. (2003). Automatic Facial Expression Analysis: A Survey. *Pattern Recognition , 36* (1), 259-275.

Field, A. (2009). *Discovering Statistics Using SPSS .* London: Sage Publications Ltd.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological. Bulletin , 76* (5), 378–382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). The Measurement of Interrater Agreement. In *Statistical Methods for Rates & Proportions* (3 ed., pp. 598-626). NJ: Wiley.

Fontaine, J., Scherer, K. R., Roesch, E., & Ellsworth, P. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science , 18* (12), 1050-1057.

Frank, M. G., Juslin, P. N., & Harrigan, ,. J. (2005). Technical Issues in Recording Nonverbal Behaviour. In J. A. Harrigan (Ed.), *The New Handbook of Methods in Nonverbal Behaviour Research.* Oxford University Press.

Fridlund, A. J. (1994). *Human Facial Expression: An Evolutionary View.* San Diego, CA: Academic Press.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology , 60* (2), 229-240.

Fridlund, A. J. (n.d.). *What do facial expressions express?* . Retrieved March 2010, from http://www.sscnet.ucla.edu/anthro/bec/papers/Fridlund_Facial_Expressions.PDF

Friedman, H. S., Prince, L. M., Riggio, R. E., & DiMatteo, M. R. (1980). Understanding and assessing nonverbal expressiveness: the Affective Communication Test. *Journal of Personality and Social Psychology , 39* (2), 333-351.

Frith, C. (2009). The role of facial expressions in social interactions. *Philosophical Transactions B , 364* (1535), 3453.

Gaver, W., Dunne, T., & Pacenti, E. (1999). Cultural Probes. *Interactions , 6* (1), 21-29.

Groom, V., Nass, C., Chena, T., Nielsen, A., Scarborough, J. K., & Robles, E. (2009). Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human-Computer Studies , 67* (10), 842-849.

Gross, J. J., & John, O. P. (1995). Facets of emotional expressivity: Three self-report factors and correlates . *Personality and Individual Differences , 19*, 555-568.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations , 11* (1), 10-18.

Har-Peled, S., Roth, D., & Zimak, D. (2003). Constraint Classification for Multiclass Classification and Ranking. (S. Becker, & K. Obermayer, Eds.) *Advances in Neural Information Processing Systems , 15*.

Hassenzahl, M., & Tractinsky, N. (2007). User Experience- a research agenda. *Behaviour & Information Technology , 25* (2).

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures , 1* (1), 77–89.

Heritage, J. (1984). *Garfinkel and Ethnomethodology.* Polity Press.

Heylen, D., Ghijsen, M., Nijholt, A., & Akker, R. (2005). Facial Signs of Affect During Tutoring Sessions. In J. Tao J, & R. W. Picard (Ed.), *Affective Computing and Intelligent Interaction . LNCS 3784*, pp. 24-31. Springer-Verlag.

Hotelling, M. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Pschology , 24*, 498-520.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies , 59*, 1-32.

Hutchinson, H., Westerlund, B., Bederson, B., Druin, A., Beaudouin-Lafon, M., Evans, H., et al. (2003). Technology probes: inspiring design for and with families. *SIGCHI conference on Human Factors in Computing Systems* (pp. 17-24). Florida, USA: ACM New York, NY, USA.

Ickes, W. (2001). Measuring Empathic Accuracy. In J. A. Hall, & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219-241). Mahwah, NJ: Erlbaum.

Immordino-Yang, M. H., & Damasio, A. (2007). We Feel, Therefore We Learn: The Relevance of Affective and Social Neuroscience to Education, Mind, Brain and Education. *Mind, Brain and Education , 1* (1).

Isomursu, M., Tahti, M., Vainamo, S., & Kuutti, K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile application. *Journal of Human-Computer Studies , 65*, 404-418.

Izard, C. E. (1993). Four Systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review , 100*, 68-90.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural Confusions Show that Facial Expressions Are Not Universal. *Curent Biology , 19* (18), 1543-1548.

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 22* (1), 4-37.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys , 31* (3), 264-323.

Jaques, P. A., & Vicari, R. M. (2007). A BDI approach to infer students' emotions in an intelligent learning environment. *Computers & Education , 49*, 360-384.

Jarvenoja, H., & Jarvela, S. (2005). How students describe the sources of their emotional and motivational experiences during the learning process: A qualitative approach. *Learning and Instruction , 15* (5), 465-480.

Johansson, G. (1973). Visual perception of biological motion & a model for its analysis. *Perception & Psychophysics , 14*, 201-211.

Kanade, T., Cohn, J., & Tian, Y. L. (2000). Comprehensive database for facial expression analysis. *IEEE FG.* France.

Kapoor, A., & Picard, R. W. (2005). Multimodal Affect Recognition in Learning Environments. *13th Annual ACM International Conference on Multimedia.* Singapore.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic Prediction of Frustration. *Journal of Human-Computer Studies , 65* (8), 724-736.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons.

King., L. A., & Emmons, R. A. (1990). Conflict over emotional expression: Psychological and physical correlates. *Journal of Personality and Social Psychology , 58*, 864-877.

Kipp, M., Neff, M., & Albrecht, I. (2007). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation-Special Issue on Multimodal Corpora*.

Kleinginna, P. R., & Kleinginna, A. M. (1981). A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition. *Motivation and Emotion , 5* (4), 345-379.

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. *IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges*, (pp. 43-48). Madison.

Kring, A., Smith, D. A., & Neale, J. M. (1994). Individual differences in dispositional expressiveness: Development and validation of the emotional expressivity scale. *Journal of Personality & Social Psychology , 66*, 934-949.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics , 33*, 159-174.

Larsen, R. J., & Diener, E. (1985). A multitrait-multimethod examination of affect structure: Hedonic level and emotional intensity. *Personality and Individual differences , 6*, 631-636.

Larsen, R. J., & Diener, E. (1987). Affect Intensity as an individual difference characteristic: A review. *Journal of Research in Personality , 21*, 1-39.

Larsen, R. J., & Fredrickson, B. L. (1999). Measurement Issues in Emotion Research. *Well-being: The foundations of hedonic psychology* , 40-60.

Lazarus, R. S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

Leahu, L., Schwenk, S., & Sengers, P. (2008). Subjective Objectivity: Negotiating Emotional Meaning. *Designing Interactive Systems (DIS)* (pp. 425-434). Cape Town, South Africa: ACM.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. (S. P. Lajoie, & S. J. Derry, Eds.) *Computers as Cognitive Tools* , 75–105.

Levenson, R. W. (1983). Personality Research and Psychophysiology: General Considerations. *Journal of Research in Personality , 17*, 1-21.

Liao W, W., Zhang, W., Zhu, Z., Ji, Q., & Gray, W. D. (2006). Toward a decision-theoretic framework for affect recognition and user assistance. *International Journal of Human-Computer Studies , 64*, 847-873.

Lisetti, C., & Schiano, D. (2000). Facial expression recognition: Where Human Computer Interaction, Artificial Intelligence and Cognitive Science Intersect. *Pragmatics and Cognition , 8* (1), 185-235.

Litman, D., & Forbes, K. (2003). Recognising emotions from student speech in tutoring dialogues. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2007). Faces of Pain: Automated Measurment of Sponaneous Facial Expressions of Genuine and Posed Pain. *International Conference on Multimodal Interfaces* (pp. 15-21). ACM.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of Facial Expression Extracted Automatically from Video. *Image and Vision Computing , 24* (6), 615.

Manusov, V. L. (2005). *Sourcebook of Nonverbal Measures: Going Beyond Words.* Lawrence Erlbaum Associates.

Martin, J., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., & Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. *Intelligent Virtual Agents*, *LNAI 3661*, pp. 405-417.

Martinez, W. L., & Martinez, A. R. (2005). *Exploratory Data Analysis with MATLAB.* US: Chapman & Hall/CRC Press.

Mavrikis, M., Maciocia, A., & Lee, J. (2007). Towards Predictive Modelling of Student Affect from Web-Based Interactions. *Artificial Intelligence in Education.* Los Angeles.

Mendolia, M. (2007). Explicit Use of Categorical and Dimensional Strategies to Decode Facial Expressions of Emotion. *Journal of Nonverbal Behaviour , 31*, 57-75.

Merill, D. C., Reiser, B. J., Trafton, J. G., & Ranney, M. (1992). Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. *Journal of the Learning Sciences , 2*, 277-305.

Meyer, D. K., & Turner, J. C. (2002). Discovering emotion in classroom motivation research. *Educational Psychologist , 37*, 107-114.

Missaoui, O., & Frigui, H. (2008). Optimal feature weighting for the Continuous HMM. *19th International Conference on Pattern Recognition (ICPR)* (pp. 1-4). IEEE.

Mori, M. (1970). The Uncanny Valley. *Energy , 7* (4), 33-35.

Muller, M. (2004). Multiple Paradigms in Affective Computing, Interacting with Computers. *16* (4), 759-768.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues , 56* (1), 81-103.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are Social Actors. *SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence (CHI)* (pp. 72-78). Boston, Massachusetts USA: ACM New York, NY, USA.

Nemiah, J. C., Freyberger, H., & Sifneos, P. E. (1976). Alexithymia: A view of the Psychosomatic Process. (O. W. Hill, Ed.) *Modern Trends in Psychosomatic Medicine , 2*, 26-34.

Nideffer, R. M. (1976). Test of Attentional and Interpersonal Style. *Journal of Personality and Social Psychology , 34*, 394-404.

Norman, D. A. (1988). *The psychology of everyday things.* Basic Books.

Nusseck, M., Cunningham, D. W., Wallraven, C., & Bulthoff, H. H. (2008). The Contribution of Different Facial Regions to the Recognition of conversatonal Expressions. *Journal of Vision , 8* (8), 1-23.

O'Regan, K. (2003). Emotion and e-learning. *Journal of Asynchronous Learning Networks , 7* (3), 78-92.

Oliver, N., & Horvitz, E. (2005). A Comparison of HMMs and Dynamic Bayesian Networks for Recognizing Office Activities. *Lecture Notes in Computer Science , 3538*, 199-208.

Ortony, A., Clore, G. L., & Collins, A. (1998). *The Cognitive Structure of Emotions.* Cambridge: Cambridge University Press.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural Universals in Affective Meaning.* Urbana: University of Illinois Press.

Pandzic, I. S., & Forchheimer, R. (2002). MPEG-4 Facial Animation: The Standard, Implementation & Applications. John Wiley & Sons.

Pantic, M. (2009). Machine Analysis of Facial Behaviour: Naturalistic & Dynamic Behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences , 364* (1535), 3505.

Pantic, M., & Bartlett, M. S. (2007). Machine Analysis of Facial Expressions. In K. Delac, & M. Grgic (Eds.), *Face Recognition* (pp. 377-416). Vienna, Austria: I-Tech Education and Publishing.

Pantic, M., & Patras, I. (2006). Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments From Face Profile Image Sequences. *IEEE Transactions on Systems, Man, and Cybernetics , 36* (2), 433-449.

Pantic, M., & Rothkrantz, L. J. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE , 91* (9), 1370-1390.

Pantic, M., & Rothkrantz, L. M. (2000). Automatic Analysis of Facial Expressions: The State of Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 12* (22).

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human Computing and Machine Understanding of Human Behaviour: A Survey. *Human Computing* , 47-71.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. *IEEE ICME.*

Park, S.-H., & Furnkranz, J. (2007). Efficient Pairwise Classification. *Lecture Notes in Computer Science , 4701*, 658-665.

Parkinson, B. (2005). Do Facial Movements Express Emotions or Communicate Motives. *Personality and Social Psychology Review , 9* (4), 278-311.

Pearson, K. (1901). On Line and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine , 6* (2), 559-572.

Pekrun, R. (2005). Progress and open problems in educational emotion research. *Learning and Instruction , 15*, 497-506.

Pekrun, R. (1992). The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators. *Applied Psychology , 41*, 359-376.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist , 37*, 91-105.

Peter, C., & Herbon, A. (2006). Emotion representation and physiology assignments in digital systems. *Interacting with Computers , 18*, 139-170.

Picard, R. W. (1997). *Affective Computing.* Cambridge: MIT.

Picard, R. W., & Klein, J. (2002). Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting with Computers , 14*, 141-169.

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., et al. (2004). Affective Learning-a manifesto. *BT Technology Journal , 22*, 253-269.

Pollick, F. E., Hill, H., Calder, A., & Paterson, H. (2003). Recognising Facial Expression from Spatially and Temporally Modified Movements. *Perception , 32*, 813-826.

Porayska-Pomsta, K., & Pain, H. (2004). Exploring Methodologies for Building Socially and Emotionally Intelligent Learning Environments. *Workshop on Social and Emotional Intelligence in Learning Environments (SEILE), Intelligent Tutoring Systems.* Brazil.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE , 77* (2), 257-285.

Rifkin, R., & Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research , 5*, 101-141.

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology , 51*, 649-660.

Riggio, R. E., & Riggio, H. R. (2005). Self-report measures of emotional and nonverbal expressiveness . In V. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 105-111). Mahwaj, NJ:: Erlbaum.

Rosenthal, R. (1979). *Sensitivity to nonverbal communication: The PONS test.* Johns Hopkins University Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology , 39*, 1161-78.

Russell, J. A., & Fernandez-Dols, J. M. (1997). *The psychology of facial expression.* Cambridge: Cambridge University Press.

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A Systems Approach to Appraisal Mechanisms in Emotion. *Neural Networks , 18* (4), 317-352.

Sarrafzadeh, A., Fan, C., Dadgostar, F., Alexander, S., & Messom, C. (2004). Frown gives game away: Affect sensitive tutoring systems for elementary mathematics. *IEEE Conference on Systems, Man and Cybernetics.* The Hague.

Scherer, K. L. (2005). What are emotions? And how can they be measured? *Social Science Information , 44* (4).

Scherer, K. R. (2005). Trends and developments: research on emotions. *Social Science Information*.

Scherer, K., & Ekman, P. (1982). *Handbook of Methods in Nonverbal Behaviour Research.* Cambridge, UK: Cambridge Univ. Press.

Schröder, M. (Ed.). (2008). *Elements of an EmotionML 1.0 W3C Incubator Group Report.* Retrieved Aug 2010, from http://www.w3.org/2005/Incubator/emotion/XGR-emotionml/

Schutz, P. A., Hong, J. Y., Cross, D. I., & Obson, J. N. (2006). Reflections on Investigating Emotion in Educational Activity Settings. *Educational Psychology Review , 18*, 343-360.

Schwaninger, A., Wallraven, C., Cunningham, D. W., & Chiller-Glaus, S. D. (2006). Processing of facial identity and expressions: a psychological, physiological and computational perspective. *Progress in Brain research , 156*, 321-343.

Sengers, P. (2005). The Engineering of Experience. *Funology* , 19–29.

Sengers, P., & Gaver, B. (2006). Staying open to interpretation: engaging multiple meanings in design and evaluation. *Proceedings of the 6th conference on Designing Interactive systems* (pp. 108-118). Pennsylvania, USA: ACM.

Sengers, P., Boehner, K., Mateas, M., & Gay, G. (2008). The Disenchantment of Affect. *Personal and Ubiquitous Computing , 12* (5), 347-358.

Sezgin, T. M., & Robinson, P. (2007). Affective Video Data Collection Using an Automobile Simulator. *Lecture Notes in Computer Science , 4738*, 770.

Skemp, R. R. (1971). *The Psychology of Learning Mathematics.* Hillsdale: NJ: Erlbaum.

Stahl, A., Hook, K., Svensson, M., Taylor, A. S., & Combetto, M. (2009). Experiencing the Affective Diary. *Personal and Ubiquitous Computing , 13* (5), 365-378.

Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions.* Cambridge University Press.

Sundstrom, P., Stahl, A., & Hook, K. (2007). In situ informants exploring an emotional mobile messaging system in their everyday practice. *International Journal of Human-Computer Studies , 65* (4), 388-403.

Takahashi, S., Aikawa, K., & Sagayama, S. (1997). Discrete mixture HMM. *IEEE International Conference On Acoustics Speech and Signal Processing*, *2*, pp. 971-974.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining* (pp. 487-567). Addison-Wesley.

Tcherkassof, A., Bollon, T., Dubois, M., Pansu, P., & Adam, J.-M. (2007). Facial Expressions of Emotions: A Methodological Contribution to the Study of Spontaneous and Dynamic Emotional Faces. *European Journal of Social Psychology , 37*, 1325-1345.

Thomas, S. M., & Jordan, T. R. (2001). Techniques for the production of point-light & fully illuminated video displays from identical recordings. *Behaviour Research Methods, Instruments, & Computers , 33* (1), 59-64.

Tian, Y.-L., Kanade, T., & Cohn, J. (2001). Recognising Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 23* (2), 97-115.

Tong, Y., Liao, W., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on Pattern Analysis and Machine Intelligence , 29* (10), 1683-1699.

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience , 3* (1), 71-86.

Turner, J. E., Husman, J., & Schallert, D. L. (2002). The importance of students'-goals in their emotional experience of academic failure: Investigating the precursors and the consequences of shame. *Educational Psychologist , 92*, 548-573.

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles Using Geometric Features. *ACM International Conference on Multimodal Interfaces* (pp. 38-45). ACM.

van Vuuren, S. (2006). Technologies that power pedagogical agents and visions for the future. *Special Issue of Educational Technology*.

VirtualDub. (n.d.). Retrieved 2007-08, from http://www.virtualdub.org/

Wagner, H. L., & Smith, J. (1991). Facial expression in the presence of friends and strangers. *Journal of Nonverbal Behavior , 15*, 201-214.

Wang, X. (2002). Feature Extraction and Dimensionality Reduction in Pattern Recognition and Their Application in Speech Recognition. *PhD Thesis* . Australia: Griffith University.

Ward, R. D., & Marsden, P. H. (2004). Affective Computing: Problems, Reactions and Intentions. *Interacting with Computers , 16*, 707-713.

Westerman, S. J., Gardner, P. H., & Sutherland, E. J. (2006). *Taxonomy of Affective Systems Usability Testing.* HUMAINE EC Network of Excellence Report D9g.

White, C. H., & Sargent, J. (2005). Researcher Choices and Practices in the Study of Nonverbal Communication. In V. L. Manusov (Ed.), *Sourcebook of Nonverbal Measures: Going Beyond Words.* Lawrence Erlbaum Associates.

Whitehall, J., Bartlett, M., & Movellan, J. (2008). Automatic Facial Expression Recognition for Intelligent Tutoring Systems. *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, *2008.*

Wierzbicka, A. (2003). *Cross-cultural pragmatics: The semantics of human interaction.* Berlin: Mouton de Gruyter.

Wosnitza, M., & Volet, S. (2005). Origin, direction and impact of emotions in social online learning. *Learning and Instruction , 15* (5), 440-464.

Zakharov, K., Mitrovic, A., & Johnston, L. (2007). Pedagogical Agents Trying on a Caring Mentor Role. *Artificial Intelligence in Education.* Los Angeles.

Zakharov, K., Mitrovic, A., & Johnston, L. (2008). Towards Emotionally-Intelligent Pedagogical Agents. *Lecture Notes in Computer Science. 5091*, pp. 19-28. Springer.

Zeng, Z., Fu, Y., Roisman, G. I., Wen, Z., Hu, Y., & Huang, T. S. (2006). Spontaneous Emotional Facial Expression Detection. *Journal of Multimedia , 1* (5), 1-8.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence , 31* (1), 39-58.

Zuckerman, M., & Larrance, D. (1979). Individual differences in perceived encoding and decoding abilities. In R. Rosenthal (Ed.), *Skill in Nonverbal Communication* (pp. 171-203). Cambridge, MA: Oelgeschlager, Gunn, & Hain.

Zuo, D. (2007). *An Efficient Framework for Video Annotation.* Cambridge: Computer Lab Summer UROP Report.