Technical Report

UCAM-CL-TR-636
ISSN 1476-2986

Number 636

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Mind-reading machines: automated inference of complex mental states

Rana Ayman el Kaliouby

July 2005

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 March 2005 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Newnham 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

People express their mental states all the time, even when interacting with machines. These mental states shape the decisions that we make, govern how we communicate with others, and affect our performance. The ability to attribute mental states to others from their behaviour, and to use that knowledge to guide one's own actions and predict those of others is known as *theory of mind* or *mind-reading*.

The principal contribution of this dissertation is the real time inference of a wide range of mental states from head and facial displays in a video stream. In particular, the focus is on the inference of complex mental states: the affective and cognitive states of mind that are not part of the set of basic emotions. The automated mental state inference system is inspired by and draws on the fundamental role of mind-reading in communication and decision-making.

The dissertation describes the design, implementation and validation of a computational model of mind-reading. The design is based on the results of a number of experiments that I have undertaken to analyse the facial signals and dynamics of complex mental states. The resulting model is a multi-level probabilistic graphical model that represents the facial events in a raw video stream at different levels of spatial and temporal abstraction. Dynamic Bayesian Networks model observable head and facial displays, and corresponding hidden mental states over time.

The automated mind-reading system implements the model by combining top-down predictions of mental state models with bottom-up vision-based processing of the face. To support intelligent human-computer interaction, the system meets three important criteria. These are: full automation so that no manual preprocessing or segmentation is required, real time execution, and the categorization of mental states early enough after their onset to ensure that the resulting knowledge is current and useful.

The system is evaluated in terms of recognition accuracy, generalization and real time performance for six broad classes of complex mental states—*agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*, on two different corpora. The system successfully classifies and generalizes to new examples of these classes with an accuracy and speed that are comparable to that of human recognition.

The research I present here significantly advances the nascent ability of machines to infer cognitive-affective mental states in real time from nonverbal expressions of people. By developing a real time system for the inference of a wide range of mental states beyond the basic emotions, I have widened the scope of human-computer interaction scenarios in which this technology can be integrated. This is an important step towards building socially and emotionally intelligent machines.

Acknowledgements

This dissertation was inspired and shaped through discussions with many people and was only made possible by the support of many more who have shared their love, prayers, time, and know-how with me. Without the support of each and every one of them it would not have been possible to undertake and successfully complete this challenging endeavour.

I would like to especially thank my supervisor, Peter Robinson, for giving me the opportunity to be part of this inspirational environment and for being supportive in every imaginable way. I would also like to thank Alan Blackwell for his valuable advice and for involving me with Crucible through which I have learnt tremendously, Neil Dodgson for his technical and moral support and Rosalind Picard and Sean Holden for their insightful comments on the dissertation.

Simon Baron-Cohen has and continues to provide the inspiration for many of the ideas in this dissertation. I would like to thank him and his group at the Autism Research Centre, especially Ofer Golan for sharing their thoughts, research and the Mind Reading DVD with us. Tony Manstead, Alf Linney, Alexa Wright and Daren McDonald for fruitful discussions on emotions and facial expressions. Special thanks are due to everyone who has volunteered in taking part in the various studies and experiments which I have undertaken, and especially those who have volunteered with their acting at the CVPR 2004 conference.

I would like to thank the former and current members of the Rainbow Group and other groups at the Computer Laboratory, especially Maja Vukovic, Mark Ashdown, Eleanor Toye, Jennifer Rode, Scott Fairbanks, Evangelos Kotsovinos and Chris Town for reviewing this dissertation at its various stages, and William Billingsley, Marco Gillies, Michael Blain and Tal Sobol-Shikler for their support. I am also grateful to the Computer Laboratory staff for their timely assistance on technical and administrative issues and for their daily encouragement, especially Lise Gough, Graham Titmus, Jiang He, Chris Hadley, Margarett Levitt, Amanda Hughes and Kate Ellis. At Newnham College, I would like to thank Katy Edgecombe and Pam Hirsch for the support they have shown me both at the research and at the personal level.

I would like to thank my "family" in Cambridge, especially Carol Robinson for making me feel part of her family and for taking care of Jana on many occasions, Catherine Gibbs, Helen Blackwell, Rachel Hewson and Amna Jarar for being there for me. I would like to thank my friends, especially Ingy Attwa for faithfully keeping in touch, and Dahlia Madgy for help with technical writing.

I have been blessed with the most supportive family, to whom I am very grateful. My parents-in-law Ahmed and Laila for never leaving me alone, and Houssam and Sahar for being supportive. My cousin Serah and my sisters Rasha and Rola for being wonderful and for help with reviewing everything I have written. My parents Ayman and Randa for always being there for me and for believing in me. Mom and dad, I wouldn't be where I am without your hard work, commitment and devotion; there is little I can say to thank you enough. Jana, thank you for being my sunshine and for putting up with a mom who spends most on her time on the "puter". Wael, thank you for helping me pursue this lifetime dream and for being the most supportive husband, best friend, advisor and mentor anyone can wish for.

This research was generously funded by the Cambridge Overseas Trust, British Petroleum, the Overseas Research Student Award, the Computer Laboratory's Neil Wiseman fund, the Marrion-Kennedy Newnham College Scholarship, and the Royal Academy of Engineering.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Mind-reading or theory of mind is the terminology used in psychology to describe people's ability to attribute mental states to others from their behaviour, and to use that knowledge to guide one's own actions and predict those of others [PW78, BRF+96, BRW+99]. It is *not*, as the word is often used in colloquial English, a mystical form of telepathy or thought reading. The mental states that people can express and attribute to each other include affective states or emotions, cognitive states, intentions, beliefs and desires. An essential component of social intelligence, mind-reading enables us to determine the communicative intent of an interaction, take account of others' interests in conversation, empathize with the emotions of other people and persuade them to change their beliefs and actions. The majority of people mind-read all the time, effortlessly and mostly subconsciously. Those who lack the ability to do so, such as people diagnosed with Autism Spectrum Disorders, have difficulties operating in the complex social world in which we live and are sometimes referred to as mind-blind.

Besides social competence, mind-reading has a central role in the processes underlying decision-making, perception and memory. Recent findings in neuroscience show that emotions regulate and bias these processes in a way that contributes positively to intelligent functioning. In decision-making, our own intentions, the social context and our appraisal of other people's mental states all affect the choices we make everyday. Indeed, the lack of emotional aptitude in people who have had traumatic brain injuries results in impaired reasoning. Consequently, there is an emerging consensus that emotions should be embedded within models of human reasoning.

Despite these important functions of mind-reading, existing human-computer interfaces are mostly mind-blind. They are oblivious to the mental states of their users, fail to reason about their actions and fail to take into account what they seem to know or not know. Such interfaces have no awareness of the user's attention, no concept of interruption and lack the ability to adapt to new circumstances or understand the context of their use. As Matthew Turk notes [Tur05], a computer may wait indefinitely for input from a user who is no longer there or decide to do irrelevant, computationally intensive tasks while a user is frantically working on a fast approaching deadline. Existing human-computer interfaces rarely take the initiative and lack persuasive power. They are mostly limited to a command and control interaction paradigm.

This interaction paradigm is especially restrictive as human-computer interaction (HCI) becomes more complex and new forms of computing emerge. Computing is no longer limited to a desktop setup and it is certainly no longer something people do at specific times during the day. Instead, computing is becoming ubiquitous, extending to mobile, embedded and wearable devices that are used by people in different interaction scenarios to perform a wide range of tasks. To utilize the full potential of these new technologies, user-aware interfaces that complement existing interaction methods are needed.

People have mental states all the time, even when interacting with a computer. These mental states can be affective, as in expressions of emotions, or cognitive as in revealing mental processes. Both shape the decisions we make and affect our performance. The ubiquity of computers, along with the fundamental role of mind-reading in communication and in decision-making combine to provide the motivation for this dissertation: building *Mind-Reading Machines*. I define these computing technologies as ones that are aware of the user's state of mind and that adapt their responses accordingly. Their goal is to enhance human-computer interaction through empathic responses, to improve the productivity of the user and to enable applications to initiate interactions with and on behalf of the user, without waiting for explicit input from that user.

The general notion of *Mind-Reading Machines*, and the specific research presented throughout this dissertation, is inspired by Rosalind W. Picard's vision and pioneering research on affective computing, a relatively new field of computing that relates to, arises from, or deliberately influences emotions [Pic97]. Since the inception of this field almost a decade ago, a number of researchers have charged ahead with building machines that have affective abilities. The expression-glasses [SFP99], psychophysiological measures such as skin conductance, heart rate and blood volume pulse [PS01, SFKP02, WS04], the pressure mouse [QP02] and automated facial expression analysis [PR03] are different approaches to the detection of emotion from nonverbal cues. In addition, several applications that adapt their responses based on the emotional state of the user have been proposed. These include affective learning [PPB+05], the learning companion [KRP01], computers that detect and respond to user frustration [KMP00], educational games [Con02, CGV02], telemedicine [LNL+03], social robots [BA02, BA03, LBF+04b, FND03, Sca01], rehabilitation technologies [Dau01b, Dau01a, DB02, PCC+03] and embodied conversational agents [CT99, GSV+03]. In the automobile industry, the automated inference of driver vigilance from facial expressions is gaining a lot of research attention [GJ04, JY02] and commercial interest [EHH+04]. Other researchers are concerned with studying the extent to which human-computer interaction is social [NM00, RN96] and would benefit from affective feedback and interventions [AS02, PS04]. Despite this significant progress, the vision of an automated, robust, affect-aware system remains elusive. The reasons why this remains a challenging endeavour and the aims of this dissertation are discussed in the following section.

1.2 Aims and challenges

Consciously or subconsciously, people employ a variety of nonverbal cues, such as vocal nuances, posture and facial expressions to communicate their emotions and mental states. The automated recognition of these cues is an open research problem, making the development of a comprehensive mind-reading machine an ambitious undertaking. This dissertation addresses the problem of detecting and recognizing people's mental states from head gestures and facial expressions.

The human face possesses excellent expressive ability and provides one of the most powerful, versatile and natural means of communicating a wide range of mental states.

Facial expressions communicate feelings, behavioural intentions, show empathy and acknowledge the actions of other people [EF69]. The possibility of enabling human-computer interfaces to recognize and make use of the information conferred by facial expressions has hence gained significant research interest over the last few years. This has given rise to a number of automated systems that recognize facial expressions in images or video.

The starting point of this thesis is the observation that existing automated facial expression analysis systems are concerned with either one of two problems. The first is the problem of facial action analysis or identifying the basic units of facial activity such as an eyebrow raise. This is essentially a perceptual task, which is a necessary but insufficient component of mind-reading. The second problem is that of the recognition of basic emotions—happy, sad, angry, disgusted, afraid, surprised.

Recognizing the basic emotions from facial expressions is of limited utility in understanding the user's cognitive state of mind and intentions. These cognitive states and intentions are more relevant and frequent in an HCI context where the user is typically performing some task. For example, even though mild forms of fear of computers are common among inexperienced users or learners (fear is a basic emotion), they are less frequent in an HCI context than cognitive mental states like thinking or concentrating. The result is that the application of automated facial expression analysis to human-computer interaction is limited to primitive scenarios where the system responds with simple positive or negative reactions depending on which basic emotion the user is in.

The range of mental states that people express and identify extends beyond the classic basic emotions, to include a range of affective and cognitive mental states which are collectively referred to as complex mental states. These states encompass many affective and cognitive states of mind, such as *agreeing*, *confused*, *disagreeing*, *interested* and *thinking*. To the best of my knowledge, the analysis of complex mental states has received almost no attention compared to the automated recognition of basic emotions. The aims of this dissertation are twofold:

1. Advance the nascent ability of machines to infer complex mental states from a video stream of facial expressions of people. Recognizing complex mental states widens the scope of applications in which automated facial expressions analysis can be integrated, since these mental states are indicators of the user's goals and intentions. Hence, the automated inference of complex mental states serves as an important step towards building socially and emotionally intelligent machines that improve task performance and goal achievement.

2. Develop a working prototype of an automated mental state inference system that is specifically designed for intelligent HCI. To be useful in an HCI context this system needs to execute in real time, require no user intervention in segmentation or other forms of manual pre-processing, should be user independent, and should support natural rigid head motion.

The automated inference of complex mental states from observed behaviour in the face involves a number of challenges. Mental state inference involves a great deal of uncertainty since a person's mental state is hidden from the observer, and can only be inferred indirectly by analyzing the behaviour of that person. In addition, the automated analysis of the face in video is an open machine-vision problem that is the concern of many research groups around the world. These challenges are further accentuated by the lack of knowledge about the facial expressions of complex mental states; there is no "code-book" that describes how facial expressions are mapped into corresponding mental states.

1.3 Dissertation overview

This dissertation describes a computational model of mind-reading as a novel framework for machine perception and social-emotional intelligence. The design of the computational model of mind-reading uses the results of several studies that I have undertaken to investigate the facial signals and dynamics of complex mental states. The model is a multi-level probabilistic graphical network that represents face-based events in a raw video stream at different levels of spatial and temporal abstraction. The implementation of the model combines top-down predictions of mental state models with bottom-up vision-based processing of the face. The implementation is validated against a number of desirable properties for intelligent machine interaction.

The remainder of this dissertation describes the application of state-of-the-art computer vision and machine learning methods in the design, implementation and validation of a computational model of mind-reading to infer complex mental states in real time:

- *Chapter 2*: Background

  The research described in this dissertation draws inspiration from several disciplines. This chapter presents the different theories on how humans perceive and interpret mental and emotional states of others, surveys the research done on how to enable computers to mimic some of these functions, and highlights the shortcomings of this work in dealing with mental states other than the basic emotions.

- *Chapter 3*: Facial Expressions of Complex Mental States

  The two corpora of complex mental states used throughout this dissertation are introduced here. The chapter then reports the results of two studies that investigate the facial expressions and dynamics of complex mental states. It concludes with the implications of the findings for the design of a computational model of mind-reading.

- *Chapter 4*: Framework for Mental State Recognition

  This chapter begins by describing the computational model of mind-reading and presents an overview of the automated mind-reading system that is based on it. The chapter concludes by discussing the advantages and disadvantages of using this approach.

- *Chapter 5*: Extraction of Head and Facial Actions

  This is the first of three chapters that discuss, one level at a time, the implementation of the automated mind-reading system. This chapter presents the extraction of basic spatial and motion characteristics of the face.

- *Chapter 6*: Recognition of Head and Facial Displays

  In this chapter, consecutive actions are analysed spatio-temporally using Hidden Markov Models to recognize high-level head and facial displays. The experimental evaluation demonstrates the reliable, real time recognition of displays sampled from a wide range of mental states.

- *Chapter 7*: Inference of Complex Mental States

  This chapter describes the inference of complex mental states from head and facial displays in video. The Dynamic Bayesian Networks of complex mental states, the mechanisms for parameter and structure learning and the inference framework are presented. A post-hoc analysis of the resulting models yields an insight into the relevance of head and facial signals in discriminating complex mental states.

- *Chapter 8*: Experimental Evaluation

  The performance of the automated mind-reading system is evaluated for six groups of complex mental states in terms of accuracy, generalization and speed. These groups are *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*.

- *Chapter 9*: Conclusion

  This chapter highlights the work presented here and its major contributions. It concludes the dissertation with directions for future research.

1.4 Publications

Some of the results in this dissertation have appeared in the following publications:

1. Rana el Kaliouby and Peter Robinson. *Real-Time Vision for HCI*, chapter Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures, pages 181–200. Spring-Verlag, 2005.

2. Rana el Kaliouby and Peter Robinson. The Emotional Hearing Aid: An Assistive Tool for Children with Asperger Syndrome. *Universal Access in the Information Society* 4(2), 2005.

3. Rana el Kaliouby and Peter Robinson. Mind Reading Machines: Automated Inference of Cognitive Mental States from Video. In *Proceedings of The IEEE International Conference on Systems, Man and Cybernetics*, 2004.

4. Rana el Kaliouby and Peter Robinson. *Designing a More Inclusive World*, chapter The Emotional Hearing Aid: An Assistive Tool for Children with Asperger Syndrome, pages 163–172. London: Springer-Verlag, 2004.

5. Rana el Kaliouby and Peter Robinson. FAIM: Integrating Automated Facial Affect Analysis in Instant Messaging. In *Proceedings of ACM International Conference on Intelligent User Interfaces (IUI)*, pages 244-246, 2004.

6. Rana el Kaliouby and Peter Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *IEEE Workshop on Real-Time Vision for Human-Computer Interaction at the CVPR Conference*, 2004. Won the Publication of the Year Award by the Cambridge Computer Lab Ring.

7. Rana el Kaliouby and Peter Robinson. Real-Time Head Gesture Recognition in Affective Interfaces. In *Proceedings of the 9th IFIP International Conference on Human-Computer Interaction (INTERACT)*, pages 950–953, 2003.

8. Rana el Kaliouby, Peter Robinson and Simeon Keates. Temporal Context and the Recognition of Emotion from Facial Expression. In *Proceedings of the 10th International Conference on Human-Computer Interaction (HCII): Human-Computer Interaction, Theory and Practice*, volume 2, pages 631-635. Lawrence Erlbaum Associates, 2003.

9. Rana el Kaliouby and Peter Robinson. The Emotional Hearing Aid: An Assistive Tool for Autism. In *Proceedings of the 10th International Conference on Human-Computer Interaction (HCII): Universal Access in HCI*, volume 4, pages 68-72. Lawrence Erlbaum Associates, 2003.

Chapter 2

Background

The research described in this dissertation draws inspiration from several disciplines. In this chapter, I present the different theories on how humans perceive and interpret mental and emotional states of others, and survey the previous research done on how to enable computers to mimic some of these functions. I first present this literature for the basic emotions, which have received most of the attention to date, and then explore the research that considers other mental states. I conclude the chapter by highlighting the shortcomings of automated facial analysis systems in dealing with a wide range of mental states.

2.1 Mind-reading

Mind-reading, theory of mind, or mentalizing, refers to the set of representational abilities that allow one to make inferences about others' mental states [PW78, Wel90, BC92, O'C98, Whi91]. In colloquial English, mind-reading is the act of "discerning, or appearing to discern, the thoughts of another person" or "guessing or knowing by intuition what somebody is thinking" [Hor87]. Following the works of Baron-Cohen *et al.* [Bar94, Bar95] and others such as Realo *et al.* [RAN+03], this dissertation uses mind-reading in a scientific sense to denote the set of abilities that allow a person to infer others' mental states from nonverbal cues and observed behaviour.

From the point of view of an observer who mind-reads, the input is an array of observations, such as visual, auditory and even tactile stimuli, as well as context cues; the output is a set of mental states that are attributed to others [Cru98]. The types of mental states that people exhibit and attribute to each other include emotions, cognitive states, intentions, beliefs, desires and focus of attention. Mind-reading is often referred to in the developmental psychology literature as a specific faculty, separable from more general cognitive abilities such as general intelligence and executive function.

Interest in the functions and mechanisms of this ability has become a central and compelling question for cognitive scientists in recent years. Since Premack and Woodruff [PW78] and Dennett [Den89] first stimulated the interest of cognitive scientists in mind-reading, numerous tasks, methods and theories have accumulated in the literature on this topic. Developmental and experimental studies, as in Goldman and Sirpada [GS05], investigate theoretical models of how people mind-read. Other studies examine the neural basis of mind-reading using brain-imaging technologies like

functional Magnetic Resonance Imaging. Examples include the studies by Baron-cohen *et al.* [BRW+99], Fletcher *et al.* [FHF+95] and Gallagher *et al.* [GHB+00]. The findings from the two classes of studies contribute to our understanding of how the cognitive skills that enable high-level social cognition are organized in the human brain, and the role they play in everyday functioning. These findings also form the basis for the computational model of mind-reading in Chapter 4.

2.1.1 The functions of mind-reading

While subtle and somewhat elusive, mind-reading is fundamental to the social functions we take for granted. It is an important component of a broader set of abilities referred to as social intelligence [BRW+99]. Through mind-reading we are able to make sense of other people's behaviour and predict their future actions [Den89, MF91, Mel04]. It also allows us to communicate effectively with other people [BBLM86, HBCH99, Tro89]. In addition, mind-reading has been described as a cognitive component of empathy [BWL+02, HPB98]. A good empathizer can immediately sense when an emotional change has occurred in someone, what the causes of this change might be, and what might make this person feel better. Mind-reading is also a powerful tool in persuasion and negotiation [HBCH99]: by realizing that people's thoughts and beliefs are shaped by the information to which they are exposed, it is possible to persuade them to change what they know or how they think.

Mind-reading is also a key component of other processes such as perception, learning, attention, memory and decision-making [BDD00, Bec04, Dam94, Ise00, LeD96, MH05, SM90]. In their studies, LeDoux [LeD96], Damasio [Dam94] and Adolphs [Ado02] uncover the parts of the brain that are responsible for higher order processing of emotion. These studies and others, like that by Purves [PAF+01], have shown that these brain areas are interconnected to other brain structures that are involved in the selection and initiation of future behaviour. These findings emphasize the interplay of emotion and cognition, and have led to a new understanding of the human brain, in which it is no longer considered as a purely cognitive information processing system; instead it is seen as a system in which affective and cognitive functions are inextricably integrated with one another [Sch00]. The implications for user-modelling in human-computer interaction (HCI) are clear: an accurate model of the user would have to incorporate the affective as well as the cognitive processes that drive the user's reasoning and actions.

2.1.2 Mind-blindness

The ability to mind-read has been shown to develop during childhood. From as early as 18–30 months, children refer to a range of mental states including emotions, desires, beliefs, thoughts, dreams and pretence [HBCH99]. By the age of five, most children can attribute many mental states to other people, and use them to predict—even manipulate—these people's actions [PWS02, Wel90, Whi91].

The lack of, or impairment in, the ability to reason about mental states is referred to as mind-blindness [Bar95, Bar01, Fli00]. Mind-blindness is thought to be the primary inhibitor of social and emotional intelligence in people with Autism Spectrum Disorders (ASD) [BLF85, Bar95, Fri89b, Fri89a, Bar01]. Autism is a spectrum of neurodevelopmental conditions that is characterized by abnormalities in a triad of domains: social functioning, communication and repetitive behaviour/obsessive interests [Ass94]. The implications of being mind-blind include the inability to gauge the interest of others in conversation [FHF+95], withdrawal from social contact [HMC01], insensitivity to social cues, indifference to others' opinions and inappropriate nonverbal communication [HBCH99].

2.1.3 Mind-reading mechanisms

Mind-reading involves two components that originate in different parts of the brain and develop at distinctive ages. These components may be impaired selectively across different populations of people [BWH+01, Sab04, TFS00].

The first component encompasses the **social-perceptual** component of mind-reading [TFS00], which involves detecting or decoding others' mental states based on immediately available, observable information. For example, one could attribute the mental state *confused* to a person given their facial expressions and/or tone of voice. As its name implies, this component involves perceptual, or bottom-up processing of facial or other stimuli. It also involves cognitive abilities, or top-down processing of abstract models that depict how people's behaviour generally map to corresponding mental states [CO00, PA03].

The second component is the **social-cognitive** component of mind-reading. This involves reasoning about mental states with the goal of explaining or predicting a person's actions. Examples include distinguishing jokes from lies, or predicting peoples' behaviour on the basis of false beliefs. False belief tasks test a person's understanding that other people's thoughts can be different from one another and from reality, and are the prototypical measure of the social-cognitive aspect of mind-reading [Fri01].

It is important to note that both the social-perceptual and the social-cognitive components of mind-reading are inherently uncertain—we are never 100% sure of a person's mental state. A person's mental state (John is thinking), and its content (what John is thinking about) are not directly available to an observer; instead they are inferred from observable behaviour and contextual information with varying degrees of certainty. Moreover, people often have expressions that reflect emotions or mental states that are different than their true feelings or thoughts. The discrepancy between expressed and true feelings, such as in lying and deception, can sometimes be identified from fleeting, subtle micro-expressions [Ekm92b]. The problem of identifying deception from facial expressions is beyond the scope of this dissertation.

2.2 Reading the mind in the face

The research presented throughout this dissertation can be described as an attempt to automate the first of the two components of mind-reading. To gain a better understanding of this component, I review the various tasks that have been devised to tap into social-perceptual understanding in people. These tasks test people's ability to recognize intentional, emotional or other person-related information such as personality traits, given perceptual stimuli like vocal expression [RBCW02], actions [PWS02] and facial expressions. Out of these cues, facial expressions have received the most attention.

Facial expressions are an important channel of nonverbal communication. They communicate a wide range of mental states, such as those in Figure 2.1, which shows the promotional material of actress Florence Lawrence (1890-1938). Besides conveying emotions, facial expressions act as social signals that enhance conversations and regulate turn-taking [EF78]. A face is comprised of permanent facial features that we perceive as components of the face such as the mouth, eyes and eyebrows, and transient features such as wrinkles and furrows. Facial muscles drive the motion and appearance of permanent facial features and produce transient wrinkles and furrows that we perceive as facial expressions. Head orientation, head gestures and eye gaze have also been acknowledged as significant cues in social-perceptual understanding. For example, Haidt

Figure 2.1: Facial expressions communicate a wide range of mental states. The top four poses are labelled (from left to right) as: Piety, Concentration, Hilarity and Coquetry. The bottom four (from left to right): Horror, Mirth, Determination and Sadness. The picture shows the promotional material for actress Florence Lawrence who worked for the Biograph Company and was known as “The Biograph Girl”. For a while in 1909 she was making two films each week. This picture was taken from *A Pictorial History of the Silent Screen* [Blu53].

et al. [HK99] show that gaze aversion, a controlled smile and a head turn are signals of embarrassment. Langton *et al.* [LWB00] emphasize the role of head orientation and eye gaze as an indicator of the focus of attention.

2.2.1 The basic emotions

In 1971, Ekman and Friesan [EF71] demonstrated the universal recognition of six emotions from the face in a number of cultures. The six emotions— happiness, sadness, anger, fear, surprise and disgust—became known as the basic emotions. The facial expressions associated with these basic emotions have almost dominated the study of facial expressions for the past forty years. These six emotions are viewed as dedicated neural circuits that facilitate adaptive responses to the opportunities and threats faced by a creature. For example, the feeling of fear leads to flight, while that of anger leads to fight [Ekm92a, Ekm94, TC90]. In addition to their universality, these emotions are also recognized by very young normally developing children [Wal82, Wel90]. Since it was first proposed, the theory of basic emotions has become one of the most prevalent theories of emotion. Advocates of this "emotions view" of the face share the belief in the centrality of emotions in explaining facial movements.

Despite its prevalence, the theory of basic emotions is quite controversial. To begin with, the criteria of what constitutes a basic emotion is under debate[1]. In Ekman's model, the universality of an emotion decides whether or not it is considered basic. The problem with this criterion is that some emotions are not considered basic by virtue that their universality has never been tested. For instance, Baron-Cohen *et al.* [BRF+96] show that emotions that are conventionally thought of as complex such as guilt and scheming are recognized cross-culturally in cultures as diverse as Japan, Spain and the UK, implying that they may be eligible to join the list of basic emotions.

Other researchers argue that universality should not be the criterion used to judge whether or not an emotion is basic. For example, Fridja *et al.* [FKtS89] propose that action readiness should be used to decide if an emotion is basic. The corresponding set of basic emotions consists of desire, happiness, interest, surprise, wonder and sorrow. There is also controversy over the mere notion of a set of basic emotions. This is because a taxonomy of emotions that categorizes only a few emotions as basic, and all other emotions as complex, masks the differences in meaning and in form among these non-basic emotions [BGW+04]. A number of emotion taxonomies have been presented that include many categories as opposed to just two. For example, Johnson-Laird and Oatley [JLO89] list seven classes of emotions, which are causative, relational, generic, basic, complex, goal, and caused emotions.

2.2.2 Performance

The ability with which people recognize the facial expressions of basic emotions has been the focus of a large number of psychological studies, complemented more recently by a wealth of studies in neuroscience that use brain mapping and neuroimaging technologies. These studies often report the **accuracy** with which a human population correctly recognize emotions in facial stimuli; they predict how the findings **generalize** to real world stimuli; and they analyse the **speed** with which the brain processes the stimuli.

[1] As a matter of fact, there is no singular or even preferred definition of emotion. A discussion of contending definitions of what is an emotion is outside the scope of this dissertation. Interested readers are referred to Cabanac [Cab02] for a recent survey of alternative definitions.

Figure 2.2: Facial expressions of the six basic emotions—happy, sad, afraid, angry, surprised and disgusted—plus a neutral face. From Ekman and Friesen's Pictures of Facial Affect (POFA) [EF76].

Together, these three criteria provide a good measure of people's performance on facial emotion recognition tasks.

To measure the accuracy of recognition, a group of subjects are shown a series of facial stimuli; for each stimulus they are asked to pick an emotional label from a number of choices that best matches what the face is conveying. Open-ended procedures in which subjects are allowed to respond freely have also been used. However, they are less common than forced-choice procedures. The facial stimuli are either static or dynamic. Static stimuli consist of still images as in the Pictures of Facial Affect (POFA) [EF76]. POFA contains 110 black and white photographs of 14 actors portraying prototypic expressions that are reliably classified as happy, sad, afraid, angry, surprised, or disgusted by naive observers. The expressions made by one of those actors are shown in Figure 2.2.

Dynamic stimuli, as shown in Figure 2.3, typically consist of computer-generated morphs between two emotion faces that start with a neutral face and end with a peak emotion. Note how the motion is very controlled with no head motion at all. The recognition of basic emotions improves with dynamic stimuli because the information inherent in facial motion is encoded. Examples of studies that have used dynamic stimuli include Edwards [Edw98], Kamachi *et al.* [KBM$^+$01], Kilts *et al.* [KEG$^+$03], Krumhuber and Kappas [KK03] and Sato *et al.* [SKY$^+$04].

Figure 2.3: Example of dynamic stimuli. From Kamachi *et al.* [KBM$^+$01].

The majority of children and adults recognize the facial expressions of basic emotions highly accurately from stimuli of the whole face [EF86] and in stimuli consisting of blends of two or more emotions [YRC$^+$97]. Even individuals diagnosed on the high end of the Autism Spectrum (High Functioning Autism or Apserger Syndrome) were, in some cases, able to correctly identify the basic emotions [BWJ97, BWH$^+$01, Gro04].

While numerous studies have been repeated over the years with similar findings, it is unclear whether they generalize to natural stimuli from real world scenarios. Generalization is an issue because the stimulus used in lab settings differs substantially from those people get exposed to in their everyday lives, and it is not necessarily easier to recognize. To start with, the images or sequences used in almost all the studies undertaken so far are posed, that is, the "actors" are asked to act the various emotions. Posed

expressions have different configurations and dynamics compared to spontaneous ones and are mediated by separate motor pathways [KEG+03, SCT03]. The stimuli used in laboratory experiments, unlike that in the real world, are stripped out of context in order to control the parameters in these studies. That is, only the face is made available to the subjects of the study; other nonverbal cues and the social context in which the facial expressions were made are not provided. Context has been shown to assist interpretation of facial expressions [Wal91], so stripping the stimuli out of context adds complexity to the recognition task.

Each stimulus is also carefully segmented so that there is a one-to-one mapping between an emotion and the corresponding facial expression. A smile, for example, is always used as the face of happiness. This oversimplifies the recognition task since in real-world scenarios a smile may be common to different kinds of emotions, such as pride, as well as to other psychological processes that are not distinctly emotions, such as in greeting someone [Fri97, FGG97]. It is also possible that different people express the feeling of happiness in different ways, or not at all [FKtS89]. The combination of these factors make it hard to predict from existing studies alone, whether people perform better or worse in real world scenarios. Further studies are needed to quantify people's ability to recognize emotions in natural stimuli.

The speed with which facial emotions are processed has been investigated using event-related potentials (ERPs) [BT03, KSFVM01]. ERPs are a series of positive and negative voltage deflections in the ongoing electrical activity of the brain measured from scalp electrodes. The ERPs are obtained by time-locking the recording of brain activity to the onset of events such as viewing facial stimuli. ERPs have shown differences in timing, amplitude and topographic layout of activation to different facial expressions. The components that occur prior to 100 ms reflect information processing in the early sensory pathway arising from rapid global processing of the facial stimuli. The components that occur after 100 ms are referred to as long-latency ERP components. Of these components, those that occur between 100 to 250 ms represent late sensory and early perceptual processes. Components that occur after 250 ms are thought to reflect higher level cognitive processes such as memory and language. The fact that humans process facial stimuli at multiple levels of abstraction provides the basis for implementing the computational model of mind-reading in Chapter 4 as a multi-level one.

2.2.3 The Facial Action Coding System

While the meaning communicated by some facial signal can be interpreted in different ways, its description should be indisputable. In 1978, Ekman and Friesan [EF78] published the Facial Action Coding System (FACS) as an attempt to unify how facial movements are described. The system describes 44 unique action units (AUs) to correspond to each independent motion of the face. For example, a lip corner pull is AU12. It also includes several categories of head and eye positions and movements. Tables 2.1 and 2.2 illustrate the AUs coded in FACS and their descriptions. AUs can occur either singly or in combination. When AUs occur in combination they may be additive, which means that the resulting action does not change the appearance of the constituent AUs. They can also be non-additive, in which case the appearance differs from that of the constituent actions. Even though Ekman and Friesen [EFA80, EF86] proposed that specific combinations of AUs represent prototypic expressions of emotion, emotion-labels are not part of FACS; this coding system is purely descriptive and does not encode how facial actions map into emotional or mental states.

Table 2.1: The upper action units (AUs) in the Facial Action Coding System (FACS) [EF78]. From Tian *et al.* [TKC01].

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.

FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way. It can also be used to discriminate between subtle differences in facial motion. For these reasons, it has become the leading method in measuring facial behaviour. FACS-coding requires extensive training and is a labour intensive task. It takes almost 100 hours of training to become a certified coder, and between one to three hours of coding for every minute of video [DBH+99].

2.3 Automated facial expression recognition

The need for an objective and inexpensive facial action coding system has been a key motivation for the development of automated Facial Expression Recognition (FER) systems. Typically, an FER system is presented with images or video clips of a person's face or facial movement, and is required to produce a description of that person's facial expression and corresponding emotional state. Advances in real time machine vision and machine learning methods, have produced a surge of interest in developing FER systems for emotionally intelligent human-computer interfaces. This research is supported by evidence that suggests that people regard computers as social agents with whom "face-to-interface" interaction may be the most natural [RN96, NM00].

2.3.1 Intelligent human-computer interaction

For an FER system to be used effectively in an HCI context a number of conditions need to be satisfied. Each of these conditions, presented below, has implications on the choice of methods used to implement such a system. Table 2.3 summarizes these characteristics:

Table 2.2: The lower action units (AUs) in (FACS) [EF78]. From Tian *et al.* [TKC01].

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			
				

1. **Range of mental states**: an FER system should reliably classify a wide range of mental states that are communicated in the face. The more mental states a system is able to recognize, the better its social intelligence skills, and the wider the scope of HCI applications that can be integrated with this technology.

2. **Fully automated**: it is imperative that all stages of an FER system not require manual intervention. While many face processing methods exist, only a subset work without the need for manual pre-processing.

3. **Real time**: most user interfaces require real time responses from the computer for feedback to the user, to execute commands immediately, or both [TK04]. In an HCI context, a system is real time if, from the user's perspective, it responds to an event without a noticeable delay [MW93].

4. **Rigid head motion**: ideally, an FER system would perform robust facial expression analysis in the presence of rigid head motion as well as other challenging conditions such as non-frontal poses.

5. **Continuous and asynchronous expressions**: facial expressions are continuous and overlap each other. They may represent mental states that are also overlapping. Carefully segmented sequences, that correspond to a single facial expression, and that start with a neutral face, peak, then end with a neutral face, rarely occur naturally. While overlapping, facial actions often co-occur asynchronously. The FER system of choice needs to be able to process these overlapping, asynchronous expressions.

6. **User-independent**: an FER system should yield reliable results when presented with new users without the need for retraining or calibration. New users are those whose faces are not included in any of the examples used to train the system; no retraining or calibration means that one can acquire the system, install it and it would be immediately ready for use.

7. **Deals with occlusion of the face**: occlusion occurs when a portion of the face image is missing such as when hand gestures occlude parts of the face or facial features, or when the face is momentarily lost. Humans are able to read facial expressions even when part of the face is occluded.

8. **Neutral expression not required**: Many FER systems rely on the availability of a neutral expression to compare a facial expression to. Most systems also assume that an expression starts and ends with a neutral face. As it is often the case that a neutral face or neutral first frame is not available, developing a system that does not have that limitation is important.

9. **Talking heads**: Finally, natural communication involves speech in addition to nonverbal cues. Typically the input is a continuous video sequence where the person is talking and expressing his/her mental state too [Pet05].

In addition to satisfying the above requirements, there are several other factors that add to the challenge of automatically analyzing facial expressions. First, there is the complexity inherent in processing faces in video. This entails the automatic detection and alignment of features in faces that vary in age, ethnicity, gender, facial hair and occluding objects such as glasses. Furthermore, faces appear disparate because of pose and lighting changes. Finally, cultural and inter-personal variation in emotional expression adds to the complexity of the problem.

Table 2.3: List of desired functions in automated facial analysis systems.

#	Criteria
1	Supports many mental states
2	Fully-automated
3	Real time
4	Deals with rigid head motion
5	Continuous and asynchronous expressions
6	User-independent
7	Deals with occlusion of the face
8	Neutral expression not required
9	Supports talking

2.3.2 Automated recognition of basic emotions

The majority of systems that automate the analysis of facial expression are concerned with the recognition of basic emotions and/or the automated recognition of facial actions. Table 2.4 compares previous FER systems against the functions listed in the previous section. Rows [1–5] describe early systems that were key to shaping the research in this area; for detailed surveys of other early FER systems, the reader is referred to Pantic and Rothkrantz [PR00a]. Rows [6–20] compare recent automated facial analysis systems, many of which have not yet appeared in literature surveys.

Prevalent FER systems are discussed in the light of a real time emotion recognition system for basic emotions that Philipp Michel worked on under my supervision for his Computer Science Tripos Project Dissertation [Mic03], the final-year undergraduate project at the Computer Laboratory, University of Cambridge. The work described in his dissertation was published in Michel and el Kaliouby [MK03b, MK03a]. The findings from this preliminary research have informed my approach to implementing an automated mind-reading system. For comparison purposes, the bottom row of Table 2.4 shows the functions that are supported by the automated mind-reading system that I have developed. The system is described throughout Chapters 4 to 7.

Implementation overview

An FER system typically consists of two components. The first component deals with the extraction of facial features from still images or a video stream. Facial feature extraction is a challenging problem because of intra-class variations that arise from factors such as rigid head motion, differences in physiognomies and changes in recording conditions. In addition, feature extraction methods need to be fully-automated and have to execute in real time. The second component is the classification of a feature vector into one of a possible set of emotion classes. Classifiers assign a class to the feature vector by maximizing inter-class variation and minimizing intra-class differences. Within this general framework of feature extraction and classification, different methods have been applied. A comprehensive survey of these methods can be found in Fasel and Leuttin [FL03].

An FER system can be static or dynamic. Static approaches consider still images for classification. Calder *et al.* [CBM$^+$01] use principal component analysis to analyse facial expressions in the still pictures of the POFA dataset [EF76]. In Fasel [Fas02b] facial expressions are recognized in still images at multiple scales using a combination of feature extractors in a convolutional neural network architecture. The system accounts for in-plane head motion. The system presented by Pantic and Rothkrantz [PR00a] is

Table 2.4: Comparison of automated facial expression recognition systems against desired functions for HCI. Rows [1-5] are key early works; rows [6-20] are recent systems; row [21] is the system described in this dissertation. Criteria [1-9] as listed in Table 2.3. The last column indicates if FACS [EF78] is used to code facial actions. Legend: • fully supported functions; ○ semi-supported functions; x it is unclear whether the function is supported; – function does not apply.

#	References	1	2	3	4	5	6	7	8	9	FACS
1	Black & Yacoob [BY97]				•		•				
2	Cohn *et al.* [CZLK98]						x				•
3	Donato [DBH+99]						x				•
4	Essa & Pentland [EP95]						x				
5	Yacoob & Davis [YD96]				•		x				
6	Bartlett *et al.* [BLFM03]		•				•				
7	Calder *et al.* [CBM+01]			–			x		–		
8	Chandrasiri *et al.* [CNH01]			•	–						
9	Cohen *et al.* [CSC+03a]						•				
10	Cohn *et al.* [Coh04]		•	•	○		•	x	x		•
11	Fasel [Fas02b]			–			○		–		
12	Hoey [Hoe04]		•				•		x		•
13	Hu *et al.* [HCFT04]										
14	Kapoor *et al.* [KQP03]		•	○	•	•	•	○	•		•
15	Littlewort *et al.* [LBF+04a]		•	•			•				
16	Michel & El Kaliouby [MK03b]		•	•			•				
17	Pantic & Rothkrantz [PR00b]		•	–	○		•		–		•
18	Pardas *et al.* [PBL02]						•			○	
19	Tian *et al.* [TBH+03]		•	•	•		•		•		•
20	Zhang and Ji [ZJ03]		•	x		•	•	○			•
21	Automated mind-reading system	•	•	•	○	•	•		○		•

also a static one. Hybrid facial feature detection is performed on a 24-point, 2D face model that is either in a frontal or a profile view. A rule-based system determines the facial actions described by the features, and classifies these actions into weighted emotion labels.

Static systems can be modified for use with video by invoking the classifiers on every frame, or as permitted by the speed of the classifier. Chandrasiri *et al.* [CNH01] use the difference in low global frequency coefficient of a person's facial expression and neutral face to define one of three emotions—happy, sad, surprised. It is then possible to map an unknown frame to one of the three emotions. The principal drawback of this method is that it is user-dependent. The system in Bartlett *et al.* [BLB+03, BLFM03] extracts Gabor wavelets of the face, which are then presented to Support Vector Machines (SVMs) for classification into one of the basic emotions. The SVMs are invoked on every frame of the video stream.

Dynamic systems consider facial motion in the classification of facial expressions. In the simplest case, the change over consecutive frames or the change with respect to a neutral frame is used to determine the underlying expression. In Michel and el Kaliouby [MK03b, MK03a], we use a fully-automated, real time facial feature tracker for feature extraction [Fac02]. For each expression (Figure 2.4), a vector of feature displacements is calculated by taking the Euclidean distance between feature locations in a neutral and a peak frame representative of the expression. This allows characteristic feature motion patterns to be established for each expression as shown in Figure 2.5. SVMs are

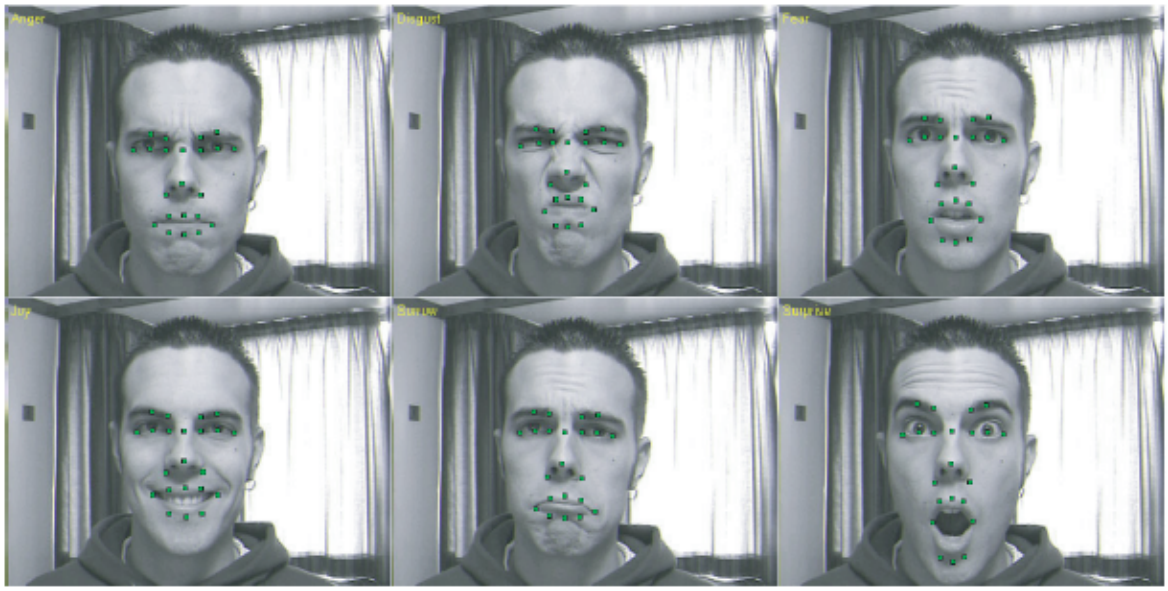


Figure 2.4: Peak frames for each of the six basic emotions, with the features localized. From Michel and el Kaliouby [MK03b].

then used to classify unseen feature displacements in real time, for every frame in the video stream.

Other systems measure how expressions develop over time, and use that information for classification. In Essa and Pentland [EP95], a detailed physics-based dynamic model of the skin and muscles is coupled with optical flow estimates. The facial muscle activation associated with each expression is determined probabilistically. Yacoob and Davis [YD96] detect motion in six predefined and hand initialized rectangular regions on a face and then use FACS rules for the six universal expressions for recognition. Black and Yacoob [BY97] extend this method using local parameterized models of image motion to deal with large-scale head motions. Cohen *et al.* [CSC+03a] use Hidden Markov Models (HMMs) to segment and recognize facial expressions from video sequences. They compare this dynamic classifier to two static ones: a Naive-Bayes and a Gaussian Tree-Augmented Naive Bayes classifier. Their results favour dynamic modelling, although they do point out that it requires more training sequences. In all these cases, sequences start and end with a neutral frame and correspond to a single facial expression.

Whether static or dynamic, the above systems do not consider the transitions between facial expressions. In addition, facial events are considered at a single temporal scale, which corresponds to the duration of a frame or sequence. Hierarchical systems that represent facial expressions at multiple temporal scales, such as the model in Hoey and Little [HL04], work well with limited training and are more robust to variations in the video stream.

FER systems also differ in their assumption about how facial expressions map to emotions. In the simplest case, a one-to-one mapping between a carefully segmented, single, facial expression and its corresponding emotion is assumed. For example, a smile typically corresponds to happiness. Actors who are asked to pose these prototypic facial expression are often required not to move their head at all. Many of the systems in Table 2.4 implicitly make this assumption. For example in Hu *et al.* [HCFT04], the facial expression of each basic emotion is mapped to a low dimensional manifold. The feature space of the manifold is described by a set of facial landmarks that are defined manually on the first frame of a sequence. The system is user-dependent, requiring roughly 1000

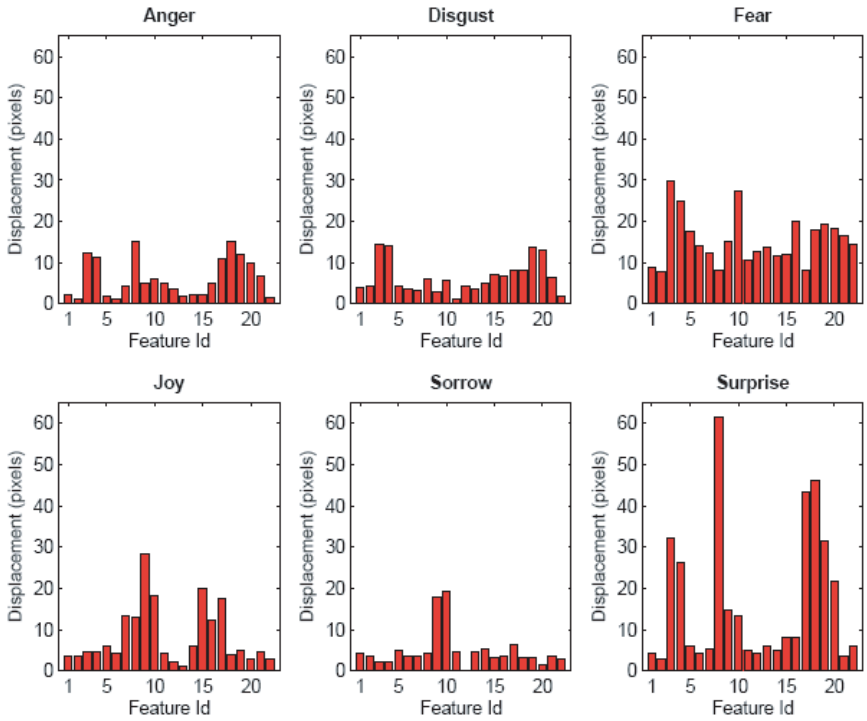


Figure 2.5: Signatures of the six basic emotions. From Michel & el Kaliouby [MK03b].

images to represent the basic expressions for each subject, and at the moment, does not run in real time.

Other systems have relaxed the constraint on head motion, using head pose estimation methods to distinguish between feature motion due to rigid head motion and that due to facial expression. In Tian *et al.* [TBH+03], pose estimation is performed to find the head in frontal or near-frontal views. The facial features are extracted only for those faces in which both eyes and mouth corners are visible. The normalized facial features are input to a neural network classifier to recognize one of the six basic emotions. Bartlett *et al.* [BML+04] investigate 3D head pose estimation to be combined with Gabor-wavelet representation.

The abstraction that there is a one-to-one mapping between facial expressions and emotions does not account for the fact that facial expressions are continuous in nature, occur asynchronously and may vary in duration. In Zhang and Ji [ZJ03], an image sequence may include multiple expressions and two consecutive expressions need not be separated by a neutral state. They use Dynamic Bayesian Networks (DBNs) to classify action units in an image sequence into one of the basic emotions. Although their work is similar to mine in that they too use DBNs as the choice of classifier, their framework for facial expression representation is fundamentally different. In their work, the AUs that occur at one moment in time and the immediately preceding ones, are used to infer a mental state. In contrast, as described in Chapter 4, facial expressions are represented at multiple temporal scales, so that the temporal transitions between, as well as within, expressions are accounted for.

Finally, a number of researchers are particularly concerned with the recognition of facial actions. The motivation behind their work is that facial actions are the building blocks of facial expressions, and can be used within a hierarchical model to describe higher-level facial events. Donato *et al.* [DBH+99] automatically recognize six single upper face AUs and six lower face AUs in sequences of images. Littlewort *et al.* [LBF+04a] extend this

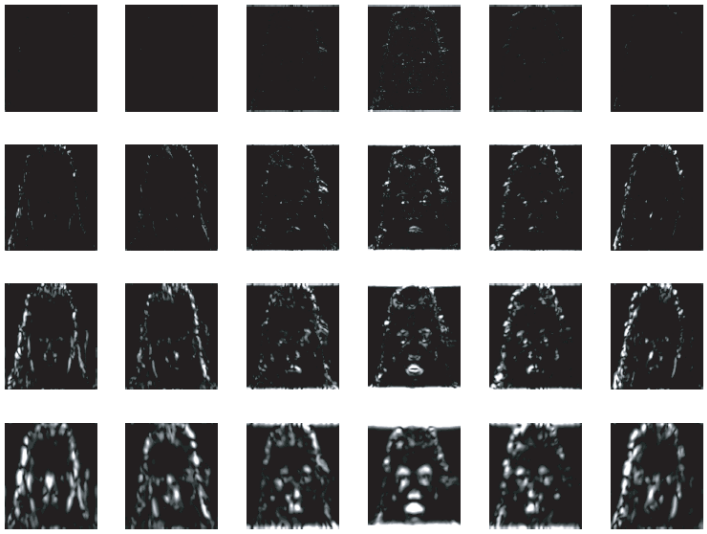


Figure 2.6: A face image after convolution with Gabor filters at different orientations (left-right) and different scales (top-down). From Michel [Mic03].

to support more AUs. The CMU/Pittsburgh group [CZLK98, CZLK99, Coh04, CRA+04, LZCK98, TKC00a, TKC01] have developed a system that recognizes 20 AUs including some non-additive AUs. Kapoor and Picard [KQP03] present a fully automatic method to recognize five upper AUs using an infrared sensitive camera for pupil detection, followed by principal component analysis to recover the shape of the eye and eyebrow regions, and SVMs for classification. Hoey [Hoe04] uses Zernike polynomials in the unsupervised learning of head and facial displays.

Feature extraction

A range of vision-based feature extraction methods exist. Face representation can be 2D or 3D. While more accurate, 3D models such as those used in Essa and Pentland [EP95], Blanz and Vetter [BV99] and Cohen *et al.* [CSC+03a], are computationally intensive and do not execute in real time with current processing power. To maximize correspondence, the facial features are selected interactively on the first frame, and fit to a generic face model for tracking.

Faces can be processed holistically or locally. Holistic methods consider the face as a whole and do not require precise knowledge of specific features. They are, however, sensitive to background clutter. Zhang *et al.* [ZLSA98] use a set of multi-scale, multi-orientation Gabor wavelet coefficients at 34 facial feature points in a two-layer perceptron. The points are extracted manually. A similar representation of faces has been used by Wiskott *et al.* [WFKdM97], where they use a labeled graph based on a Gabor wavelets to represent faces. Figure 2.6 shows a face image after convolution with Gabor filters at different orientations and scales. Lien *et al.* [LZCK98] and Hoey [Hoe04] perform dense optical flow on the whole face.

Local processing methods focus on facial features or areas that change with facial expressions. They are sensitive to subtle expression changes [CZLK98] and are also better-suited to dealing with occlusion of parts of the face [BCL01]. Face models based on feature points [CZLK98, K00, GK01, LZCK98, MK03b, WCVM01], a local geometric representation of the face [Fas02a, PR00b, TKC00a, TKC00b, TKC01], or optical flow of local features [HL03], can be easily adapted to deal with partial occlusion. Even if some model parameters are missing due to occlusion, the parameters not affected by

occlusion are still used for classification. Cohen *et al.* [CSC+03b] and Hoey [Hoe04] have investigated formal methods for learning from missing and unlabelled data.

Feature extraction methods can be classified as appearance-based or feature-based. In appearance-based methods, features are extracted from images without relying on extensive knowledge about the object of interest. Examples include Gabor filters [BLB+03, TKC02, ZJ03], principal component analysis [CYKD01, CBM+01, TP91] and independent component analysis [BHES99, PC95]. Feature-based methods include feature point tracking and geometric face models [LZCK98, PR00a, TKC00a, TKC00b, TKC01]. In its early days, feature-point tracking required dot markers to be attached to the face of subjects using the system. Today, tracking is entirely vision-based and unobtrusive and fully-automated, real time trackers are commercially available. In Michel and el Kaliouby [MK03b], we use `FaceTracker`, part of Nevenvision's commercial facial feature tracking SDK [Fac02]. The tracker is described in more length in Chapter 5. Model-based approaches, such as active contours [BI98], active shape models [CJ92] and active appearance models [CEJ98] have been successfully used for tracking facial deformation. These methods, however, tend to fail in the presence of non-linear image variations such as those caused by rigid head motion and large facial expression changes.

In several comparative studies, Donato *et al.* [DBH+99], Bartlett *et al.* [BHES99] and Zhang *et al.* [ZLSA98] find that appearance-based methods yield better recognition rates than feature-based ones. However, appearance-based methods require extensive, often manual, pre-processing to put the images in correspondence. In addition, in the presence of rigid head motion and a diverse population of subjects, the manual aligning of images for appearance-based feature extraction is laborious, time-consuming and unwieldy, suggesting that appearance-based methods are not well-suited to real time HCI applications. Facial feature tracking, on the other hand, can cope with a large change of appearance and limited out-of plane head motion. It can be complemented with facial component models that are extracted based on motion, shape and colour descriptors of facial components, as well as execute in real time.

Classification

Many classifiers have been applied to the problem of facial expression recognition. Many of these explicitly construct a function that maps feature vectors to corresponding class labels, such as rule-based classifiers, neural networks and SVMs. In a rule-based system, domain knowledge is encapsulated in rules that have the form IF A THEN B, where A is an assertion or group of assertions, and B may be an assertion or action [CDLS99]. The main problem with using rule-based systems for emotion recognition, as in the system described in Pantic and Rokhkrantz [PR00b], is that often the input available may be inadequate or insufficiently reliable to enable a conclusion to be reached, or the rules themselves may not be logically certain. Neural networks consist of input units, hidden units and output units [Bis95]. Connections between units imply that the activity of one unit directly influences the activity of the other, with a specified strength or weight. Neural networks learn from examples by modifying these connection strengths or weights. A number of systems use neural networks for facial action and emotion recognition: Fasel [Fas02a, Fas02b], Pantic and Rothkrantz [PR00b] and Tian *et al.* [TBH+03].

SVMs are based on results from statistical learning theory [Vap95, JDM00]. They perform an implicit embedding of data into a high-dimensional feature space, where linear algebra and geometry may be used to separate data that is only separable with

nonlinear rules in input space. In Michel and el Kaliouby [MK03b, MK03a], SVM-classifiers were implemented using the C++ version of libsvm [CL01]. Bartlett *et al.* [BLB+03, BLFM03] also use SVMs for expression classification. While intuitive and efficient to use, SVMs like rule-based systems and neural networks have no mechanism for representing or making use of dynamic information. There is nothing to link previous classifications with current ones; instead, dynamic information is represented at the feature level. This suggests that dynamic classifiers, such as HMMs, are more appropriate.

Bayesian classifiers take a somewhat different approach to the problem of classification. Often, they do not attempt to learn an explicit decision rule. Instead, learning is reduced to estimating the joint probability distribution of the class and the feature vector describing it [FGG97]. A new instance of a feature vector is classified by computing the conditional probability of each class given that feature vector and returning the class that is most probable. Bayesian classifiers have several advantages. First, they make use of prior knowledge to determine the model and to estimate the prior probabilities. Second, probabilistic methods provide a principled method for dealing with missing data. This is done by averaging over the possible values that data might have taken. Third, when used with decision theory, Bayesian classifiers provide a principled method for combining probability estimates with the utility or cost of different decisions. Bayesian classifiers are discussed in more detail in Chapters 4 and 7.

2.3.3 Facial expression corpora

Table 2.5 summarizes the key differences between four corpora of facial expressions in terms of four groups of factors: 1) general characteristics, 2) the stimuli and recording setup, 3) the actors and 4) the general pose of the actors in the video.

The first two corpora encompass enactments of the six basic emotions. I have already described the first of those—the POFA dataset—in Section 2.2.2. POFA was originally designed for experimental studies. It has, since then, been used to test FER systems, for example in Littlewort *et al.* [LBF+04a]. The second one, the Cohn-Kanade database [KCT00], contains 2105 recordings of posed sequences of 210 adults who are aged 18 to 50 years old, from diverse ethnic origins. The sequences are recorded under controlled conditions of light and head motion, and range between 9-60 frames per sequence at an average duration of 0.67 seconds. Each sequence represents a single facial expression that starts with a neutral frame and ends with a peak facial action. Transitions between expressions are not included. Several systems use the Cohn-Kanade database for training and/or testing. These include Bartlett *et al.* [BLFM03], Cohen *et al.* [CSC+03a], Cohn *et al.* [Coh04], Littlewort *et al.* [LBF+04a], Michel & El Kaliouby [MK03b], Pardas *et al.* [PBL02] and Tian *et al.* [TBH+03].

The third and fourth corpora are the ones that I have used throughout this dissertation. Both cover mental states that extend beyond the basic emotions. A detailed description of each appears in Chapter 3. The third is the video library of the Mind Reading DVD [BGWH04], which is publicly available for a nominal fee. My research pioneers the use of this corpus in automated facial analysis systems. The fourth, the CVPR 2004 corpus was developed specifically to test the generalization ability of the system that I developed throughout this research.

It is important to note that all four databases are posed. In databases of posed expressions, subjects are asked to "act" certain emotions or mental states. For the POFA and the Cohn-Kanade databases, the head motion of the actors is strictly controlled and

Table 2.5: Comparison between Pictures of Facial Affect (POFA) [EF76], the Cohn-Kanade facial expression database [KCT00], the Mind Reading DVD [BGWH04] and the CVPR 2004 corpus. Legend is as follows: • included; x not known; – does not apply. *Described in detail in Chapter [**?**].

	Characteristics	POFA	Cohn-Kanade	Mind Reading*	CVPR2004*
General	Publicly available	•	•	•	
	Complex States			•	•
	FACS-coded	•	•		
	Emotion-labelled	•		•	•
Stimuli	Video stimuli		•	•	•
	Min-Max duration (sec)	–	0.3-2.0	5.0-8.0	0.9-10.9
	Number of videos	–	2105	1742	96
	Dynamic background				•
	Uncontrolled lighting				•
Actors	Number of actors	14	210	30	16
	Male-female ratio (%)	x	31-69	50-50	81.2-18.8
	Diverse ethnic origin		•	•	
	Children & seniors			•	
	Glasses				•
	Moustache				•
Pose	Rigid head motion			•	•
	Non-frontal pose				•
	Non-neutral initial frame			•	•
	Talking				•

a frontal pose is maintained throughout the sequence. In comparison, actors on the Mind Reading DVD and the CVPR 2004 corpus were allowed to exhibit natural head motion. The CVPR 2004 corpus is the most challenging of the four. A few actors wore glasses and some had moustaches. Several were talking throughout the videos and gestured with their hands. Many videos included rigid head motion and non-frontal poses. The background and recording conditions were relaxed in comparison to the other three data-sets, all of which had uniform, static backgrounds. Like the stimuli used in experimental psychology and brain studies, the examples on all four data-sets are context-free. That is, only facial information is available.

2.3.4 Performance

Besides implementation, three factors affect the performance of a vision-based system. The first is the variability of FER systems with respect to the images and/or sequences of facial expressions used to evaluate the systems. The second factor is the degree of variance between the stimuli used to train the system and that used to test it. Systems that are trained and tested on the same corpus typically report better recognition accuracies than those that are tested on a corpus different than the one used in training.

The third is the experimental methodology. Even when training and testing are done on the same corpus, the choice of resampling method has an impact on the result. In addition, some systems are evaluated at each frame, while others are evaluated once for the entire sequence. Another factor that needs to be considered in comparing results is the number of classes that a classifier can choose from and the decision rule used. While these factors make it hard to make a direct comparison between the performance of systems in terms of implementation, an analysis of the results reveals general trends of performance in terms of accuracy, generalization and speed.

Accuracy

Most automated facial analysis systems are tested on a single corpus of images or video, using re-sampling methods such as cross-validation and bootstrapping [Koh95]. In person-dependent tests, examples of an emotional expression of the same subject are used in both the training and the test set, for example, Chandrasiri *et al.* [CNH01]. Recognition accuracies in this case are quite high: between 80-90%. In the more challenging person-independent tests, the sequences of one particular actor are held-out for testing, and the system is trained with the rest of the sequences. Accuracy drops to a range of 55-66%. For instance, Cohen *et al.* [CSC+03a] report an accuracy drop from 82.5% in person-dependent tests to 58.6% in person-independent ones.

Even though in person-independent evaluations, test sequences are previously unseen by a system, there is a database bias introduced in the results by virtue that training and test sequences were subject to the same recording procedures and conditions. A more accurate measure of performance is obtained if systems are tested across corpora.

Generalization

Generalization considers the system's performance when trained on one corpus and tested on previously unseen examples from a different corpus. It is an important predictor of the system's performance in a natural computing environment. The better the generalization ability of the system, the more feasible it is to train the system on some (limited) data-set then deploy it in different interaction scenarios, with many users, without having to re-train or calibrate the system. Evaluating the generalization of a system, however, is an expensive task, especially in vision-based systems where the collection, filtering and labelling of videos is a time-consuming task.

Littlewort *et al.* [LBF+04a] test the accuracy of their system in recognizing the facial expressions of basic emotions when trained on the Cohn-Kanade facial expression database [KCT00] and tested on the Pictures of Facial Affect [EF76] and vice versa. An average of 60% was reported compared with 95% when the system was trained and tested on the same corpus. Tian *et al.* [TBH+03] train their system on the Cohn-Kanade database, and test its ability to recognize smiles on the PETS 2003 evaluation dataset. For a false positive rate of 3.1%, the recognition accuracy of smiles was 91%. This is compared to an accuracy of 98% for smiles when tested on the Cohn-Kanade database [TKC01].

In Michel and el Kaliouby [MK03b], the system's accuracy dropped from 87.5% when trained and tested on the Cohn-Kanade database, to 60.7% when users who were oblivious of the prototypic faces of the basic emotions tested the system. In Pardas *et al.* [PBL02] the system's accuracy dropped from 84% when trained and tested on the Cohn-Kanade database to 64% when sequences containing speech were included. The divergence of results when testing is done between datasets, rather than on a single one, emphasizes the importance of evaluating the generalization of FER systems.

Speed

Speed is a crucial aspect of an FER system, especially if it is intended to be used with interactive interfaces. For systems in general, the time it takes a system to produce its output, starting from the moment all relevant inputs are presented to the system, is called the latency or lag. A system is real time if its latency satisfies constraints imposed by the application. In vision-based systems, a 45 ms visual delay is not noticeable; anything above that, progressively degrades the quality of interaction [MW93].

Significant progress has been made in developing real time vision-based algorithms [KP05]. Feature-based approaches on 2D face models, such as feature-point tracking, are well-suited to real time systems in which motion is inherent and places a strict upper bound on the computational complexity of methods used in order to meet time constraints. 3D face representations, on the other hand, are computationally intensive and do not run in real time with current processing powers. The evaluation stage of the majority of classifiers runs in real time, with algorithms like Kalman filters being extremely fast [KP05]. Exceptions include classifiers that use approximate inference algorithms or implement online learning.

2.4 Beyond the basic emotions

While there is a consensus that the face communicates a wide range of mental states, the prevalence of the theory of basic emotions has meant that almost no attention has gone into studying the facial expression of other emotions or mental states. Some of the reasons why these other states of mind have not been systematically investigated is because they are context-dependent and their definitions are not as clear-cut as the basic emotions.

Our everyday social experiences however, involve much more than just these six emotions, and the ability to recognize them needs to be studied. Rozin and Cohen [RC03] describe a study in which college students were instructed to observe the facial expressions of other students in a university environment and to report the emotion being expressed. The most common facial expressions reported were those of confusion, concentration and worry. Despite their prevalence in everyday interactions, these facial expressions have not been investigated because they do not correspond to generally recognized emotions, leading the authors of the study to call for more studies that explore the facial expressions of mental states that are not typically thought of as emotions.

Simon Baron-Cohen and his group at the Autism Research Centre at the University of Cambridge, have undertaken a series of studies to investigate the facial expressions of mental states other than the basic emotions. The principal objective of these studies is to investigate the differences in emotion processing between a general population of people and those with ASD. Because these differences were not apparent on basic emotion recognition tasks, yet were clearly demonstrated in natural interaction contexts, more challenging tasks were needed.

Baron-Cohen and Cross [BC92] show that normally developing four-year-old children can recognize when someone else is thinking from the direction of that person's gaze. That is, when a person's eyes are directed away from the viewer, to the left or right upper quadrant, and when there is no apparent object to which their gaze is directed, we recognize them as thinking about something. In Baron-Cohen *et al.* [BRF+96], the cross-cultural recognition of paintings and drawings of the face was shown among normal adults and children for mental states such as scheme, revenge, guilt, recognize, threaten, regret and distrust. In two other studies, Baron-Cohen *et al.* [BWJ97, BWH+01] show that a range of mental states, cognitive ones included, can be inferred from the eyes and the face. Figure 2.7 shows several examples of the face stimuli of complex mental states used in Baron-Cohen *et al.* [BWJ97]. The findings of these studies show that many mental states are like virtual print-outs of internal experience, simply waiting to be read by an observer (with a concept of mind).

Figure 2.7: Four examples of the complex mental states face stimuli used in Baron-Cohen *et al.* [BWJ97]: (from left to right) GUILT vs. Arrogant; (b) THOUGHTFUL vs. Arrogant; (c) FLIRTING vs. Happy; (d) ARROGANT vs. Guilt. The correct responses are shown as uppercase letters.

2.5 Limitations of automated facial analysis systems

The field of automated facial expression analysis is a relatively young field of research. While significant progress has been made over the past decade, there is still a long way to go before automated FER systems can be seamlessly integrated with human-computer interfaces. Existing facial analysis systems have the following shortcomings:

1. Focus on the basic emotions.
2. Assume a one-to-one correspondence between an expression and an emotion.
3. Consider the dynamics within a single facial expression only.
4. Account for facial actions but not head gestures or head orientation.

First and foremost, the majority of automated facial expression recognition systems are either concerned with the recognition of basic emotions or with the automated coding of facial actions. There are two recent exceptions. Gu and Ji [GJ04] present a facial event classifier for driver vigilance. The mental states of inattention, yawning as well as the state of falling asleep are represented and classified using DBNs. Kapoor *et al.* [KPI04] devise a probabilistic framework for the recognition of interest and boredom using multiple modalities. These modalities are: facial expressions, which in the current version of the system are coded manually, posture information and the task the person is performing. Their results show that classification using multiple information channels outperforms that of individual modalities.

Applications that are integrated with these FER systems are inherently limited to the basic emotions, and are, as a result, restricted to only a few scenarios [TW01]. For instance, Chandrasiri *et al.* [CNH01] present a system that augments internet chatting with animated 3D facial agents that mimic the facial expressions of the user. Their system only recognizes three of the six basic emotions: happy, sad and surprised. The same application concept would be more powerful if it had knowledge of the user's attention, level of engagement, and cognitive mental states.

Second, naturally occurring facial expressions are continuous; they are not separated by a neutral state and may occur asynchronously. Approaches to automated facial expression analysis that assume a one-to-one correspondence between a carefully segmented

sequence and an emotional state do not generalize well to sequences of natural facial expressions. More importantly, these approaches would not perform well with mental states other than the basic emotions because the same facial expression may mean different emotions and the same emotion may be expressed through different expressions.

Third, while many systems are dynamic, they at most consider the progression of facial motion within a single expression. The transition between one facial expression and another is not considered. In addition, facial events are represented at a single timescale, which is often close to the capture rate of the raw video stream. A more useful model of facial events and emotions requires the abstraction of larger, more meaningful, elements at temporal scales that are progressively greater than the frame rate sampled.

Fourth, facial actions occur as part of coordinated motor routines [CZLK04]. Recently, Cohn *et al.* [CRA+04] studied the temporal relation between facial action, head motion and eye-gaze. They found that considering these additional cues added to the predictive power of their system. For instance, they found that an eyebrow raise was more likely to occur as the head pitched forward. Most existing automated facial analysis systems do not take account of meaningful head gestures and head orientation in their analysis.

In summary, there is an opportunity in developing a system for the automated recognition of mental states beyond the basic emotions. That system would have to account for asynchronous facial cues that occur within a video stream and for the uncertainty in the relationship between facial expressions and underlying mental states. Finally, it would have to consider the transition between facial expressions and would fuse multiple cues including those available from head gestures and head orientation.

2.6 Summary

In this chapter, I have presented the theory of mind-reading, which is at the center of the current thinking on how people read the minds of others from their face. I then presented a survey of the efforts made in automating facial expression recognition of basic emotions. From a technical standpoint these systems, and the machine vision and learning methods they use, are the most relevant to our work in automated mental state recognition. I have critiqued these systems against a list of characteristics that are desirable in an HCI context.

I then explored the limitations of the research that focuses strictly on basic emotions, and presented the shortcomings of existing systems in dealing with a wider range of mental states. Finally, I argued that as of the time of this writing, little attention has been devoted to building a mental state recognition system capable of recognizing complex mental states in an HCI context. The next chapter explores the characteristics and nuances of these complex mental states that will serve as a foundation for the design of the computational model of mind-reading presented in this dissertation.

Chapter 3

Facial Expressions of Complex Mental States

The survey in Chapter 2 has shown that the majority of automated facial analysis systems are concerned with the recognition of the facial expressions of the six basic emotions. This dissertation tackles the emotions that are not part of the basic emotions set. I refer to these states collectively as "complex mental states", rather than complex emotions, to encompass both the affective as well as the cognitive states of the mind.

This chapter explores the characteristics of the facial expressions specific to complex mental states that are central to the design of my automated mind-reading system. The chapter introduces the two corpora of videos that I use extensively throughout this dissertation: the Mind Reading DVD [BGWH04] and the CVPR 2004 corpus. It then presents two experiments that I have undertaken to explore the facial expressions of these mental states. The first is a preliminary study to investigate if there is a single key facial expression within a video of a complex mental state that is a strong discriminator of that state. The results motivate the second study, which explores the temporal dynamics of facial signals in complex mental states. The chapter concludes with a discussion of the implications of the experimental results for the design of an automated mind-reading system.

3.1 Corpora of complex mental states

3.1.1 The Mind Reading DVD

Many people diagnosed with ASD correctly recognize the basic emotions, but often fail to identify the more complex ones, especially when the facial signals are subtle and the boundaries between the emotional states are unclear [BWL+02]. Existing corpora of nonverbal expressions are of limited use to autism therapy, and to this dissertation, since they encompass only the basic emotions.

The Mind Reading DVD is an interactive computer-based guide to emotions (Figure 3.1). It was developed by a team of psychologists led by Professor Simon Baron-Cohen at the Autism Research Centre at the University of Cambridge, working closely with a London multimedia production company. The objective was to develop a resource that would help individuals diagnosed with ASD recognize facial expressions of emotions. It is based on a taxonomy of emotion that covers a wide range of affective and cognitive mental states [BGW+04]. This makes it a valuable resource in developing an automated inference system for computer user interfaces.

Figure 3.1: Mind Reading DVD [BGWH04].

Taxonomy

The taxonomy by Baron-Cohen *et al.* [BGW+04] consists of 412 emotion concepts. The emotion concepts are classified taxonomically into 24 distinct emotion groups, shown in Table 3.1, such that each of the concepts is assigned to one and only one group. The 24 groups in this taxonomy were chosen such that the semantic distinctiveness of the different emotion concepts within each group is preserved. In other words, each group encompasses the fine shades of that mental state. For instance, *brooding*, *calculating* and *fantasizing* are different shades of *thinking*; likewise, *baffled*, *confused* and *puzzled* are different classes within the *unsure* group. The ability to identify different shades of the same group reflects one's empathizing ability [Bar03]. Note that even though the mental state groups are semantically distinctive, this does not mean they can not co-occur. For instance, it is possible to imagine that one is both *thinking* and *confused* at the same time. The co-occurrence of the mental state groups within the taxonomy of Baron-Cohen *et al.* is an interesting and open research question.

Table 3.1: The 24 mental state groups that constitute the taxonomy of Baron-Cohen *et al.* [BGW+04]. Basic emotions are indicated with a •; the groups that are addressed throughout this dissertation are indicated with a ⋆.

afraid •	excited	liked	surprised •
angry •	fond	romantic	thinking ⋆
bored	happy •	sad •	touched
bothered	hurt	sneaky	unfriendly ⋆
disbelieving	interested ⋆	sorry	unsure ⋆
disgusted •	kind	sure ⋆	wanting

Out of the 24 groups, I consider the ones that are not in the basic emotion set, and which, as a result have not been addressed by automated facial analysis systems. Of those 18 groups, I focus on an interesting and challenging subset that is relevant in an HCI context. These groups are marked with a ⋆ in Table 3.1. The five groups constitute the first level of the mental state tree diagram in Figure 3.2. The second level of the tree in Figure 3.2, shown in italics for emphasis, describes the six mental state groups that I particularly focus on throughout the dissertation; the child nodes are the mental state concepts they encompass. For example, the *agreeing* group is derived from the

Table 3.2: The actors on the Mind Reading DVD [BGWH04] characterized by gender, ethnicity, age, accessories and pose.

		C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	%
Gender	Male	•		•		•		•		•		•		•		•		•		•		•		•		•		•		•		50.0
	Female		•		•		•		•		•		•		•		•		•		•		•		•		•		•		•	50.0
Ethnicity	White	•	•	•	•	•		•	•	•		•		•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•		83.3
	Black						•																				•					6.7
	Asian										•		•																		•	10.0
Age	< 18	•	•	•	•	•	•	•	•																							26.7
	$18 - 60$									•	•	•	•	•	•	•	•							•	•	•	•	•	•	•	•	53.3
	> 60																	•	•	•	•	•	•									20.0
Access.	Glasses																															0.0
	Moustache																															0.0
Pose	Frontal	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100.0
	Looking down																															0.0
	Talking																															0.0

sure group in Baron-Cohen *et al.* [BGW+04], and encompasses the following emotion concepts: *assertive*, *committed*, *convinced*, *knowing*, *persuaded* and *sure*.

Note that *concentrating* and *interested* are both derived from the broad group of *interested* in Baron-Cohen *et al.* [BGW+04]. Typically though, in the literature of cognition and emotion, *concentration* is referred to as a separate mental state than that of interest [Ekm79, RC03, Ell03]. I have selected two classes from the same group to test the system's ability to discriminate the finer shades of mental states. Also, in an HCI context, *concentrating* denotes that the user's attention is directed to the machine, while *interested* does not by necessity impose that. Accordingly, *concentrating* encompasses *absorbed*, *concentrating* and *vigilant*, while *interested* covers *asking*, *curious*, *fascinated*, *impressed* and *interested*.

Video library

The emotions library of the DVD provides a corpus of audio and video clips portraying the 412 emotion concepts. Each emotion is captured through six audio and video clips, and is also explained through six stories to give a flavour of the kinds of contexts that give rise to that emotion. The process of labelling the videos involved a panel of 10 judges who were asked 'could this be *the emotion name*?' When 8 out of 10 judges agreed, a statistically significant majority, the video was included. To the best of my knowledge, this corpus is the only available, labelled resource with such a rich collection of mental states and emotions, even though they are posed. In addition, it was not developed with automation in mind, so the videos are much more naturalistic compared with prevalent facial expression databases.

The videos were acted by 30 actors, mostly British, of varying age ranges and ethnic origins[1]. As shown in Table 3.2, there is an equal number of male and female actors, mostly of White origin. The ages of the actors range mostly between 16 and 60. In addition, there are eight actors under 16 and six that are over 60. These two groups are typically not represented in facial expression databases. None of the actors wore any glasses or had a beard or moustache. The resulting 2472 videos were recorded at 30 fps,

[1]The ethnic groups were defined as in the latest UK census, The Focus on Ethnicity and Identity, produced by the UK National Statistics Office, 2004.

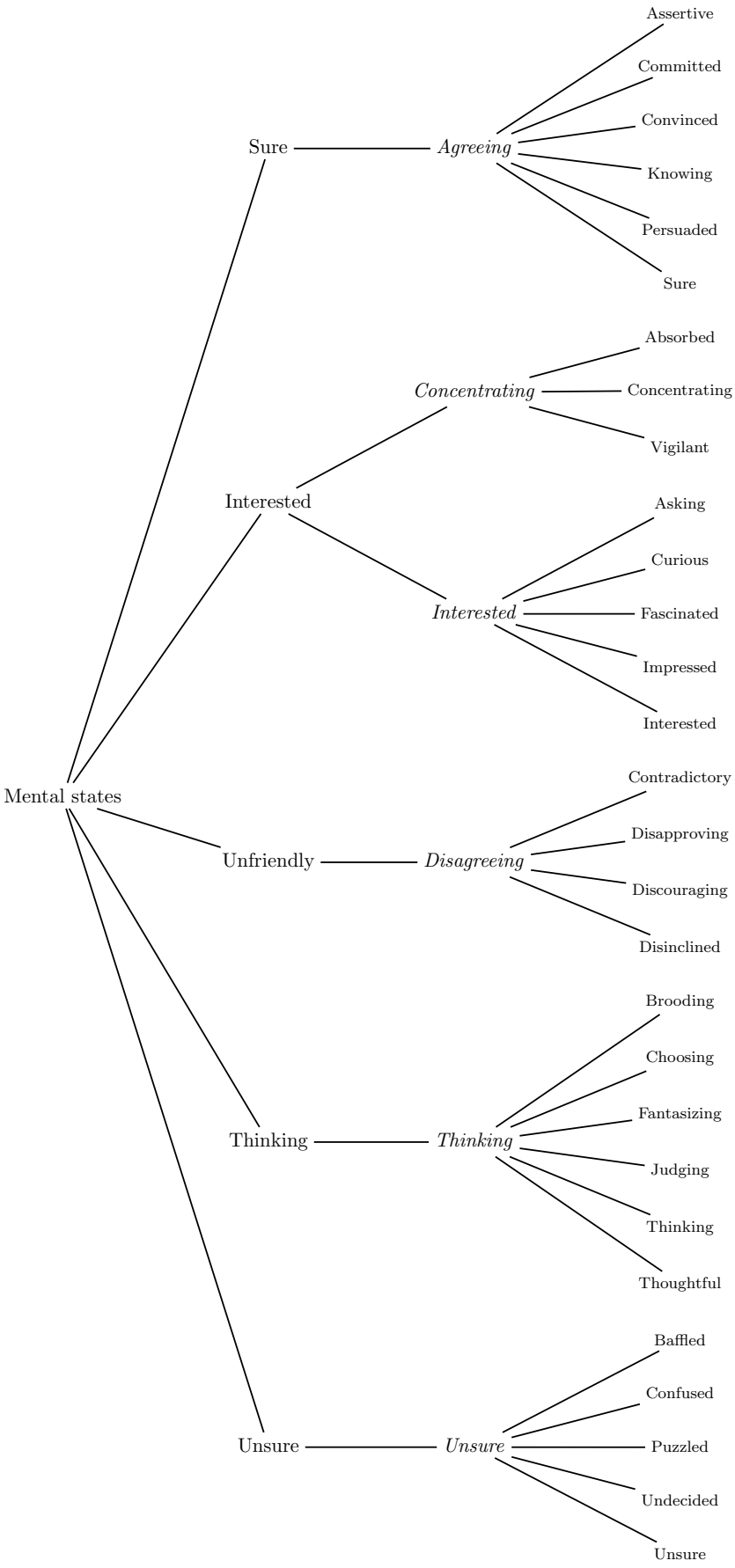


Figure 3.2: Tree diagram of the mental state groups that are addressed throughout the dissertation, shown in *italics*. The groups they belong to in the emotion taxonomy of Baron-Cohen *et al.* [BGW+04] are parent nodes, while the emotion concepts they encompass are child nodes. The complete list of mental states can be found in the emotion taxonomy of Baron-Cohen *et al.* [BGW+04].

and last between five and eight seconds. This is considerably longer than a typical sequence in the Cohn-Kanade database, where the mean duration of a sequence is 0.67 seconds. The resolution is 320x240. All the videos were frontal with a uniform white background. All the actors were looking into the camera and none of them were talking. The videos do not by necessity start, or end for that matter, with a neutral frame.

The instructions given to the actors for an emotion were limited to an example scenario in which that emotion may occur. The actors were *not* given any instructions on *how* to act that emotion. Hence, there is considerable within-class variation between the six videos of each emotion. Moreover, there were no restrictions on the head or body movements of the actors, so the resulting head gestures and facial expressions are naturalistic, even if the mental state is posed. Contrast this to prevalent databases, such as the Cohn-Kanade database [KCT00], in which head motion is strictly controlled, and the facial expressions are prototypic and exaggerated. Finally, while each video is given a single mental state label, it consists of a number of asynchronous head and facial displays. For example, a video of *impressed* includes the following displays throughout the video: a head nod, a head turn, a jaw drop, an eyebrow raise and a smile.

3.1.2 The CVPR 2004 corpus

The videos on the Mind Reading DVD comprise a valuable resource of facial enactments of a wide range of complex mental states. They were, however, taken under controlled recording conditions. An additional corpus of videos representing complex mental states was needed in order to test whether the research presented throughout this dissertation generalizes beyond the controlled videos in the Mind Reading DVD or not. Since the Mind Reading DVD is currently the only available corpus of complex mental states, I decided to construct a second corpus myself: the CVPR 2004 corpus.

Recording setup

During a demonstration of the automated mind-reading system at the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2004) during July 2004, I asked volunteers to act the six mental states shown in italics in Figure 3.2. The volunteers were only given example scenarios to help them perform the mental state (Table 3.3).

Table 3.3: Examples scenarios for the CVPR 2004 corpus.

Mental state	Example scenario
Agreeing	Oh YES! I absolutely agree this is the right way of doing it!
Concentrating	hmmm! ... I have to figure out what this means
Disagreeing	No! No! This is not the right way of doing it
Interested	REALLY? WOW! That's very interesting
Thinking	hmmm! I wonder if this is the right thing to do
Unsure	This is very confusing, not sure what to do

They were *not* given any instructions or guidance on *how* to act a particular mental state. They were, however, asked to maintain a frontal pose as much as possible, but were allowed to move their head freely. The volunteers were also asked to state the specific mental state they will be acting immediately before they started. This too was recorded on the video and was later used to label the videos. While the volunteers were not given instructions about talking, they were allowed to do so if they asked. They were not given any instructions regarding the duration of a recording.

Table 3.4: The 16 volunteers from the CVPR 2004 Conference characterized by gender, ethnicity, age, accessories and pose.

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	%
Gender	Male	•	•		•	•	•	•	•		•		•	•	•	•	•	81.3
	Female			•						•		•						18.7
Ethnicity	White	•	•	•	•	•		•	•	•	•	•	•	•	•	•		87.5
	Black																	0.0
	Asian						•										•	12.5
Age	< 18																	0.0
	$18 - 60$	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100.0
	> 60																	0.0
Access.	Glasses					•	•					•						18.7
	Moustache	•										•						12.5
Pose	Frontal	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	93.8
	Looking down				•					•								12.5
	Talking		•	•					•		•							25.0

The videos were recorded in a relatively uncontrolled environment. The background of this setup was another demonstration booth so people were moving in and out of the scene all the time. The lighting was not controlled; I just relied on the lighting in the conference room at the time. The videos were recorded in a single take, in real time, with one sitting for each subject.

Video library

Table 3.4 summarizes key features of the 16 volunteers. All the volunteers work in a computer-science or engineering discipline. Most were American males of a White ethnic origin. There were two Asians, three Europeans, and only three females. All the volunteers were aged between 16 and 60. Three volunteers had glasses on, and two had moustaches. Only one volunteer had a non-frontal pose, and two were looking down rather than into the camera. Four of the volunteers were talking throughout the videos.

The videos were captured using a standard commercial camcorder and digitized at 30 fps using off-the-shelf video editing software. The resolution is 320x240. The resulting 96 videos, or 12374 frames at 30 fps, have a mean duration of 4.0 seconds or 121.3 frames, standard deviation of 13.66. The longest video is 10.9 seconds long and is labelled as *concentrating*; the shortest is 0.93 seconds and is labelled as *interested*. The *thinking* videos were the longest at an average duration of 5.15 seconds, while the *interested* videos were the shortest at an average duration of 3.92 seconds. Each video was labelled using the actor's subjective label from the audio accompanying the footage.

Since the volunteers were not given any instructions on how to act a mental state, there is considerable within-class variation between the 16 videos of each emotion. Like the Mind Reading DVD, there were no restrictions on the head or body movements of the actors, and each video has a number of asynchronous head and facial displays. The idea is to train the automated mind-reading system once it has been developed on the videos from the Mind Reading DVD and test its generalization power with the CVPR 2004 corpus. To design the framework for automated mind-reading, I explore the characteristics of the facial signals of complex mental states.

Table 3.5: The list of 24 mental state videos used throughout the experiment, and the groups they belong to under the taxonomy in Baron-Cohen *et al.* [BGW+04]. The basic emotions are marked by a •; a ▲ indicates that two videos of the mental state were used in the test.

#	Mental state	Group	#	Mental state	Group
1	Afraid•▲	Afraid	14	Hesitant	Unsure
3	Angry•	Angry	16	Impressed▲	Interested
4	Bored	Bored	18	Interested	Interested
5	Choosing	Thinking	19	Sad•	Sad
6	Comprehending	Thinking	20	Surprised•	Surprised
8	Confused▲	Unsure	21	Sure	Sure
9	Disgusted•▲	Disgusted	22	Thoughtful	Thinking
10	Empathic▲	Kind	22	Tired	Sad
12	Enthusiastic	Excited	24	Undecided	Unsure
13	Happy•	Happy			

3.2 Experiment 1: Facial signals of complex mental states

Classifiers that attempt to infer underlying emotion from a single facial expression are designed with the assumption that there exists a single, peak, facial expression that is representative of the underlying mental state. In this preliminary study, I explore the discriminative power of facial signals for both basic emotions and complex mental states. I used the findings from this study to inform the design of a more powerful experiment that explores the facial dynamics of complex mental states (Section 3.3). In Chapter 7, I describe a more principled approach to identifying highly discriminative facial signals using statistical machine learning.

3.2.1 Objectives

In this preliminary study, I investigate if there is a single key facial expression within a video of a complex mental state that is a strong discriminator of that state. From an observer's point of view, seeing this key facial expression would "give-away" the mental state. Finding whether a key expression within a video is a strong indicator of a mental state would guide both the choice of features and classifier in automated mind-reading.

3.2.2 Experimental design

A number of videos are divided into segments of distinct facial expressions. Participants are then shown the segments in isolation and asked, in a forced-choice procedure, what mental state does the facial expression in the segment represent. A comparison of the percentage of correct answers reported for each of the five segments representing a video should indicate whether any of the segments was particularly discriminative of the mental state.

Stimuli

The stimuli used throughout this experiment were developed using 24 videos from the Mind Reading DVD: 16 videos represented 13 classes of complex mental states, while eight videos represented the six classic basic emotions. Table 3.5 lists the mental states picked for the study and their respective groups. The duration of the videos varied from three to seven seconds (mean=5.3, SD=0.45).

Each video was divided into five separate segments, with a mean duration of 1.06 seconds per segment. The segments were numbered S_1 to S_5 starting from the beginning as

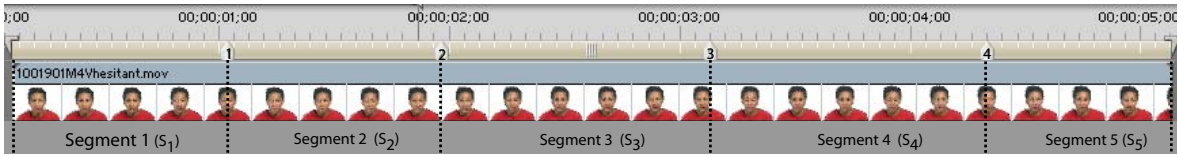


Figure 3.3: Segmenting a video into five segments.

shown in Figure 3.3. To date, automatic facial action segmentation remains a very challenging problem, and no off-the-shelf software that performs this type of segmentation is available. One possible approach, albeit resource intensive, would have been to employ a certified FACS-coder to code the action units and segment the videos accordingly. Instead, I segmented the videos visually according to the following guidelines:

- Consecutive segments should entail essentially different head and/or facial expressions. In other words, a change in pose, head gesture or facial expression signals a new segment.
- If the transition from one expression to another entails a neutral state, then that neutral frame marks the end of a segment and the beginning of a new one, otherwise the frame that lies half-way between the two expressions can be used to divide the segments.
- In the case of asynchronous, but overlapping expressions, then the beginning of the overlapping expression defines the beginning of a new segment.
- If this process results in more than five segments, consecutive segments that have similar facial events are concatenated. Likewise, if segmentation yields less than five segments, longer ones are divided further.

Experimental tasks and procedure

A between-subjects measure was used for the task: each participant was shown one of the five segments for each of the 24 videos. Even though a within-subjects set-up would have minimized the differences in task responses that can be attributed to varying emotion-reading abilities of the participants, it was not an option for this task because of memory-effect. In other words, since the objective is to measure the discriminative ability of a single segment when viewed on its own, seeing more than one segment of the same video would bias and invalidate the results.

To specify which segments of a video each participant got to view, participants were randomly assigned to one of two groups. Participants in the first group viewed either the first or second segment of each video, while those in Group B viewed one of the third, fourth or fifth segments of each video. Note that no presumption was made about which segment out of the five is the most representative one. Indeed, it is entirely possible that the key segment differs across different emotions: S_1 is the most critical to identifying *interest*, S_2 in identifying *boredom*, S_3 in identifying *thinking*, and so on.

A forced-choice procedure was adopted for this experiment. Three foil words were generated for each emotion, for a total of four choices on each question. In picking the foils, I made sure that none was an exact opposite of the target emotion since that might over-simplify the test. Note that the videos on the Mind Reading DVD are already

Table 3.6: The list of target mental state terms for each item (in italic) and their distractors.

1	Sure	Surprised	*Enthusiastic*	Glad
2	*Interested*	Amused	Confused	Affectionate
3	*Confused*	Distraught	Bored	Angry
4	*Tired*	Bored	Distraught	Unimpressed
5	Playful	Sad	Irritated	*Bored*
6	Disappointed	*Sure*	Surprised	Interested
7	*Undecided*	Grumpy	Distraught	Tense
8	Teasing	Hurt	Suspicious	*Choosing*
9	Surprised	Enjoying	*Impressed*	Decided
10	Upset	Cruel	Bored	*Thoughtful*
11	Irritated	*Confused*	Disgusted	Lying
12	*Comprehending*	Calm	Admiring	Enthusiastic
13	Betrayed	Ashamed	*Empathic*	Heartbroken
14	Surprised	Upset	Arrogant	*Impressed*
15	Sneaky	*Hesitant*	Disbelieving	Impatient
16	*Empathic*	Troubled	Guilty	Disappointed
17	Dreamy	*Sad*	Shy	Thinking
18	*Afraid*	Surprised	Cruel	Irritated
19	*Disgusted*	Hurt	Cruel	Angry
20	Happy	Adoring	Excited	*Surprised*
21	Hurt	Frustrated	*Afraid*	Ashamed
22	Terrified	Deceitful	*Frustrated*	Ashamed
23	Afraid	*Disgusted*	Hurt	Deceitful
24	*Happy*	Surprised	Teasing	Adoring

labelled, so it was only a matter of picking the distractors. A complete list of the target mental state terms for each item and their distracters are shown in Table 3.6.

Participants were briefed about the experiment. They were not told anything about its objective. For each question the procedure was as follows:

- Participants were first asked to go through the four emotion words provided for that question, and were encouraged to inquire about any word meanings they were unsure of.
- Participants were shown the video on a standard multimedia player and projection screen, and then asked to circle the emotion word that they thought best matched what the actor in the video was feeling. They were told there was only one correct answer for each question.
- Participants were encouraged to request as many replays as they deemed necessary to properly identify the emotion.

3.2.3 Results

Eight participants, two males and six females, between the ages of 20 and 32 took part in the experiment. Participants were mostly university research members. All participated on a voluntary basis. The test generated 24 trials for each participant for a total of 192 responses. Two independent variables were defined. The segment number indicates which segment of the video was viewed. It has five conditions: S_1 through to S_5.

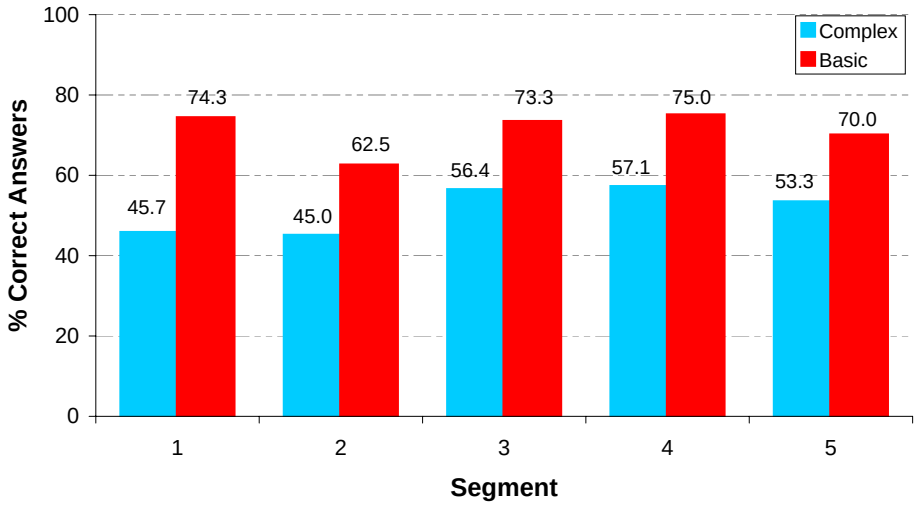


Figure 3.4: Experiment 1: Recognition results for the 192 responses for the five segments of videos of basic and complex mental states. Note the marked difference in recognition rates between the basic emotions (mean=71.0%, SD=0.051) and complex mental states (mean=51.5%, SD=0.058). The recognition accuracies are shown above the bars.

The mental state category has two conditions, either basic or complex. The dependent variable is the accuracy, measured as a percentage of correctly identified mental states.

The recognition results are summarized in Figure 3.4. The figure shows the percentage of correct answers for all the participants when seeing one of the five segments in isolation of the others. In the case of basic emotions, the recognition rate ranges between 62.5% for S_2 and 75.0% for S_4 (mean=71.0%, SD=0.051). For the videos of complex mental states, the recognition rates range between 45.0% for S_2 and 57.14% for S_4 (mean=51.5%, SD=0.058). Note that because there are four options on each question, the probability of responding by chance is 25%. Thus, the results for both the basic emotions and the complex mental states are significantly above chance level.

Statistical analysis of all five tasks

I analysed the results to determine the statistical significance of the difference in recognition results for the five segments in the case of basic emotions and in the case of complex mental states. The Kruskal-Wallis test is the standard non-parametric test for comparing three or more independent samples[2]. The hypotheses for the comparison of three or more groups are:

- The **null** hypothesis H_0: the distribution of the results are the same across the five task conditions. In other words, subjects are equally likely to score correctly when presented with any of the five segments, and any difference between the results of the five tasks is due only to chance.
- The **alternative** hypothesis H_a: the distributions across the tasks are different. The observed difference in recognition is attributed to the effect of the task condition—the discriminative power of the facial displays of that segment—on the ability to discern the mental state.

[2]The statistical tests that I have used throughout this dissertation are detailed in Robson's book on experimental design [Rob94], and were computed using WinStAT [Win04], a statistics add-in for Microsoft Excel.

The test computes the statistic H and the probability p that the distributions are the same across the five task conditions. Since the number of samples per segment is more than five, the test statistic H under H_0 can be approximated with a chi-squared distribution. The value of p is computed as follows: it is the probability under H_0 of getting a value greater than or equal to the observed H. If p is less than $\alpha = 0.05$, then it is statistically unlikely that the difference in recognition between the five tasks was the result of chance. This is evidence to reject H_0 and it is possible to attribute the difference between the results to H_a.

The test showed that the difference in recognition results between the five segments of complex emotions is not statistically significant ($H = 5.95$, $p = 0.2$). A similar result was obtained for the five segments of the basic emotions ($H = 3.6$, $p = 0.46$). Although it would have been possible to increase the power of the results by repeating this study with a larger sample size, I chose to use the findings, even if preliminary, to explore the dynamics of facial of complex mental states. This experiment is described in Section 3.3.

3.2.4 Discussion

The starting point of this study was the observation that the videos of mental state enactments on the Mind Reading DVD involved overlapping, asynchronous facial expressions and purposeful head gestures. The objective of this experiment was to find out if there was a key expression within a video that essentially gives away the mental state. The results show little difference in the contribution of any of the segments of a video to the recognition of the underlying mental state for both the basic emotions and the complex mental states.

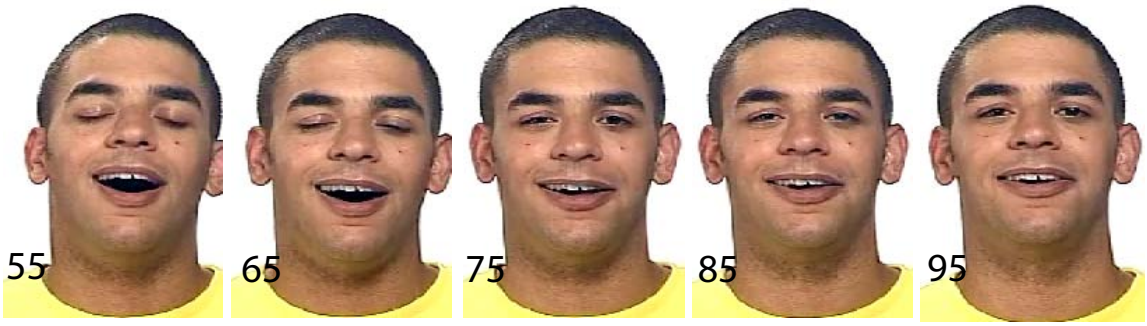


Figure 3.5: A smile is not only an indicator of the *happiness* basic emotion. This smile is extracted from a video labelled as *comprehending* from the Mind Reading DVD.

The generally low accuracy rate attained in recognizing complex mental states in this study is noteworthy. This was surprising given how readily humans identify the same mental states in everyday interactions. This finding emphasizes that even though the individual segments viewed in isolation do help identify a mental state—the recognition rates are above chance level—they are weak identifiers.

To demonstrate how facial expressions can be weak classifiers of mental states, consider the smile sequence in Figure 3.5. Given the limited set of basic emotions to choose from, most people, and most automated facial analysis systems, would easily classify this smile sequence as happy. Given a wider range of complex mental states, the same recognition task becomes much harder. Does this smile sequence signal that the person is *happy*, *excited*, *liked*, *fond* or *romantic*? This smile sequence was, in fact, extracted from a video labelled as *comprehending*.

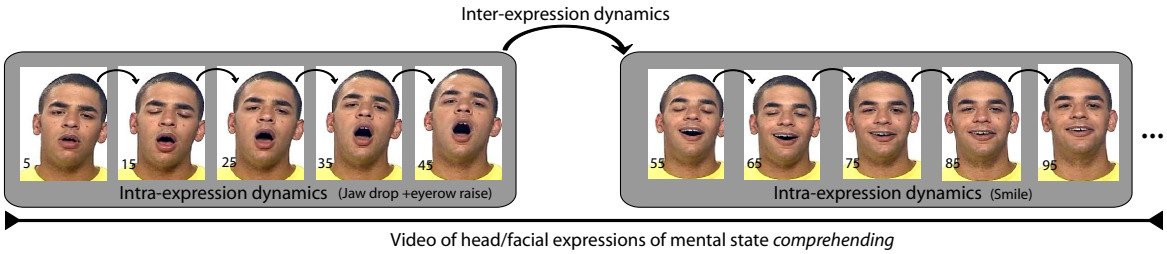


Figure 3.6: An example of intra- and inter-expression dynamics shown for selected frames at 10 frame intervals for the mental state *comprehending*. The intra-expression dynamics of the first segment describe the temporal structure of a jaw drop and eye-brow raise, while the second shows the temporal progression of a smile. The inter-expression dynamics describe the transition from the first to the second segment, or seeing the second segment in the context of the first.

The low recognition rate suggests that people use other cues such as facial dynamics, in addition to facial configuration, to recognize complex mental states. This result provided the motivation for the next experiment, which investigates the dynamics of the facial signals of complex mental states.

3.3 Experiment 2: Facial dynamics of complex mental states

The dynamics of facial expressions describe how facial actions unfold in time. I use the term intra-expression dynamics to denote the temporal structure of facial actions within a single expression. Figure 3.6 shows the intra-expression dynamics of two consecutive segments from a video labelled as *comprehending*. The first segment shows the temporal structure of a jaw drop and eye-brow raise, while the second shows the temporal progression of a smile. Inter-expression dynamics, on the other hand, refers to the temporal relation or the transition in time, between consecutive head gestures and/or facial expressions. In Figure 3.6 the inter-expression dynamics is the transition from the first to the second segment. It describes seeing the smile in the context of the jaw drop/eye-brow raise.

Recall from Chapter 2 that most of the literature on facial dynamics is concerned with intra-expression dynamics. The main problem with the existing studies, and in turn with automated facial analysis systems, is that they only consider single, isolated or pre-segmented facial expression sequences. In natural settings though, facial expressions may occur simultaneously, may overlap asynchronously in time, and often co-occur alongside purposeful head gestures and other modalities. In addition, as shown in Figure 3.6, the transition between two expressions does *not* by necessity involve passing through a neutral state. This is a simplifying assumption that is often made in automated facial analysis systems.

Within these complex patterns of facial signals, the role of inter-expression dynamics in the identification of mental states has not been studied. Cohn *et al.* [Coh04] recently pointed out that very few studies investigate the timing of facial actions in relation to each other, and to other gestures and vocalization. From an engineering point of view, designing an automated mental state classifier that takes inter-expression dynamics into account introduces an additional level of complexity to that of intra-expression dynamics. Incorporating the former when the information it provides is minimal or redundant adds unnecessary complexity to the system. It is unclear given the existing literature whether or not automated systems should be designed to consider expressions in the context of each other.

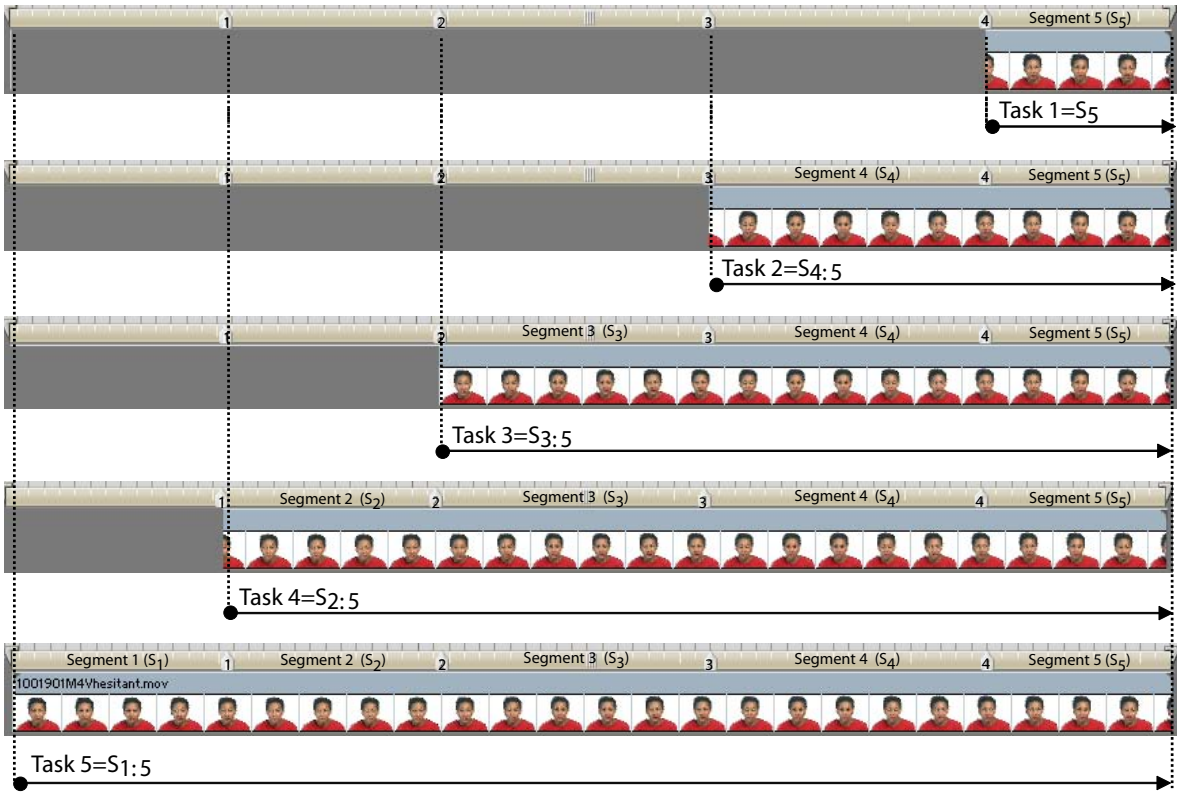


Figure 3.7: Constructing the stimuli for the five tasks in the experiment. The tasks gradually add earlier clips to test their effect on the recognition of the underlying mental states of the clips. The clips are played from start to end, not backward.

3.3.1 Objectives

This study investigates the effect of inter-expression dynamics (if any) on people's ability to recognize complex mental states. Specifically, the goal is to find what relationship exists between the amount of temporal context and recognition accuracy, whether this relationship has critical inflection points and whether it tapers off with additional context becoming irrelevant. I also test the effect of inter-expression dynamics on the recognition accuracy of the classic basic emotions for comparison purposes.

3.3.2 Experimental design

The experiment is structured as follows: 24 videos are divided into segments of distinct facial expressions. Five clips of increasing length are constructed from the segments. Participants are shown the clips in a forced-choice procedure, and are asked to pick the mental state that best describes the clip. A comparison of the percentage of correct answers reported for each of the five clips of a video should indicate whether or not there is any effect of viewing segments in the context of each other.

Stimuli

This experiment consisted of five tasks. The stimuli for the tasks were constructed as shown in Figure 3.7 using the segments of the videos from the previous study. The first task in the experiment comprises only the last segment of the videos S_5. For the second task, the 4th and 5th segments are concatenated, and the resulting clip, which begins at the start of S_4, is the stimulus for that task, and so on. For the 5th and final task, the stimulus is the entire video.

Experimental tasks and procedure

The five tasks test the effect on people's recognition of mental states, of gradually adding earlier segments of a video. Table 3.7 summarizes the span and effect being tested for the five tasks. For instance, the third task tests the effect of seeing $S_{4:5}$ in the context of S_3, which the participants had not seen during the first two tasks. The clips are played forward and not backward[3].

The reason why the tasks were designed to start with the last segment, gradually adding earlier ones, as opposed to starting from the first segment and gradually adding later ones, is to analyse the effect of recognizing a current expression given knowledge about previous ones. This effect is more formally known as the Markov chain effect, and if present warrants the use of dynamic classifiers at the facial expression level in order to represent inter-expression dynamics [Rab89].

Table 3.7: Summary of the five tasks.

Task	Span	Effect tested
1	S_5	baseline
2	$S_{4:5}$	S_5 in the context of S_4
3	$S_{3:5}$	$S_{4:5}$ in the context of S_3
4	$S_{2:5}$	$S_{3:5}$ in the context of S_2
5	$S_{1:5}$	$S_{2:5}$ in the context of S_1

The procedure was the same as that in the previous study: during each of the tasks, participants viewed 24 video clips and were asked to identify the mental state portrayed by the actor in each video. The questions in Table 3.6 were used again for this experiment.

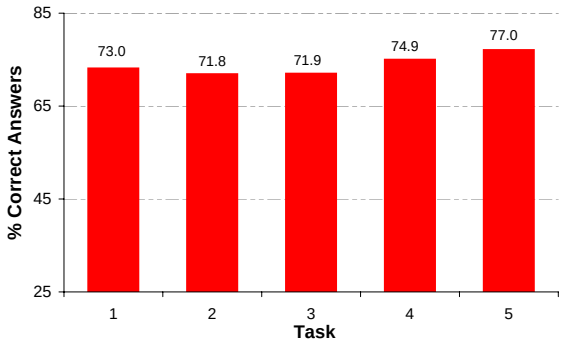
This experiment improved on the previous one in two ways. First, a larger number of participants took part in the experiment. Second, the experiment was designed as a within-subject, repeated-measures study: all participants were asked to carry out all five tasks. Whereas this was not possible with the previous experiment, in this experiment tasks were carried out in increasing order of clip length to prevent any memory effect. The order with which the videos were viewed within each task was randomized. These two factors increase the power of the study to detect real effects, avoiding phantom effects that may be caused by random variation in the emotion-reading abilities of the participants.

3.3.3 Results

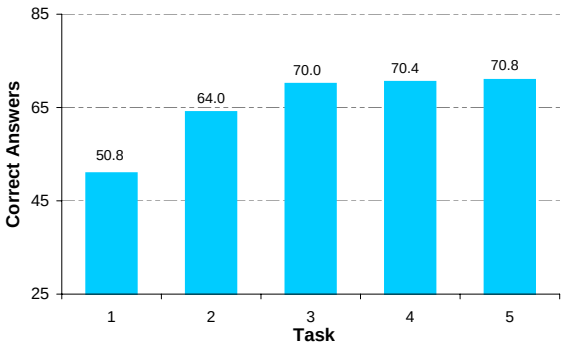
A total of 30 participants (58.0% male, 42.0% female) between the ages of 19 and 65 (mean= 31, SD= 11) took part in the experiment. Participants were either company employees who covered a wide range of occupations or university research members, mostly in Computer Science. All participated on a voluntary basis.

A typical test took 40 minutes on average to complete. There were two category conditions, basic or complex, and five task conditions that depict the span of the video clip viewed by the participants. For example, the span of the clips in Task 2 is $S_{4:5}$. The 30 participants took the five experimental tasks, viewing all 24 videos within each task. This produced 3600 trials in total. The dependent variable is the accuracy of recognition measured in percentage of correctly identified mental states.

[3]Playing the stimuli backward is an equally interesting experiment but tests a different set of hypotheses.



(a) Basic emotions: the recognition results are approximately constant across the tasks.



(b) Complex mental states: an increase in recognition accuracy is observed, which diminishes as more segments are added.

Figure 3.8: Experiment 2: Recognition results for each of the tasks for the case of basic emotions and complex mental states. The recognition accuracies are shown above the bars. Note that the y-axes start at chance responding (25%).

Table 3.8 displays the data—participant versus task condition—generated in the case of complex mental states. A similar matrix was generated in the case of basic emotions. Figure 3.8 summarizes the percentage of correct answers on each task across all the videos and for all the participants. The results show that:

- In the case of basic emotions, shown in Figure 3.8(a), the recognition accuracy is approximately constant; the mean recognition rate is 74.0% (SD=1.9). Note how this result is very similar to that obtained in the previous experiment, where the mean recognition rate for basic emotions recorded there was 71.0%.
- In contrast, for complex mental states (Figure 3.8(b)), there is an increase in recognition accuracy with each task as earlier segments are gradually added. This relationship however is not linear, as more segments are added the observed improvement in recognition tapers off.
- For complex mental states, the result of the first task, the baseline condition, in which participants viewed only the last segment of the video, was 53.6%. Again, this is comparable to the result of the previous experiment, where the mean recognition rate for complex mental states was 51.5%. Quite remarkably, the addition of earlier segments has moved the recognition rate of the complex mental states to a level comparable to that of the basic emotions ($\approx 71.0\%$).

Statistical analysis of all five tasks

I analysed the results to determine the statistical significance of the difference in recognition results for the five tasks in the case of basic emotions, and in the case of complex mental states. In this study, the data is the percentage of correct answers reported for each condition. As explained earlier, the data for this experiment is matched; each row depicts the percentage of correct answers for a single participant over the five task conditions. The distribution of the data was not apparent, suggesting the use of a non-parametric test. Non-parametric tests do not make any assumptions about the distribution of the data.

Table 3.8: Experiment 2: The percentage of correct answers scored by each of the 30 participants on all 16 clips of complex mental states for each of the five task conditions. The tasks are described in Table 3.7.

	1	2	3	4	5	Total
1	31.3	62.5	68.8	68.8	62.5	58.8
2	56.3	56.3	75.0	75.0	81.3	68.8
3	43.8	62.5	75.0	68.8	81.3	66.3
4	43.8	81.3	75.0	87.5	87.5	75.0
5	37.5	31.3	43.8	37.5	37.5	37.5
6	31.3	62.5	37.5	56.3	43.8	46.3
7	43.8	50.0	68.8	81.3	75.0	63.8
8	68.8	87.5	75.0	87.5	75.0	78.8
9	62.5	62.5	68.8	75.0	81.3	70.0
10	31.3	37.5	68.8	43.8	75.0	51.3
11	43.8	43.8	56.3	50.0	50.0	48.8
12	68.8	68.8	87.5	68.8	75.0	73.8
13	62.5	81.3	75.0	62.5	62.5	68.8
14	37.5	50.0	37.5	62.5	50.0	47.5
15	31.3	56.3	50.0	62.5	62.5	52.5
16	62.5	81.3	81.3	68.8	75.0	73.8
17	50.0	50.0	62.5	81.3	75.0	63.8
18	62.5	75.0	81.3	87.5	81.3	77.5
19	50.0	81.3	93.8	75.0	81.3	76.3
20	56.3	81.3	81.3	87.5	75.0	76.3
21	75.0	81.3	81.3	87.5	87.5	82.5
22	25.0	50.0	75.0	56.3	50.0	51.3
23	37.5	37.5	43.8	62.5	62.5	48.8
24	68.8	62.5	87.5	75.0	87.5	76.3
25	37.5	50.0	56.3	50.0	62.5	51.3
26	56.3	75.0	75.0	62.5	68.8	67.5
27	56.3	81.3	87.5	81.3	81.3	77.5
28	75.0	81.3	81.3	87.5	87.5	82.5
29	62.5	62.5	75.0	81.3	62.5	68.8
30	56.3	75.0	75.0	81.3	87.5	75.0
Mean	50.8	64.0	70.0	70.4	70.8	65.2

Table 3.9: The corresponding ranks for all five tasks used as input to the Friedman rank test: one is the lowest rank, and five is the highest.

	1	2	3	4	5
1	1	3	5	5	3
2	2	2	4	4	5
3	1	2	4	3	5
4	1	3	2	5	5
5	4	1	5	4	4
6	1	5	2	4	3
7	1	2	3	5	4
8	1	5	3	5	3
9	2	2	3	4	5
10	1	2	4	3	5
11	2	2	5	4	4
12	3	3	5	3	4
13	3	5	4	3	3
14	2	4	2	5	4
15	1	3	2	5	5
16	1	5	5	2	3
17	2	2	3	5	4
18	1	2	4	5	4
19	1	4	5	2	4
20	1	4	4	5	2
21	1	3	3	5	5
22	1	3	5	4	3
23	2	2	3	5	5
24	2	1	5	3	5
25	1	3	4	3	5
26	1	5	5	2	3
27	1	4	5	4	4
28	1	3	3	5	5
29	3	3	4	5	3
30	1	3	3	4	5
Mean	1.5	3.0	3.8	4.0	4.1

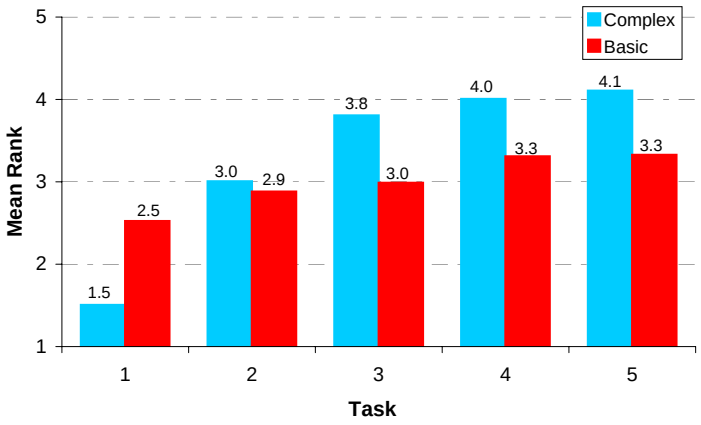


Figure 3.9: Mean ranks of the five tasks (720 responses in each) for basic and complex mental states. A Friedman test on the mean ranks shows a statistically significant difference between the ranks for the complex mental state. This effect was not observed for the basic emotions.

The Friedman rank test is the test to use in comparing three or more matched groups of non-Gaussian scores[4]. Like other non-parametric tests, it uses the ranks of the data. The rank of a number is its position relative to other values in a list. The ranks of the data for complex mental states are shown in Table 3.9. A plot of the mean ranks across the five conditions for the basic emotions and complex mental states is shown in Figure 3.9.

The hypotheses for the comparison of three or more matched groups are:

- The **null** hypothesis H_0: the distributions are the same across the five task conditions. In other words, subjects are equally likely to score correctly when presented with any of the five conditions, and any difference between the results of the five tasks is due only to chance.
- The **alternative** hypothesis H_a: the distributions across the tasks are different. The observed difference in recognition is attributed to the effect of the task condition—the degree of temporal context available—on the ability to discern the mental state.

The test computes the statistic χ^2 and the probability p that the distributions are the same across the five task conditions. Since the number of repeated measures, five in this case, is greater than four and the number of participants is more than 15, the test statistic χ^2 under H_0 can be approximated with a chi-squared distribution. The value of p is computed as follows: it is the probability under H_0 of getting a value greater than or equal to the observed χ^2. If p is less than $\alpha = 0.05$, then it is statistically unlikely that the difference in recognition between the five tasks was the result of chance. This is evidence to reject H_0 and it is possible to attribute the difference between the results to H_a.

In the case of complex mental states, the results were $\chi^2 = 53.1$ and $p = 0.000000001$. Since p is significantly less than α, there is evidence to reject H_0 and to conclude that there is a statistically significant difference between the recognition results of the five tasks. In the case of basic emotions, the results were $\chi^2 = 4.97$ for $p = 0.29$. Since p is greater than α, it is not possible to reject the null hypothesis and the difference between the recognition results was deemed statistically insignificant.

[4]The statistical tests that I have used throughout this dissertation are detailed in Robson's book on experimental design [Rob94], and were computed using WinStAT [Win04], a statistics add-in for Microsoft Excel.

Statistical analysis of consecutive tasks

Analyzing the results in further detail, I then did a pairwise analysis between consecutive tasks to determine if there is any statistical significance between each pair of tasks in the case of basic emotions and in the case of complex mental states. The Signed Wilcoxon Test is the test to use to compare matched pairs of distribution-free scores. The test assumes that there is information in the magnitudes of the differences between paired observations, as well as the signs. The hypotheses for the comparison of each pair of consecutive tasks are:

- The **null** hypothesis H_0: the distributions of the results are the same for the two tasks.
- The **alternative** hypothesis H_a: the distributions across the two tasks are different. The observed difference in recognition is attributed to adding an earlier segment, that is, viewing the two segments in the context of each other.

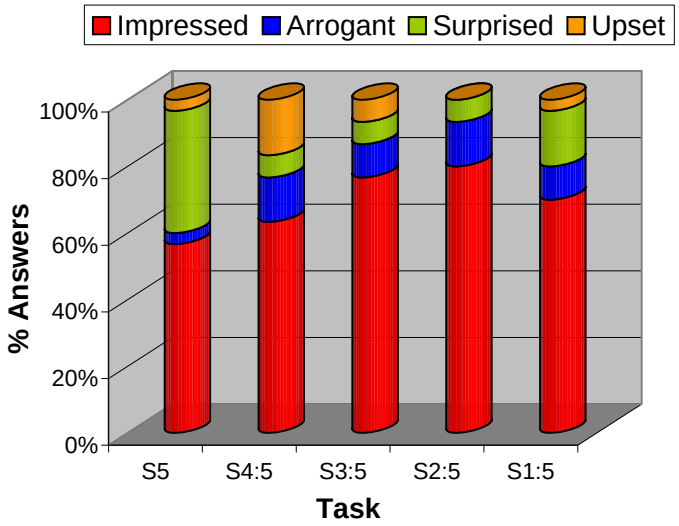
When the number of participants is greater than 10, the Wilcoxon test outputs the Z-value statistic, which under H_0 can be approximated with a normal distribution. The value p denotes the probability under H_0 that the distributions are the same across the two task conditions. If the p-value is less than $\alpha = 0.05$, then it is statistically unlikely that the difference in recognition between the two tasks was the result of chance. This is evidence to reject H_0 and it is possible to attribute the difference between the results to H_a.

To carry out the test, the difference in recognition result is calculated for all the participants between two consecutive tasks, that is, before and after a new segment is added. The differences are then ranked by their absolute value. The ranks of positive differences are summed, so are the ranks of the negative differences. If there is no marked difference in the results of the two tasks, then one would expect the rank sums for positive and negative ranks to be the same. The Z-value is a measure of the difference between the two sets that is based on the rank sums.

At the significance level of $\alpha = 0.05$, a pair-wise analysis of the complex mental state samples shows the following:

- An improvement of 13.2% in accuracy moving from the S_5 condition to $S_{4:5}$, Z-value=-3.77, $p = 0.0001$. Since p is significantly less than α, there is evidence to reject H_0 and it is possible to conclude that there is a statistically significant difference between the recognition results of the two tasks.
- A smaller improvement of 6.0% is observed between the condition $S_{4:5}$ and $S_{3:5}$, Z-value=-2.39 , $p = 0.017$. Since p is less than α, there is evidence to reject H_0, and to conclude that there is a statistically significant difference between the recognition results of $S_{4:5}$ and $S_{3:5}$.
- The percentage improvement between $S_{3:5}$ and $S_{2:5}$ is almost negligible (0.4%) and is not statistically significant (Z-value=-0.04, $p = 0.97$).
- The improvement in moving from $S_{2:5}$ to $S_{1:5}$ is also negligible (0.4%) and statistically insignificant (Z-value=-0.32, $p = 0.75$).

Consistent with the Friedman rank test, a pair-wise analysis of the responses of basic emotions showed negligible improvement between clip spans (mean improvement=2.8%, SD=0.8) and the differences were not statistically significant.



Summary of the five segments

Segment	Head/facial display
5	eye-brow raise
4	head dip
3	head nod, lip purse
2	eye-brow raise
1	jaw drop

Figure 3.10: Distribution of responses for a video labelled as *impressed* and a summary of the underlying head and facial displays in the five segments of the video.

3.3.4 Discussion

The starting point of this study was the observation from the previous experiment that the recognition rate of complex mental states was markedly lower than that of the basic emotions. The objective of this experiment was to investigate the role of dynamics, the inter-expression ones in particular, on the recognition accuracy of complex mental states. To do so, the experiment tested the effect of viewing segments of a video in the context of other segments. The findings can be summarized as follows:

- Viewing segments of a video in the context of each other yields a significant improvement in the recognition accuracy of complex mental states. A similar effect was not observed for the basic emotions.
- Viewing segments of a video in the context of each other boosts the recognition rate of complex mental states to a level comparable to that of the basic emotions.
- The pattern of improvement is not linear with the number of segments: the percentage of improvement tapers off as more segments are added.

In order to explain the results consider the following example from the test. The video is an example of the mental state *impressed*. The three distractors for that item were *surprised*, *upset* and *arrogant*. A summary of the head/facial expressions within each of the five segments and a distribution of responses for this question is shown in Figure 3.10. During the first task, participants were shown only the last segment of the video, which was essentially an eye-brow raise. The responses were split between *impressed* (56.7% of the responses) and *surprised* (36.7%). During the second task the participants were shown the last two segments $S_{4:5}$, a head dip followed by an eye-brow raise. For this task, the recognition rate was 63.3%, an increase of 6.7% from the previous task. Upon adding the head nod in S_3 this jumps to 76.7%, and so on. The highest recognition rate obtained for this video was 80.0%.

There are three possible explanations for the improvement in recognition seen in this example. The first explanation is that one of the segments is representative of that mental state and gives it away. The results of the previous study have already shown

that this is unlikely and does not account for the boost in recognition accuracy. The second explanation suggests that improvement in recognition is due to the fact that, with the longer clips, participants are exposed to facial stimuli for a longer duration. For instance, one could argue that the improvement in accuracy between seeing S_5 and $S_{4:5}$ is because $S_{4:5}$ is the longer clip, and has nothing to do with seeing the expressions in relation to each other. It would have been possible to manipulate the presented clips so that they lasted for the same amount of time. Instead, I chose to allow participants to repeat clips at will until they were satisfied with their answers. This preserves the integrity of the material being presented while eliminating the impact of duration. In addition I refer to a number of previous experiments that have specifically addressed this issue. Experiments by Edwards [Edw98], Kamachi *et al.* [KBM$^+$01] and Lander and Bruce [LB03] show that dynamic information embedded in the motion, not time, is responsible for the differences in expression judgment. Finally, the third and most plausible explanation is that seeing the facial expressions in the context of each other accounts for the pronounced improvement. So in the example, it would be the information embedded in seeing a head dip followed by an eye-brow raise.

3.4 Implications for automated inference of mental states

The studies presented serve as a first step toward developing an automated mind-reading system. The findings of this analysis are summarized in Table 3.10. They are presented in the light of existing approaches to automated facial analysis systems. Their implications for automated inference of complex mental states are also included.

3.4.1 Facial signatures

The theory of basic emotions maintains that each of the basic emotions has a distinct and universal facial expression [EF71, Ekm92a]. For instance, a smile is typically the facial signature of *happiness*. This theory has influenced how automated facial analysis systems of basic emotions are designed. Most FER systems encode a one-to-one mapping between facial signals and emotions: each basic emotion is inferred from a single facial expression that is a strong identifier of it.

As discussed in Section 3.2.4, facial expressions are strong classifiers of emotions only if one is limited to the set of basic emotions. Once the set is expanded to incorporate complex mental states, the problem becomes more complex. Facial expressions will still provide people with cues to discriminate between mental states. However, these expressions are not distinct, and on their own are weak identifiers of these mental states, resulting in a low recognition rate. Unlike basic emotions, the facial signals of each complex mental state encompass multiple asynchronous facial expressions, head gestures and orientation cues, all of which may occur asynchronously. This confirms the literature on overlapping facial actions [GSS$^+$88] and on the role of head orientation in communicating various mental states [Bar94, BRF$^+$96, LWB00, CZLK04]. In terms of an automated mind-reading system, these findings suggest the use of an ensemble of head and facial event classifiers.

3.4.2 Facial dynamics

So far, automated FER systems that are dynamic only deal with intra-expression dynamics. The second study showed that if the same approach to basic emotions is to be followed for complex mental states, accounting only for intra-expression dynamics, the

Table 3.10: The implications of the findings of the two studies for the design of an automated mental state inference system. The findings are considered in the light of existing approaches to automated facial analysis systems.

	Previous approaches	Findings	Design implications
Facial Signals	• 1-1 mapping between facial expression classifiers and basic emotions	• No evidence for the presence of a key segment • Facial expressions are weak classifiers	• Use ensemble of weak classifiers of facial expressions and head displays (or other modalities)
	• Only facial expressions are considered	• Multiple asynchronous facial, head, and eye-gaze cues occur	• Incorporate multiple cues
Facial Dynamics	• Limited to intra-expression dynamics of pre-segmented facial expressions	• Inter-expression dynamics count in complex mental states	• Implement multi-level temporal abstraction (intra- and inter-expression dynamics)
		• Rate of improvement diminishes with the amount of context considered	• Account for two immediately preceding segments (equivalent to two seconds of video on the Mind Reading DVD)

recognition rate would be at least 20.0% less than that of the basic emotions. The recognition results of complex mental states are significantly improved with the incorporation of inter-expression dynamics and are comparable to those of basic emotions. This finding strongly suggests accounting for inter-expression dynamics in an automated mental state inference system, particularly the two immediately preceding segments, which are equivalent to two seconds of video on the Mind Reading DVD. The results can also be extended to using a person's previous mental state in recognizing the current one.

3.5 Summary

This chapter presented a discourse on the facial expressions of complex mental states, both the affective and the cognitive ones. I undertook two studies to investigate the facial signatures and dynamics of various mental states. The first study showed that there were no key segments within a video of a complex mental state that acted as a strong discriminator of that state; rather, the isolated facial expressions were weak classifiers of the complex mental states. The second study showed a marked improvement in the recognition rate of the complex mental states when consecutive head and facial displays were viewed in the context of each other. There was no similar effect reported for basic emotions. With inter-expression dynamics the recognition rate of complex mental states was comparable to that of the basic emotions.

Compared with the facial expressions of basic emotions, the number of studies that address the facial signals of complex mental states is very limited, so there are many open research problems one could explore. To start with, the studies could be repeated with more stimuli that ideally would be sampled from different corpora to determine whether the results reported here are general characteristics of complex mental states, or specific to the Mind Reading DVD. Further experiments are needed to investigate in more detail the signatures and dynamics of a wider range of mental states, with a larger sample of participants. Finally, it is important to note that even with inter-expression dynamics, the results obtained for basic and complex mental states reach a mean upper ceiling of 80%. Further studies are needed to experiment with additional cues beyond those in the face, such as situational context or the inclusion of other modalities.

In terms of an automated mental state inference system, the results suggest that an ensemble of classifiers be used to represent a mental state, and that multi-level temporal abstraction be implemented to account for the intra- and inter-expression dynamics. Beyond the implications for automated mind-reading, the findings presented on the facial expressions of complex mental states can be applied to the design of embodied conversational agents that perceive the mental states of other agents and humans.

In the next chapter, I draw on the results presented here to introduce a computational model of mind-reading. I also describe how that general model is implemented in an automated mind-reading system.

Chapter 4

Framework for Mental State Recognition

In this chapter, I present a computational model of mind-reading to address the problem of mental state recognition. The model uses the theories of how people read the mind in the face to formulate a mapping between high-level hidden mental states and low-level observable facial behaviour. Drawing on findings in the previous chapter on the facial expressions of complex mental states, the model abstracts video input into three levels. Each level conveys face-based events at different levels of spatial and temporal abstraction. I also present an overview of the automated mind-reading system. The system implements the model by combining top-down predictions of mental states with bottom-up vision-based processing of the face to infer complex mental states from video in real time.

4.1 Computational model of mind-reading

The framework that I present for mental state recognition draws on the literature of mind-reading. Mind-reading, described in Chapter 2, is the ability to attribute a mental state to a person from the observed behaviour of that person. The theory of mind-reading describes a coherent framework of how people combine bottom-up perceptual processes with top-down reasoning to map low-level observable behaviour into high-level mental states. In bottom-up processing, facial expressions exhibit rich geometric and dynamic properties sufficient to select a corresponding mental state [EFA80, CBM+01, SC01]. In top-down reasoning, people utilize mental models that map observations of particular facial configurations to mental state labels [PA03, GS05]. The theory of mind-reading also considers the uncertainty inherent in the process of reading other people's minds. This uncertainty results from the stochastic nature of facial behaviour: people with the same mental state may exhibit different facial expressions, with varying intensities and durations.

To simulate the process by which humans mind-read, a computational model of mind-reading would first have to build or "learn" mappings between mental state classes and patterns of facial behaviour as observed in video sequences. These mappings are then used during classification to infer the probability of an incoming video sequence being "caused" by each of the states.

Table 4.1: The key characteristics of the computational model of mind-reading.

Characteristic	Summary
Probabilistic	Bayesian inference framework
Hierarchical	video, actions, displays, mental states
Multi-level temporal abstraction	different degrees of temporal detail
Multi-cue integration	integration of asynchronous sources

To account for the variability in expressing these states, the patterns underlying any generic set of mental states are learned from data using statistical machine learning instead of rule-based methods. In other words, assuming the data exists, the model will learn a mapping from fast continuous video input to any generic set of slowly changing discrete mental states. This is crucial given the lack of expert domain knowledge on the subject. Being data-driven constitutes one aspect of this model of mind-reading. The rest of the characteristics are summarized in Table 4.1 and are presented in detail in the sections that follow.

4.1.1 Probabilistic framework

A probabilistic framework provides a principled approach to combine multiple sources of information and to handle the uncertainty inherent in facial behaviour. Within a Bayesian inference framework, the relationship between a hidden mental state X_i and observed data $\mathbf{D}$ is given by Bayes' rule as follows:

$$P(X_i|\mathbf{D}) = P(X_i)\frac{P(\mathbf{D}|X_i)}{P(\mathbf{D})} \tag{4.1}$$

The prior $P(X_i)$ represents the belief about mental state i before observing data $\mathbf{D}$. The likelihood $P(\mathbf{D}|X_i)$ denotes the probability of data $\mathbf{D}$ being generated from mental state i. The posterior $P(X_i|\mathbf{D})$ represents the updated belief about X_i after observing data $\mathbf{D}$. Symbols that denote a set of events are in bold face, while those that refer to single events are not. $\mathbf{D}$ is in bold face since it denotes the set of facial events extracted from the raw video input.

4.1.2 Hierarchical model

One possible approach to estimate the likelihood $P(\mathbf{D}|X_i)$ is to define a direct mapping from the raw video input to the high level mental states. Monolithic models of this type however, generate a large parameter space, requiring substantial amounts of training data for a particular user. These models also do not generalize well to new users or settings, and typical classification accuracies are not high enough for real time applications [OHG02]. An alternative approach entails representing raw video input at multiple levels. In contrast, hierarchical frameworks work well with limited training, and are more robust to variations in the low-level video input [OHG02]. They also map naturally onto the problem domain since many studies suggest that human behaviour is hierarchically structured such that lower-level units combine to form higher-level ones [Bir70, ZT01].

Ideally, a computational model of mind-reading should work with all users, be robust to variations in facial expressions and generalize well to previously unseen examples of mental states. To satisfy these constraints, I pursued a multi-level representation of the

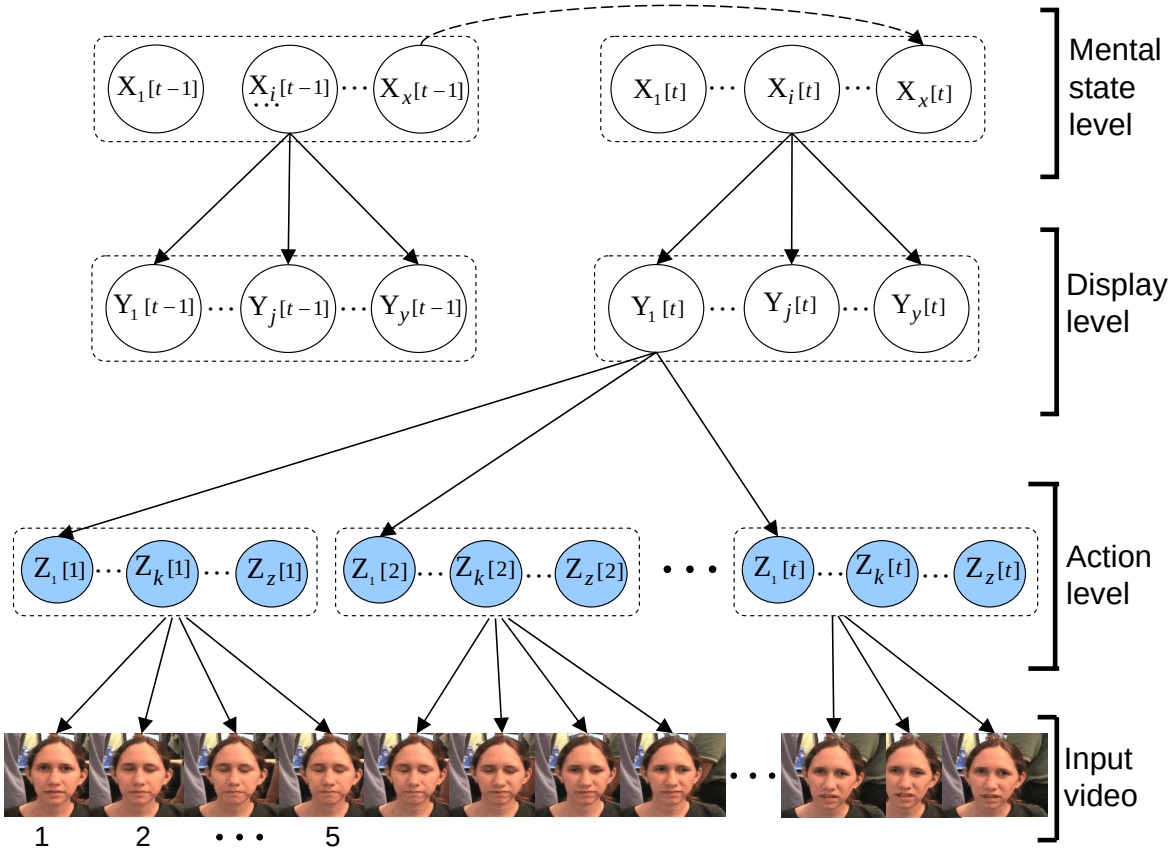


Figure 4.1: Computational model of mind-reading as a multi-level probabilistic graphical model. The observation is a video sequence portraying facial activity shown at the bottom. The three levels are for spatially and temporally abstracting the video. **Action-level**: **Z** $= \{Z_1, \ldots, Z_z\}$ represents z head or facial action events, each extracted from five frames. The actions are based on the FACS AUs [EF78]. **Display-level**: **Y** $= \{Y_1, \ldots, Y_y\}$ represents y head/facial display events, each spanning a sequence of t actions. **Mental state-level**: **X** $= \{X_1, \ldots, X_x\}$ represents x mental state classes each derived from observations of head/facial displays accumulated over a span of two seconds. Solid arrows indicate probabilistic influence, dashed arrows denote temporal influence. Shaded nodes are fully observable. Note: some of the arrows have been left out of the figure for simplification.

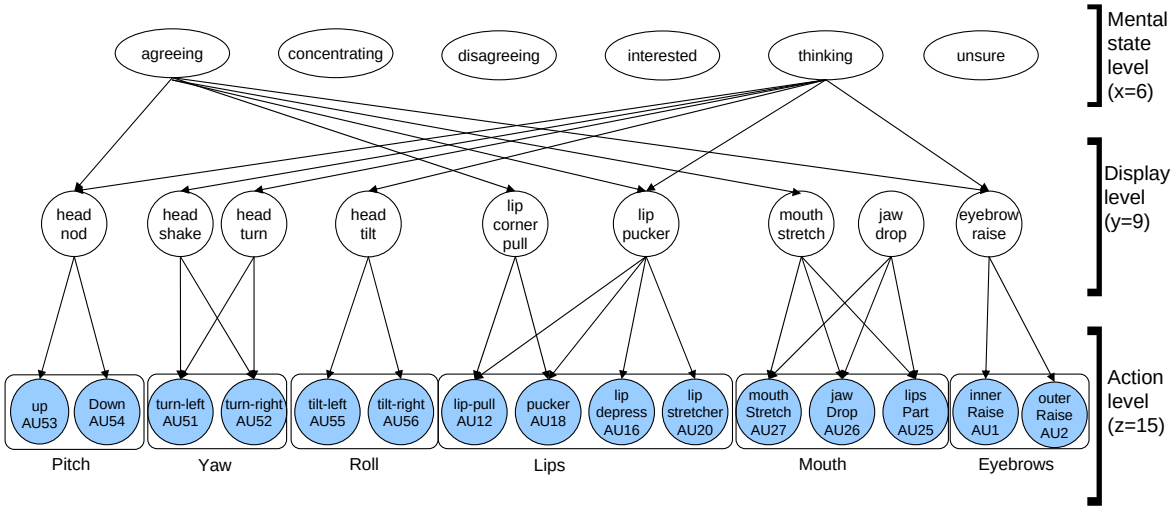


Figure 4.2: A video stream is abstracted spatially into head rotation on three axes and facial components. Each spatial abstraction is described by a number of actions. The actions are in turn abstracted into displays and mental states. The displays present in a model of a mental state are determined by a feature selection mechanism. For clarity, the displays of only two mental states are shown. Figure 7.9 shows the displays present in each of the six mental states.

video input. Figure 4.1 shows the computational model of mind-reading represented as a multi-level probabilistic graphical model (PGM).

PGMs such as Dynamic Bayesian Networks (DBNs), complement probability theory by allowing for the representation of prior knowledge of the causal probability and conditional independence among events in a system in the form of a probabilistic graph. Directed arcs between events capture probabilistic influences. Note that in a hierarchical model each level only influences the one below it, so the parameters are greatly simplified. In my model, video input is abstracted into three levels, each conveying face-based events at different granularities of spatial and temporal abstraction. On each of the levels, more than one event can occur simultaneously. A bold-face symbol represents the entire set of events at any one level, normal-face symbols with subscripts denote specific events. The specific time, or time range, of an event is given in square brackets []:

- **Head and facial actions**: The first level of abstraction models the basic spatial and motion characteristics of the face including the head pose. These are described by z facial events $\mathbf{Z} = \{Z_1, \ldots, Z_z\}$, where each event represents some spatial abstraction, and describes the underlying motion of that abstraction across multiple frames. Figure 4.2 summarizes the $z = 15$ spatial abstractions currently supported by the model. These are head rotation along each of the three rotation axes—pitch, yaw and roll—and lips, mouth and eyebrow facial components. The motions are described by FACS AUs [EF78], which are the building blocks of facial activity. For example, $Z_1[t]$ may represent the head pose along the pitch axis at time t; the possible values of Z_1 are $\{\mathrm{AU53}, \mathrm{AU54}, \textit{null}\}$, which represent a head-up, head-down or neither. Note that the non-additive AUs that are also supported are not shown in Figure 4.2.

- **Head and facial displays**: Head and facial actions are in turn abstracted into $y = 9$ head and facial displays $\mathbf{Y} = \{Y_1, \ldots, Y_y\}$. Displays are communicative facial events such as a head nod, smile or eyebrow flash [Fri92]. I refer to these

communicative conjunctions of facial behaviour as displays rather than expressions to avoid automatic connotation with the facial expressions of basic emotions. Each display is described by an event that is associated with a particular spatial abstraction as in the action level. Like actions, display events can occur simultaneously. The term $P(Y_j[t])$ describes the probability that head/facial display event j has occurred at time t. For example, Y_1 may represent the head nod event. $P(Y_1[t]|Z_1[1:t])$ is the probability that a head nod has occurred at time t given a sequence of head pitch actions. If a head nod event occurs, a sequence of alternating head-up, head-down actions, or some variation of that, such as $\{Z_1[1] = \text{AU53}, Z_1[2] = \text{AU54}, \ldots, Z_1[t] = \text{AU53}\}$ would be observed.

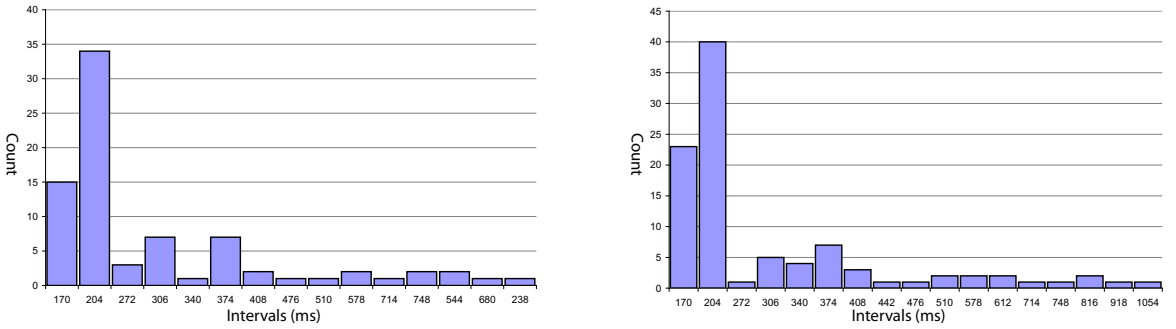
- **Mental states**: Finally, at the topmost level, the model represents $x = 6$ mental state events $\{X_1, \ldots, X_x\}$. For example, X_1 may represent the mental state *agreeing*; $P(X_1[t])$ is the probability that *agreeing* was detected at time t. The probability of a mental state event is conditioned on the most recently observed displays and previous inferences of the mental state: $P(X_i[t]|\mathbf{Y}[1:t], P(X_i[1:t-1]))$. Note that a separate classifier is constructed for each of the x mental states, rather than having only one classifier with x possible values. Having a classifier for each class means that the system can represent mental states that may co-occur.

4.1.3 Multi-level temporal abstraction

Several studies demonstrate how facial expressions are temporally structured in a way that is both perceptible and meaningful to an observer [Edw98, KBM+01, SC01, KRK03]. The experiment in Chapter 3 on the facial dynamics of complex mental states showed the importance of incorporating inter-expression, as well as intra-expression dynamics to boost recognition results. Accordingly, each of the three layers is defined as a dynamic classifier where current events are influenced by previous ones. In addition, each level of the model captures a different degree of temporal detail based on the physical properties of the events at that level. The observation (input) at any one level is a temporal sequence of the output of lower layers. Hence, the higher the level, the larger the time scale, and the higher the level of abstraction.

The automatic estimation of time scales of events from data is a challenging problem. Possible solutions include searching for the most likely time scale [WBC97, WCP00], using the temporal structure of events to automatically segment a sequence [WPG01] or synchronizing with other co-occurring events and context cues [Hoe04]. To determine the time scale of each level, I draw on the characteristics of the videos on the Mind Reading DVD and the results of the experiment in Chapter 3 on the facial dynamics of complex mental states:

- **Head and facial actions**: To determine the time scale of head and facial actions, I timed the temporal intervals of 80 head-up (AU53) and 97 head-down (AU54) motions in head nod gestures. The head nods were sampled from 20 videos by 15 people representing a range of complex mental states such as *convinced*, *encouraging* and *willing*. Figure 4.3 shows that purposeful head-up and head-down movements in a head nod lasted at least 170 ms. This result is similar to the literature on the kinematics of gestures [Bir70, DV01]. Accordingly, I have decided that facial or head actions (later referred to simply as actions) are abstracted as spanning five video frames at 30 fps (approximately 166 ms).



(a) Duration of head-up motion in head nods. (b) Duration of head-down motion in head nods.

Figure 4.3: Distribution of head-up and head-down durations in head nod gestures.

- **Head and facial displays**: Recall from Chapter 3 that the videos from the Mind Reading DVD were segmented into clips containing single head and/or facial displays (later referred to as displays) that were approximately one second long. Accordingly, the time scale of a single display is 30 frames at 30 fps, or six actions. The output progresses one action at a time, that is, every 166 ms.
- **Mental states**: Recall from Chapter 3 that two seconds is the minimum time required for a human to reliably infer a mental state; video segments of less than two seconds result in inaccurate recognition results. I chose to sample these two seconds (60 frames) using a sliding window of 30 frames, sliding it six times, five frames at a time. In display units, the sliding window spans one display and progresses six times one display at a time to constitute a mental state.

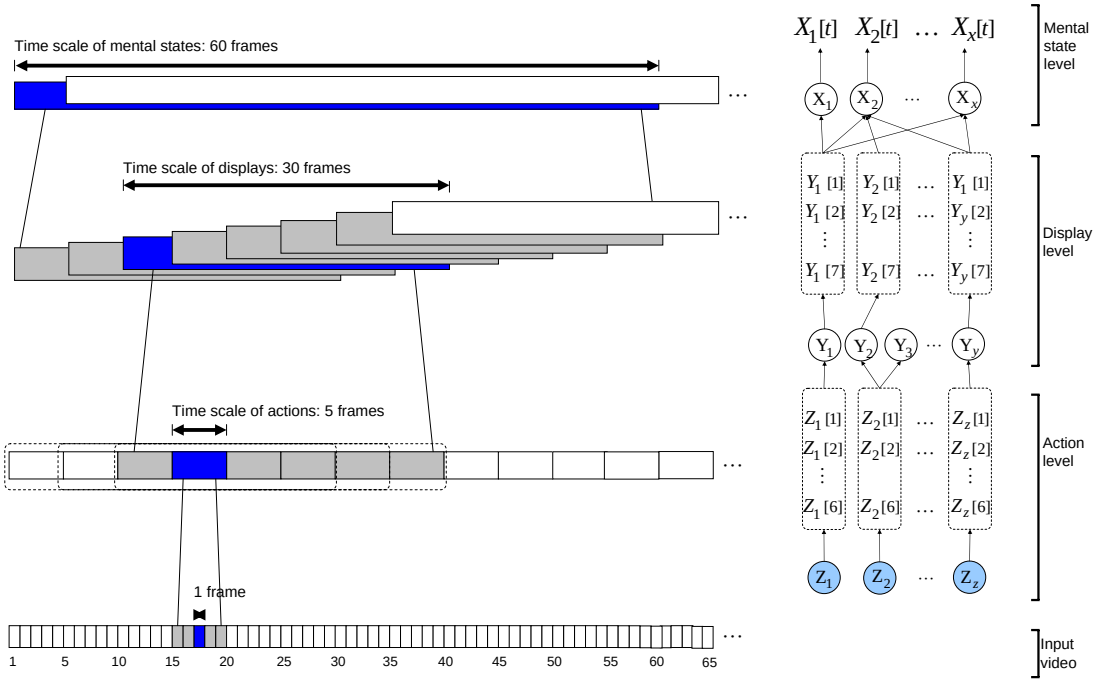
In a sense, each layer performs time compression before passing data upward. For instance, a video of 160 frames at 30 fps is described in terms of 33 instances of head/facial actions, 27 instances of head/facial displays and 22 mental state inferences. The temporal relationship between each of the levels is summarized schematically in Figure 4.4(a). A matrix representation of the output of each of the levels, which constitutes the input to the next level up, is shown in Figure 4.4(b).

4.1.4 Fusion of multiple cues

When mind-reading, people make considerable use of a variety of communication channels and context cues in a complementary and redundant manner to interpret facial expressions [Wal91, BY98, Edw98]. These include information conferred from head and facial behaviour, other modalities and context cues.

To exploit the information available from the different head and facial components, a mental state classifier treats each of the display classifiers as a separate information source. The assumption is that displays are independent of each other. For example, a head nod and a smile are independent given mental state *agreeing*. The resulting mental state model—known as a Naive Bayes model—can be seen as an ensemble of classifiers. This means that each of the display classifiers can be trained independently of each other and combined based on their joint performance on some training data. Thus, mental state i influences y displays as follows:

$$P(X_i|Y_1,\ldots,Y_y) = P(X_i)\frac{P(Y_1,\ldots,Y_y|X_i)}{P(Y_1,\ldots,Y_y)}$$



(a) Time scales at each level of the model. On level L a single event is shown in blue ■. The input to this event is a sequence of events from level $L-1$ (shown in grey ■). At 30 fps, a single action spans 166 ms (five frames), a display spans one second (30 frames), and a mental state spans two seconds (60 frames).

(b) Matrix representation of the output at each level of the model. The direction of the arrows denotes the flow of information.

Figure 4.4: Multi-level temporal abstraction in the computational model of mind-reading.

Assuming conditional independence between displays, the joint distribution over all of the displays is factored into the product: $P(Y_1, \ldots, Y_y|X_i) = \prod_j P(Y_j|X_i)$, where $P(Y_j|X_i)$ is the conditional distribution of display j given its parent, the mental state i. The Naive Bayes classifier yields surprisingly good results in many classification problems even though the independence assumption usually does not reflect the true underlying model generating the data. Domingos and Pazzani [DP97] explore the conditions for the optimality of the Naive Bayes classifier and show that under zero-one loss (misclassification rate) the classifier's region of optimal performance is far greater than that implied by the independence assumption. It is also possible to selectively choose those sources that are most informative to a particular mental state. Because expert domain knowledge on complex mental states is limited, automatic model selection is needed to find the head and facial displays that are the most discriminative of each mental state. Model selection is described in Chapter 7. Within a Bayesian framework, efficient algorithms are available to incorporate new observations in the network.

4.2 Overview of the automated mind-reading system

The automated mind-reading system implements the computational model of mind-reading by combining bottom-up, vision-based processing of the face with top-down, predictions of mental state models to interpret the meaning underlying some head/facial signal. The idea is to use the system as a means for understanding and responding to a user's mental state in a natural computing context. Each level in the hierarchical model maps to a major component in the implementation of the system. The major training and inference stages of the system at each level are summarized in Algorithm 4.1.

Algorithm 4.1 Overview of training and inference stages of the mind-reading system

Extraction of head and facial actions

Offline: Derive motion, shape and colour models of head/facial components

1. Track feature points

2a. Estimate the head pose → Extract head action units

2b. Extract facial features → Extract facial action units

Recognition of head and facial displays

Offline: Train display HMMs

1. Accumulate six actions for each display HMM

2. Input actions to display HMMs

3. Output likelihood for each display

4. Quantize output to binary

Inference of complex mental states

Offline: Learn model parameters and select best model structure

1. Slide the window of display-observations 6 times, 1 display at a time.

2. Input observations and previous mental states to DBN inference engines

3. Output the probability of each mental state

4. Classify output into the most likely mental state (if needed)

4.2.1 Assumptions and characteristics

The automated mind-reading system draws on research in several open machine vision problems. Examples include head pose estimation, feature tracking under varying illumination conditions and 3D pose, and dealing with occlusion. To focus on the principal challenge that this dissertation addresses—the inference of complex mental states in video—I apply state-of-the-art machine vision and machine learning methods with few modifications to implement each level of the system. I also make several simplifying assumptions: 1) the face is assumed to be frontal or near-frontal in the video, although the pose can vary within this view, 2) the lighting conditions are reasonable, and 3) there is no occlusion of the face or facial features.

In addition to complex mental state recognition, the implementation of the system has to satisfy several functions for use in natural computing contexts. These are summarized in Table 4.2. To start with, the system accounts for rigid head-motion while recognizing meaningful head gestures. The implementation of facial action recognition is robust to intra-class variations that arise from rigid head motion. Second, the system is user-independent, yielding reliable results with new users without the need for re-training or calibration. Third, the system requires no manual preprocessing or segmentation of frames in a video. Fourth, the system executes in real time. In the context of mind-reading, real time means that facial events are processed as they occur at minimal latency. The latency is the difference between the instant a frame is captured and the time when the system infers the mental state, and should be comparable to the time it takes humans to mind-read. The system is also unobtrusive in that no special equipment or face-markers are required.

Table 4.2: The key characteristics of the automated mind-reading system.

Characteristic	Summary
Complex states	supports cognitive and affective (complex) mental states
Head-motion	accounts for rigid head motion while recognizing head gestures
User-independent	new users, no calibration
Fully-automated	no pre-processing or segmentation of video needed
Real time	events processed as they occur
Unobtrusive	no special equipments, no face-markers required

4.2.2 Implementation of the three levels

The characteristics listed in Table 4.2 governed my choice of implementation methods at each of the three levels of the automated mind-reading system. The action level is described in Chapter 5. Feature points are first located and tracked across consecutive frames using `FaceTracker` [Fac02], part of Nevenvision's facial feature tracking SDK. The co-ordinates of these points are then used to extract head actions and facial actions. The training phase at this level involves deriving the head pose and facial feature models. This is a time-consuming task because to incorporate a new head/facial action, one would have to design the corresponding dynamic model. An alterative approach is to use Gabor wavelets as in Littlewort *et al.* [LBF+04b]. Though feature independent, Gabor wavelets are less robust to rigid head motion and require extensive, sometimes manual, alignment of frames in a video sequence.

The head and facial displays (Chapter 6) are implemented as Hidden Markov Models (HMMs). These are essentially a quantization of a system's configuration space into a small number of discrete states, together with probabilities for the transitions between these states. The training phase at this level involves defining and training an HMM for each of the supported head and facial displays. HMMs are a good choice of classifiers for representing and classifying displays because they encode the temporal regularity inherent in the head/facial actions that constitute these displays.

The output of the display level forms the observation vector or evidence for the mental state models. At the mental state level (Chapter 7), a DBN is trained for each mental state class from example videos and prior domain knowledge. DBNs are a good choice because they act as an ensemble of classifiers fusing multiple asynchronous information sources over multiple temporal scales.

4.2.3 Inference framework

While training is done off-line during a one-off process, inference is done in real time. A procedural description of how inference is carried out in the automated mind-reading system in shown in Figure 4.5. When a previously unseen video is presented to the system, processing proceeds bottom-up, combining the top-down models that are learned from training data for classification of the facial event. The inference framework is implemented as a sliding window, so head/facial action symbols, head/facial displays and mental state inferences are output approximately every 166 ms. The exception is the first instance of display recognition, which occurs at time one second when six head/facial actions have been accumulated, and the first instance of mental state inference, which occurs at two seconds. The levels execute simultaneously and facial actions are incorporated for mental state inference as soon as they occur. The categorization of mental states early enough after their onset ensures that the resulting knowledge is current and useful to target applications.

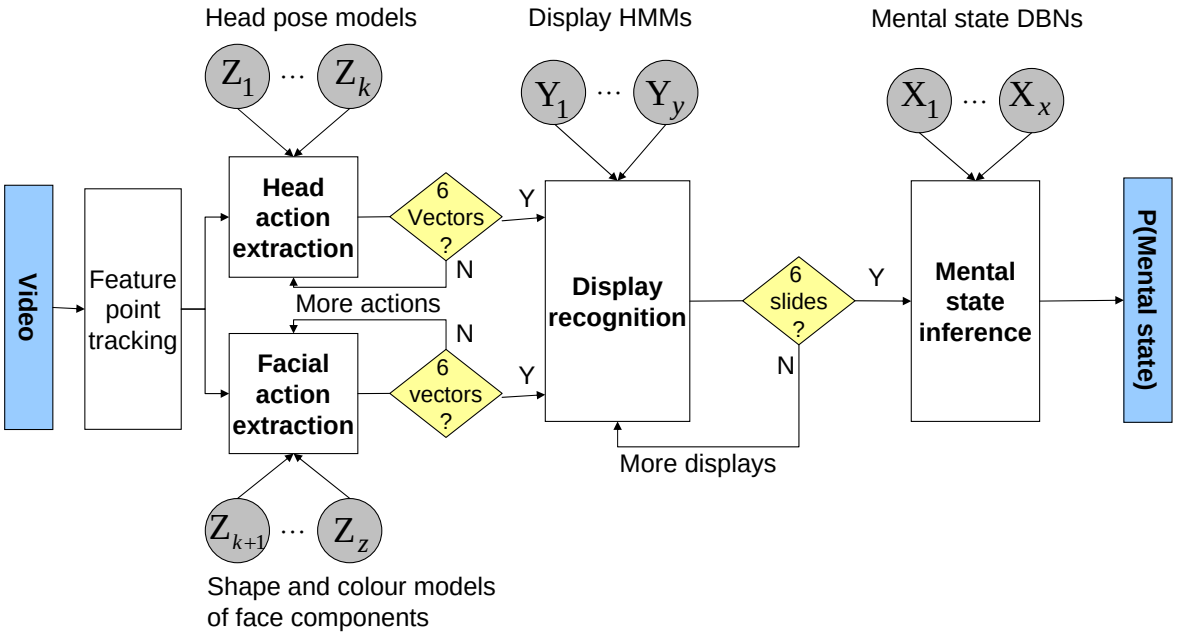


Figure 4.5: Procedural description of inference in the automated mind-reading system.

4.3 Automated mind-reading software

`Auto-MR` is the software that I have developed to code the functionality of the automated mind-reading system. `Auto-MR` has two modes of execution: an online and an off-line mode. In the online mode, the input is a live stream of video: users interact with the system by exhibiting various head and facial expressions and the system processes and recognizes these displays in real time. In the off-line mode, pre-recorded video sequences in Audio Video Interleave (AVI) format are presented to the system. Facial events are recognized and displayed to the monitor as they are processed and are also saved on disk for further analysis.

4.3.1 Hardware setup

Figure 4.6 illustrates the current setup of the online mode. A commercial digital camcorder is connected to a desktop machine. The camcorder is mounted at monitor level at a distance of approximately 120 cm from the user. The camera is situated such that it has a frontal view of the user's face, even though the user's pose may vary within this view. Typically, the video is captured at 30 fps and the frames are presented to the system in real time. I have also used the system with a webcam, in which case the accuracy of tracking drops because of the lower frame rate and resolution. The system is unobtrusive in that the user is not required to wear any tracking device and no special markers on the user's face are needed to interact with the system. Although the current implementation of the system is on a conventional desktop machine, it could be theoretically extended for use with mobile devices such as camera phones.

Figure 4.6: Hardware setup of `Auto-MR`.

4.3.2 User interface

The idea behind implementing the interface of Auto-MR was to provide researchers with some feedback about the processing that is taking place at the different levels of automated mind-reading. I started by implementing the single-frame view. In that view, a snapshot of the state of the system is taken at a single point in time t and drawn to the screen (Figure 4.7). The frame number being processed is written at the top-left of the screen. The current video frame is drawn to the center of the screen and the localized feature points and extracted features are overlayed on the frame. The amount and verbosity of the overlay can be tuned based on the needs of the user/target application. The probability $P(\mathbf{Y}[t])$ of the head/facial display events at the current time are shown by vertical bars to the left of the frame. The length of the bar is proportional to the probability and is colour-coded to indicate the degree of confidence. The probability $P(X_i[t]|\mathbf{Y}[1:t], P(X_i[1:t-1]))$ of the mental states at the current time are shown horizontally by circles above the frame. Like the vertical bars, the size of a circle is proportional to the probability and is colour-coded to indicate the degree of confidence. This view is updated in real time with every new frame in the video stream.

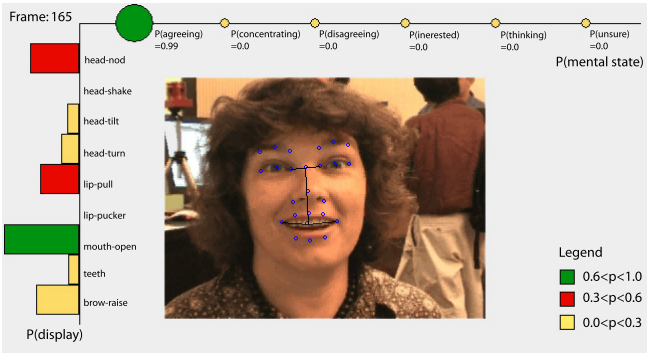


Figure 4.7: The single-frame view in Auto-MR.

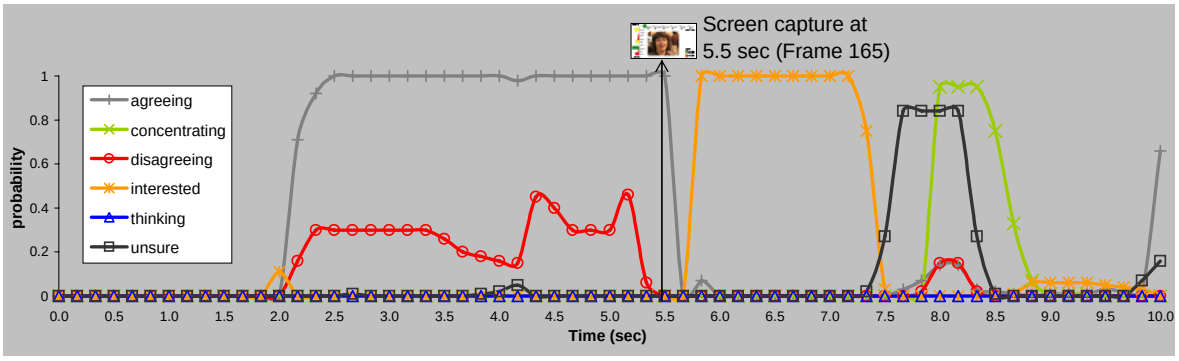


Figure 4.8: The temporal view in Auto-MR shows the mental state inferences, $P(\mathbf{X}[1:t])$ over the course of a video.

The single-frame view, while detailed, does not encode any temporal information. In particular, it does not show how the mental state inferences progress over the course of a video stream. Hence, I implemented another output view that shows the history of inferences up to the current time: $P(\mathbf{X}[1:t]|\mathbf{Y}[1:t], P(\mathbf{X}[1:t-1]))$. An example of this view is shown in Figure 4.8, with a thumbnail of the snapshot shown in Figure 4.7 superimposed to show its position in the video.

In addition, a menu-bar allows the user to toggle back and forth between the live and the pre-recorded mode of Auto-MR. In the live mode, the user can start and stop the system, and can also play the video in a non-processing mode to adjust the view. In the pre-recorded mode, a standard Windows interface allows the user to select video-files for the system to process.

4.3.3 Structure of the software

Figure 4.9 shows the architecture of Auto-MR. All the modules, with the exception of the DBN, are implemented in C++ using Microsoft Visual Studio.NET 2003 as a single-windowed application. All data processing and event handling are done within a single document class, which maintains all state information, while one or more view classes are responsible only for presenting the information provided by the document visually within a GUI.

The online and off-line frame grabbers extract a frame from the video stream, which is processed by the feature point tracker, FaceTracker. The localized feature points are then used by the head pose estimator and facial action extraction module. The actions are then passed onto the display recognizers. I have used Myers' HMM software [Mye94], a C++ implementation of HMMs used mainly in speech recognition to implement display recognition, and I integrated it with the rest of Auto-MR. In terms of data dependencies, a generic face-model needs to be made available to Facetracker along with the parameters of all the trained models. The trained models can be learned at the beginning of a system run or can be loaded from disk.

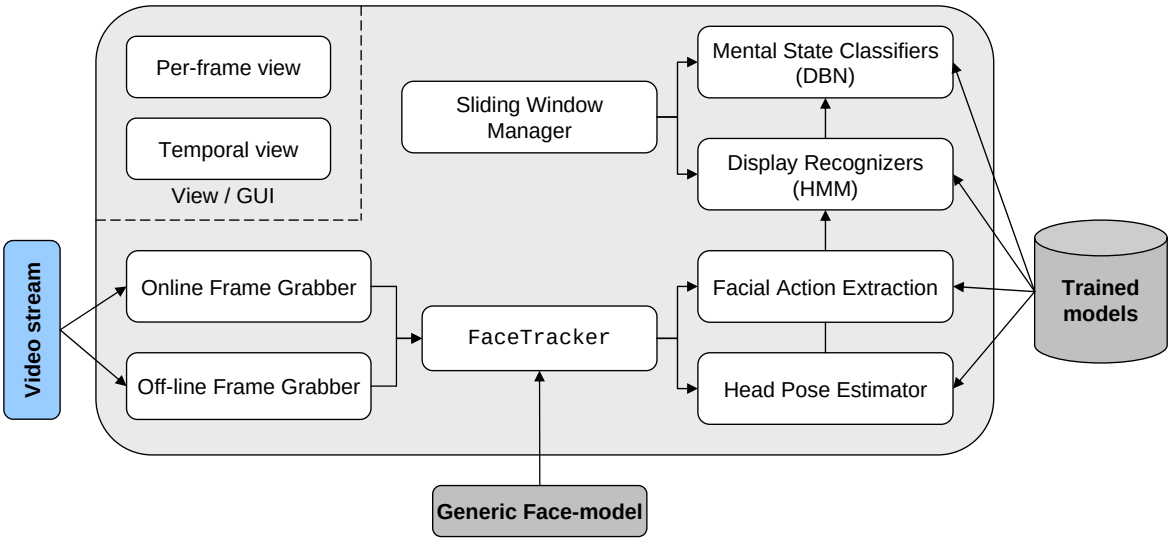


Figure 4.9: Structure of Auto-MR.

The mental state DBNs are implemented using the Bayes Net Toolbox (BNT) [Mur01], an open-source Matlab package for directed graphical models. Even though the rest of the automated mind-reading software has been implemented directly in C++, I decided to use BNT nonetheless for its simplicity and power in representing DBNs. Also at the time of implementation, BNT was the only available package for graphical models that made the source code available and supported DBNs. The integration between the display level and the mental state level is done through intermediate file-outputs.

The application is coded as a sliding window to incorporate event-history in an efficient way. Other than basic efficiency considerations, AutoMR has not been optimized for speed or for variations in recording conditions.

4.4 Discussion

The proposed framework for mental state recognition describes several contributions compared with existing facial analysis systems. First, using the theory of mind-reading as the basis for the representation and inference of complex mental states is a novel approach to the problem of mental state recognition. Secondly, the resulting computational model of mind-reading is biologically inspired: it combines bottom-up perceptual vision processes with top-down reasoning to map observable facial behaviour to hidden mental states. Third, the model also represents facial expression events at different time granularities, simulating the hierarchically structured way with which humans perceive the behaviour of others [ZTI01]. Fourth, this multi-level abstraction accounts for both intra-expression and inter-expression facial dynamics.

The computational model of mind-reading enjoys several characteristics that, as shown in Chapter 8, yield favourable results. Being hierarchical, higher-level classifiers are less sensitive to variations in the lower levels because their observations are the outputs of the middle classifiers, which are less sensitive to variations in the environment. The multi-level representation also means that the dimensionality of the state space that needs to be learned from data is much smaller than that of a corresponding monolithic model. This results in more robust performance in cases of limited training data. In addition, with each of the layers being trained independently, the framework is easier to interpret and improve at different levels. For instance, one could retrain the action level—the most sensitive to variations in the environment—without having to retrain any other level in the hierarchy.

By combining dynamic modelling with multi-level temporal abstraction, the model fully accounts for the dynamics inherent in facial behaviour. In addition, by selecting the most informative observation channels for each mental state, inference is specialized and very efficient. The probabilistic framework for inference means that it is possible to monitor how the probabilities of the mental states progress over time. This information is valuable for directing the response of applications.

In terms of implementation, the automated mind-reading software is user-independent. This is achieved by training each layer independently of the others so that the higher-level mental state models are completely independent of the facial physiognomies in the video. It is also one of the few automated facial analysis systems that account for rigid head motion while recognizing meaningful head gestures.

Finally, it is fairly easy to compose large Bayesian network models by combining subgraphs. This makes it possible to reuse and extend modelling components without retraining an entire model for each new problem, making it possible to extend the model to more mental state concepts, facial actions and displays, even other modalities and context cues.

The model however, makes a number of simplifying assumptions in the representation of facial events. First, the model is a Naive Bayes model that assumes that displays are conditionally independent of each other given the mental state. In some cases, this may not be entirely true, as in the case of a head nod and a head shake, which are mutually exclusive. Second, the causal relationship between mental states is also not represented. For instance, the mutual exclusiveness of the *agreeing* and *disagreeing* mental states is not currently accounted for by the model. Finally, the specific probabilities output by the HMMs are quantized to binary, which in turn results in loss of detail. These assumptions have been made in an attempt to favour real time execution over an accurate generative model. As mentioned earlier, a Naive Bayes Model, though simplified, is amenable to efficient learning and inference.

4.5 Summary

In this chapter I have described a computational model of mind-reading as a novel approach to the learning and inference of complex mental states. The computational model of mind-reading, and an overview of its implementation in a fully automated, real time system were presented. The model is hierarchical where each level captures a different granularity of spatial and temporal abstraction. To represent mental states from video, facial components and head pose models are defined. The relationships between these components, both spatially and temporally, constitute the models of complex mental states. Inference is then performed within a probabilistic Bayesian framework.

The computational model of mind-reading is a general framework that depicts how observed behaviour are combined to infer mental states. It is therefore not tied to a specific implementation of the event-classifiers at each level of the model. To verify the model, I have developed an automated mind-reading system that implements the computational model of mind-reading. The system combines bottom-up, vision-based processing of the face with top-down predictions of mental state models to interpret the meaning underlying video input. The system is designed to be fully automated, user-independent, and to execute in real time, making it suitable for application in HCI.

The following three chapters of this dissertation describe each of the levels that constitute the automated mind-reading system in detail.

Chapter 5

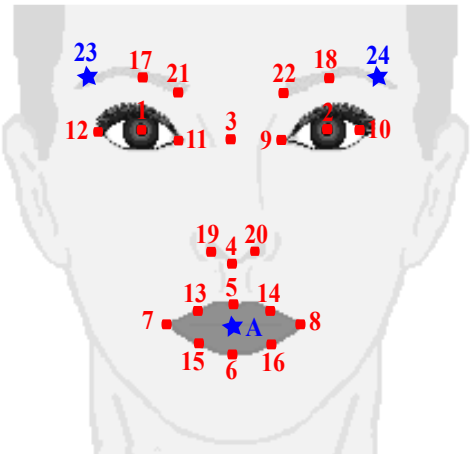
Extraction of Head and Facial Actions

Head and facial actions make up the bottom level of the computational model of mind-reading. They encode the basic spatial and motion characteristics of the head and facial features in video input. This chapter discusses the extraction of head and facial actions from a video stream, and shows how the approach adopted is suited to a system that is required to be automated, real time, and user-independent.

5.1 Face model

A number of studies have shown that the visual properties of facial expressions can be described by the movement of points belonging to facial features [EF78, Bru86, PR00a]. These feature points are typically located on the eyes and eyebrows for the upper face, the lips and nose for the lower face. Figure 5.1 illustrates the 2D face model of the 25 feature points used throughout this work. By tracking these feature points over an image sequence and analyzing their displacements over multiple frames, a characteristic motion pattern for various action units (AUs) can be established. Cohn *et al.* [CZLK99] have shown that the results of automated extraction of AUs using methods based on feature point tracking match that of manual FACS coding [EF78].

Table 5.1 describes how the head AUs that are currently supported are measured. These are divided into different "sensing" channels based on rotation axis. Table 5.2 describes the facial AUs, which are grouped into lip, mouth and eyebrow sensors. Of the complete list of AUs in FACS, I have chosen those that were straightforward to identify and that, through observation, I deemed relevant to the mental states on which I chose to focus. As explained throughout the chapter, the measurements I describe for the lips and mouth AUs are more precise and more robust to rigid head motion compared to similar measurements that also use feature-point tracking. The table includes both additive and nonadditive facial AU combinations. In a nonadditive combination, the resulting facial action has a different appearance altogether than the individual AUs, and hence requires an analysis rule of its own. For instance, the combination of a lip corner pull (AU12) with the lips parted (AU25), has a different appearance from either AUs when occurring singly.



1. Right pupil
2. Left pupil
3. Nose root
4. Nose tip
5. Upper lip center
6. Lower lip center
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 center
18. Left eyebrow center
19. Right nostril
20. Left nostril
21. Right inner eyebrow
22. Left inner eyebrow
23. Right outer eyebrow
24. Left outer eyebrow

A. Anchor point

Figure 5.1: 2D model of the face. The 22 feature points defined by `FaceTracker` are shown as a ■. For example, P_1 is the right pupil; points that are extrapolated are shown as a ★. For example, P_{24} is the left outer eyebrow.

Table 5.1: Head AUs. The measurements refer to feature points on the 2D face model in Figure 5.1. The examples are from the Mind Reading DVD [BGWH04].

Sensor	AU	Description	Example	Measurement
Yaw	51	Head turn left		Ratio of left to right eye widths $\frac{\overline{P_9P_{10}}}{\overline{P_{11}P_{12}}}$
	52	Head turn right		
Pitch	53	Head up		Vertical displacement of nose tip $P_4[t-1,t]$
	54	Head down		
Roll	55	Head tilt left		Angle of line P_9P_{11} with horizontal axis
	56	Head tilt right		

Table 5.2: Facial AUs. The measurements refer to feature points on the 2D face model in Figure 5.1. The examples are FACS-coded image sequences from the Cohn-Kanade database [KCT00] of facial expressions. Symbols are as follows: c is the threshold for lip motion n_α, n_τ are the number of aperture and teeth pixels within the mouth polygon; c_α, c_τ minimum amount of aperture and teeth. As explained throughout the chapter, the measurements described here for the lips and mouth AUs are more precise and more robust to rigid head motion compared to similar measurements that also use feature-point tracking.

Sensor	AU	Description	Example	Measurement
Lips	12	Lip corner pull		Increase in polar distance $(\overline{AP_7} + \overline{AP_8})[0,t] > c$
	20	Lip stretch		
	18	Lip pucker		Decrease in polar distance $< -c$
Mouth	26	Jaw drop		Significant presence of aperture $(n_\alpha \geq n_\tau) \wedge (n_\alpha \geq c_\alpha)$
	27	Mouth stretch		
	12+27	Lip corner pull and mouth stretch		
	12+25	Lip corner pull and lips part		Significant presence of teeth $n_\tau \geq c_\tau$
	20+25	Lip stretch and lips part		
	20+26	Lip stretch and jaw drop		
	25	Lips part		$n_\alpha + n_\tau \approx 0$
Eyebrows	1	Inner brow raise		Increase in eye-eyebrow distance $(\overline{P_{11}P_{21}} + \overline{P_1P_{17}} + \overline{P_{12}P_{23}})[t-1,t]$
	1+2	Brow raise		

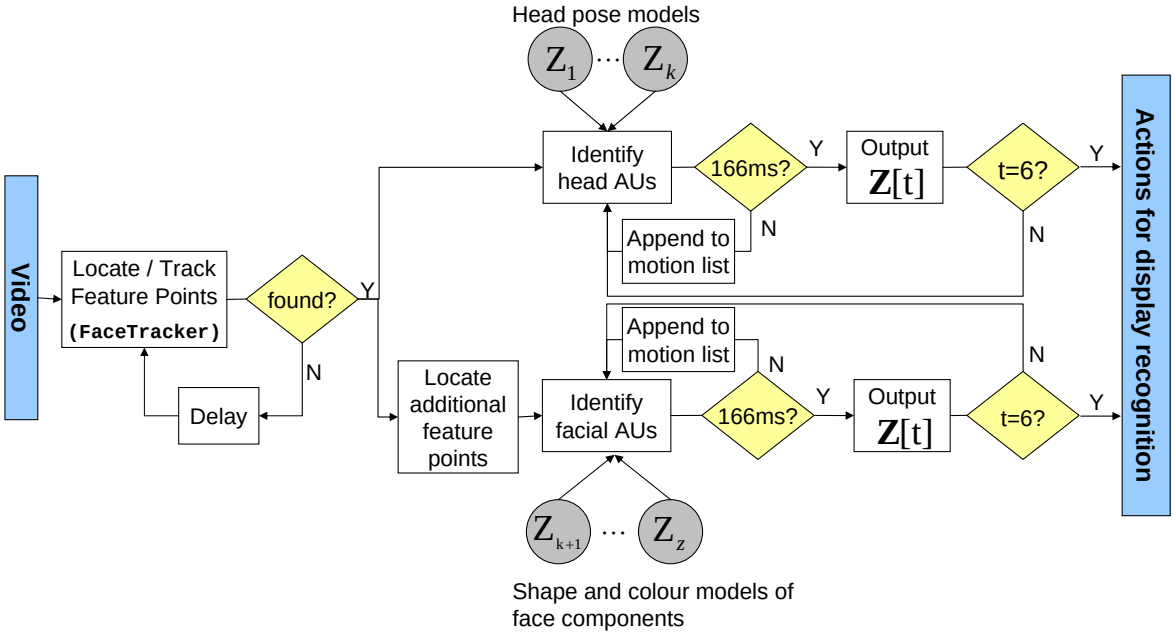


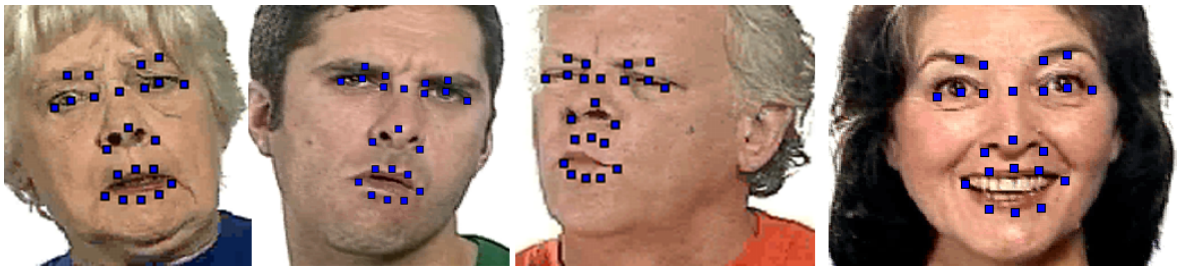
Figure 5.2: Procedural description of facial and head action extraction from video. On each frame, feature points are located and tracked. If successful, head and facial AUs are identified and accumulated over a time span of 166 ms after which a vector $\mathbf{Z}[t]$ of z action symbols is output. Whenever 6 consecutive symbols of each action have been accumulated, display recognition (Chapter 6) is invoked.

Figure 5.2 shows a procedural description of head and facial action extraction. On each frame of a video, feature points are located and tracked. If successful, head AUs are identified along each rotation axis. Shape and colour models of face components such as the lips and mouth are used to identify the underlying facial AUs. As mentioned in Chapter 4, the AUs are analysed over 5 consecutive frames in a video sampled at 30 fps. The output $\mathbf{Z}[t]$ is a vector of z head and facial action symbols at time t. Whenever $t = 6$, the vector of consecutive actions $\mathbf{Z}[1 : 6]$ is presented to the next level of the system: display recognition (Chapter 6).

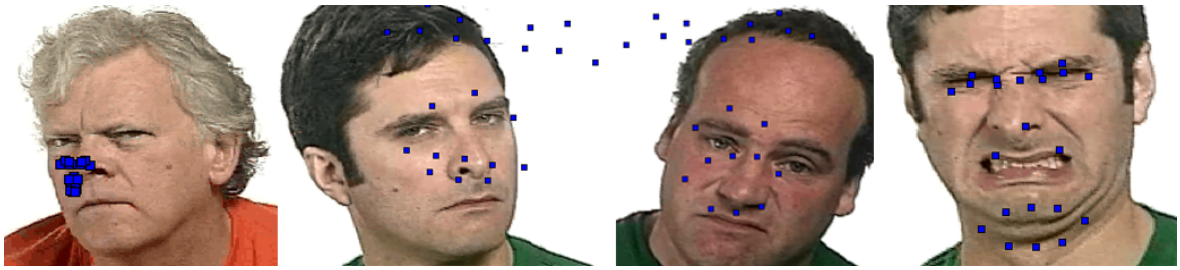
5.2 Interface to `FaceTracker`

For feature point tracking I use `FaceTracker` [Fac02], part of Nevenvision's facial feature tracking SDK. `FaceTracker` uses a generic face template to bootstrap the tracking process, initially locating the position of 22 facial landmarks (shown as ■ in Figure 5.1). To track the motion of the points over a live or recorded video stream, the tracker uses a combination of Gabor wavelet image transformations and neural networks. While tracking proceeds on 2D video input, a learned 3D model of the human face is used to correct tracking errors and cope with pose variations. In the event that the tracking process fails, as in a sudden large motion of the head, tracking is delayed for 5 ms before re-attempting to locate the feature points.

`FaceTracker` deals with a wide range of face physiognomies and skin colour, and tracks users that wear glasses and/or have facial hair. The tracker also deals with non-initial neutral frames, a key feature that most other existing tracking systems do not currently support. Figure 5.3(a) shows examples of initial frames on which the tracking system correctly localizes the feature points. These frames include non-frontal poses and a variety of expressions such as a smile. However, the tracker will still fail to locate feature points under certain conditions such as those shown in Figure 5.3(b).



(a) Initial frames on which FaceTracker correctly locates facial landmarks. Note that the frames are of non-neutral expressions and non-frontal poses. Also note the range of facial physiognomies, age range and skin colour that the tracker is able to deal with.



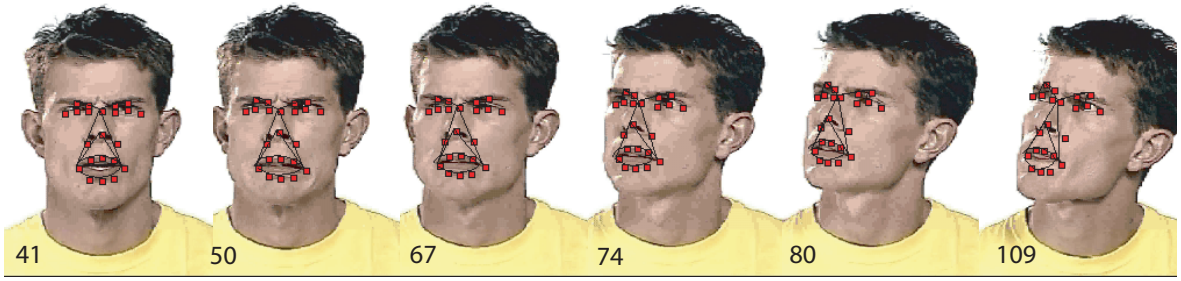
(b) Initial frames on which FaceTracker fails to locate facial landmarks.

Figure 5.3: FaceTracker's localization of facial landmarks for a variety of initial frames. The frames are extracted from videos from the Mind Reading DVD.

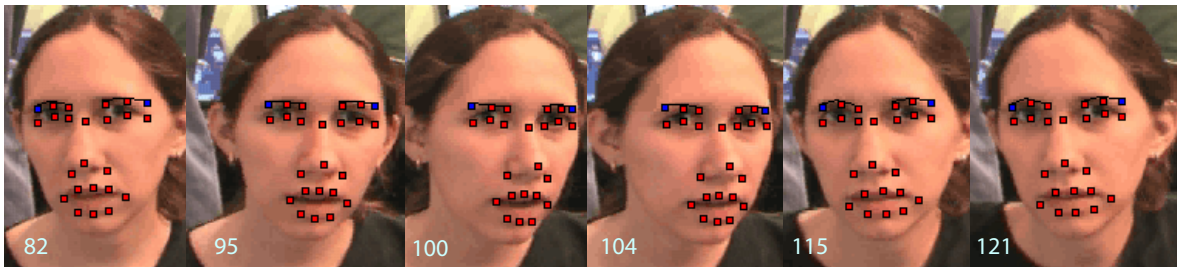
The tracker is also robust to a wide range of magnitude and velocity of in-plane and out-of-plane rigid head motion. Natural human head motion ranges between $70 - 90°$ of downward pitch, $55°$ of upward pitch, $70°$ of yaw (turn) and 55^o of roll (tilt) [Kur03]. Out of this range, the tracker supports up to $50°$, $45°$ and $30°$ of pitch, yaw and roll respectively subject to factors such as changes in illumination, the underlying facial expression and velocity of rigid head motion.

Through experimentation, I found that the FaceTracker is able to correctly track videos with head rotation speed of up to $4°$ per frame and translation of up to 20 pixels per frame. At 30 fps this accounts for most naturally occurring facial expressions, except for severe and sudden jerkiness, which is infrequent in most typical HCI contexts. Figure 5.4 shows examples of the robustness of FaceTracker to different degrees and velocities of in-plane and out-of-plane head motion. Figure 5.4(a) shows selected frames from a video that has up to $7°$ of pitch, $26°$ of head yaw and $16°$ of head roll, at a maximum angular velocity of $2°$ per frame. Figure 5.4(b) has up to $16°$ of pitch, $20°$ of head yaw and $20°$ of head roll, at a maximum angular velocity of $1.5°$ per frame. Rigid head motion also typically occurs as a combination of motion along all three rotation axes, as is evident in the examples in Figure 5.4. This contrasts with the highly controlled head motion patterns that are prevalent in existing facial expression databases such as the Cohn-Kanade database [KCT00].

Figure 5.5 shows examples when feature point tracking fails. In Figure 5.5(a), the video has up to $32°$ of pitch, $10°$ of head yaw and $15°$ of head roll, at a maximum angular velocity of $6°$ per frame. Tracking is successful up to frame 21. Between frames 22 and 25, when the head pitches downward at a velocity of $6°$ per frame, the lower, then upper, feature points are gradually lost. Similarly, Figure 5.5(b) shows selected frames from a video that has up to $23°$ of pitch, $24°$ of head yaw and $9°$ of head roll, at a maximum angular velocity of $3°$ per frame. Tracking is successful up to frame 100. Between frames 100 and 108 the feature points are gradually lost.



(a) Selected frames from a video that has up to 7° of pitch, 26° of head yaw and 16° of head roll, at a maximum angular velocity of 2° per frame. The video (30 fps) shows the mental state *undecided* from the Mind Reading DVD.



(b) Selected frames from a video that has up to 16° of pitch, 20° of head yaw and 20° of head roll, at a maximum angular velocity of 1.5° per frame. The video (30 fps) shows the mental state *unsure*, from the CVPR 2004 corpus.

Figure 5.4: Robustness of `FaceTracker` to a combination of in-plane and out-of-plane rigid motion and angular velocity of the head.

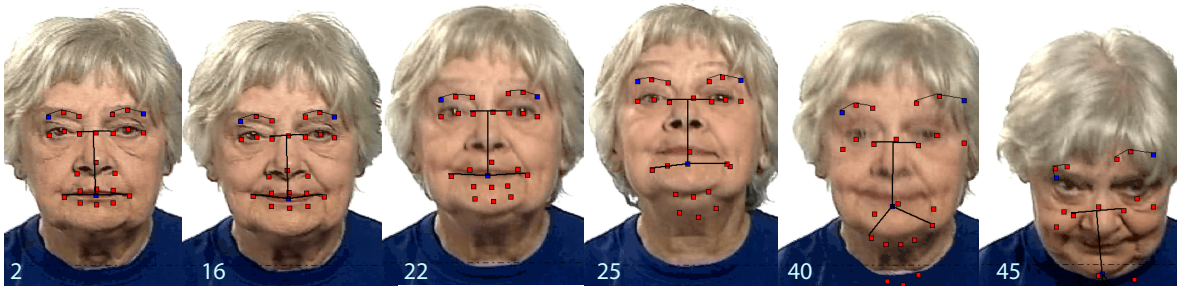
`FaceTracker` is accessed through an SDK referred to as a Core Technology Interface. The interface allows the position of the landmark points to be queried and provides an interface to the actual video stream. To start the tracking process, a tracker object is created and the 22-point face model files are loaded. Once a tracker object is instantiated and initialized, an observer class generates a number of events whenever a new video frame is present or new landmark positions are available.

5.3 Additional feature points

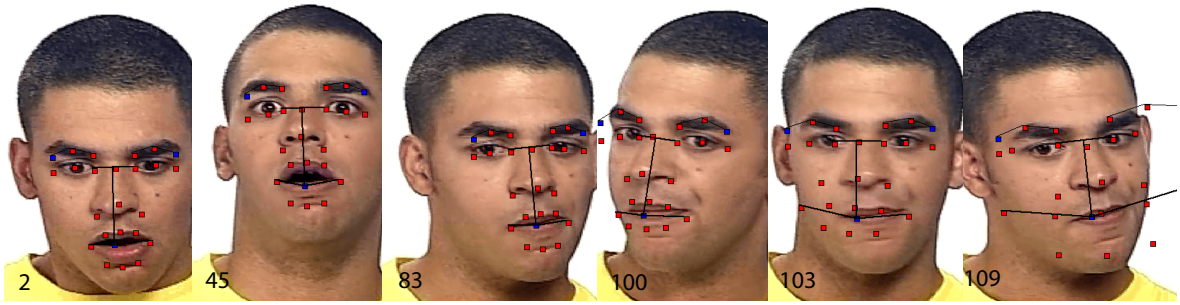
In addition to the 22 facial landmarks output by `FaceTracker` on each successful frame, I define three points that are calculated on each frame: the anchor point A and two outer eyebrow points (P_{23} and P_{24}). These were shown on Figure 5.1 as a ★.

5.3.1 Anchor point

The anchor point A serves a center of projection for the face; it is a 2D projection of the point around which the head rotates in 3D space. The point remains stable with respect to the face and is independent of the deformation of facial features. Algorithm 5.1 describes how the anchor point is localized on the initial frame and calculated on subsequent frames. The anchor point is initially defined as the midpoint between the two mouth corners when the mouth is at rest, and is at a perpendicular distance d from the line joining the two inner eye corners L_{eyes}. In subsequent frames the point is measured at distance d from L_{eyes} (Figure 5.6). The anchor point is, by definition, insensitive to in-plane head motion (head rolls) and is also resilient to a range of out-of-plane head motion. With head pitch motion, the midpoint of the distance d on L coincides



(a) Selected frames from a video that has up to 32° of pitch, 10° of head yaw and 15° of head roll, at a maximum angular velocity of 6° per frame. Tracking is successful up to frame 21. Between frames 22 and 25, when the head pitches downward at 6° per frame, the lower, then upper, feature points are gradually lost. The feature points are completely lost starting frame 25. Note how due to speed of motion the frames are out of focus. The video (30 fps) shows the mental state *decided*.



(b) Selected frames from a video that has up to 23° of pitch, 24° of head yaw and 9° of head roll, at a maximum angular velocity of 3° per frame. Tracking is successful up to frame 100. Between frames 100 and 108 the feature points are gradually lost. Frame 109 onwards tracking is unsuccessful. The video (30 fps) shows the mental state *cheated*.

Figure 5.5: The effect of the degree and velocity of rigid head motion on the performance of `FaceTracker`. The videos are from the Mind Reading DVD.

with the hypothetical point around which the head pitches upward or downward. Hence the effect is similar to that of a perspective transformation. The current version of the localization algorithm does not account for scale variations within a video sequence, although it is possible to normalize the distance d against changes in the length of L_{eyes}. Because facial actions need to be resilient to intra-class variations that arise from rigid head motion, the anchor point serves as a good reference point for shape-based analysis of lower face actions, including asymmetric ones. The use of the anchor point in the extraction of facial actions is described in more detail in Section 5.5.

The anchor point has several advantages over other common centers of projection, such as the nose tip. It lies on the same plane of other facial features. It is more robust to occlusion of the lower face typically caused by the hand being placed in front of the

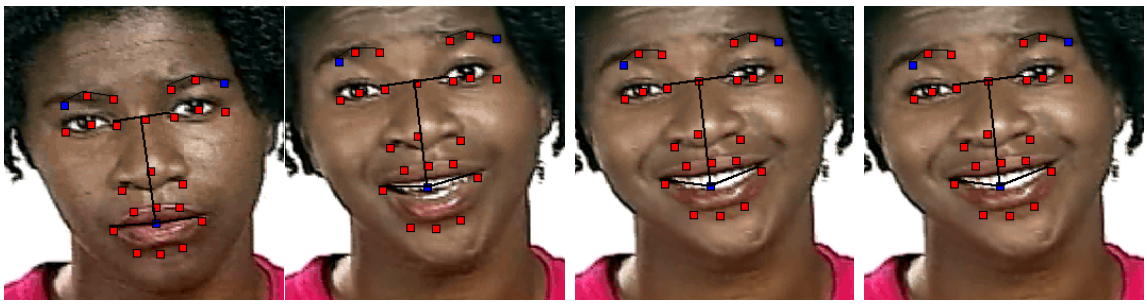


Figure 5.6: Example of anchor point on initial and subsequent frames of a video.

Algorithm 5.1 Anchor point localization

Objective: Calculate the initial and subsequent locations of anchor point A

Given: The set of 22 localized feature points $\{P_i\}_{i=1}^{22}$ at frame t and the line L_{eyes} connecting the two inner eye corners P_9 and P_{11}, where the midpoint of L_{eyes} coincides with the nose root P_3

On initial frame:

Let L_{mouth} be the line connecting the mouth corners P_7 and P_8
Drop a line L from P_3 such that $L \perp L_{eyes}$
A is initially where L intersects L_{mouth}, at a distance d from P_3

On subsequent frames: A lies on L at a distance d from P_3

mouth, because it is computed using feature points in the upper face region. The eye corners and lip corners, which are used in the computation of the anchor point, are generally easier to track on the face than is the nose tip. This is due to their physical appearance typically having more colour contrast, regardless of ethnicity, than the nose tip. Finally because the anchor point lies in the middle of the mouth, the polar distance from the anchor point is particularly well suited to capturing the lip corner deformation (lip corners move out and up, out and down or inward).

5.3.2 Outer eyebrow points

The default face template of `FaceTracker` locates and tracks only the inner and central eyebrow points. I use the height of the inner eyebrow points, measured from the inner eye corners to extrapolate the location of the outer eyebrow points measured from the outer eye corners.

5.4 Extraction of head actions

With the exception of a few facial expression analysis systems such as Colmenarez *et al.* [CFH99] and Xiao *et al.* [XKC02], the problem of rigid head motion is avoided altogether by requiring subjects to limit their head movements strictly. This is an unrealistic assumption because head movements occur frequently in spontaneous interactions. More importantly, head gestures and head orientation, like facial expressions, play a prominent role in communicating social and emotional cues [CLK$^+$02, LWB00]. Hence, one of the objectives in implementing the automated mind-reading system, was to extract head actions to enable the recognition of intentional head gestures.

A simple and computationally efficient approach to head pose estimation uses the motion of feature points over successive frames to extract head rotation parameters [MYD96, JP00, KO00, TN00, TKC00b, DV01, KP01]. The alternative approach involves tracking the entire head region using a 3D head model. Different 3D head models have been investigated including anatomical models [BV99, ES02], planar-based [BY95], ellipsoidal [BEP96] and cylindrical models [CSA00, XKC02]. While more accurate, 3D head models require precise or manual initialization to work well, are more computationally intensive and do not always run in real time. Since the objective here is to identify head actions automatically given a frontal view of the face and to do so in real time, rather than come up with a precise 3D estimate of the head pose, a feature-point based approach was deemed more suitable than a model-based one.

Algorithm 5.2 Head yaw extraction*

Objective: Extract 6 head yaw symbols $Z_k[1:6]$

Given: The set of 25 localized feature points $\{P_i\}_{i=1}^{25}$ at frame t, and $\mathbf{M} = \{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_k\}$ vector of k head motions identified along the yaw axis so far, where M_k comprises the following set of variables: the direction ($\mathrm{AU}51$, $\mathrm{AU}52$ or *null*), start frame, end frame and intensity (energy) of the motion

TrackYaw

for all frames i **do**
$\angle_{\text{yaw}} = \frac{\overline{P_9P_{10}}}{\overline{P_{11}P_{12}}}$, the ratio of left to right eye widths or as output by `FaceTracker`
if $\angle_{\text{yaw}} > \varepsilon$, where ε is a minimum-rotation threshold **then**
if direction of $\angle_{\text{yaw}}$ and M_k (most recent stored motion) are equal **then**
if duration of M_k is less than 166 ms **then**
AppendMotion
else
OutputHeadAction
Start a new motion M_{k+1}
else
OutputHeadAction
Start a new motion M_{k+1}

AppendMotion: Update the end frame and intensity of M_k

OutputHeadAction: Output a symbol $Z_k[t]$ with direction equal to that of M_k and intensity one of low ($0-15°$), medium ($15-30°$) or high (above $30°$)

* The same algorithm applies to head pitch and head roll. The pitch and roll angles, $\angle_{\text{pitch}}$ and $\angle_{\text{roll}}$ are given by `FaceTracker`. They can also be estimated as described in Table 5.1

The following head AUs are extracted: the pitch actions $\mathrm{AU}53$ (up) and $\mathrm{AU}54$ (down), yaw actions $\mathrm{AU}51$ (turn-left) and $\mathrm{AU}52$ (turn-right), and head roll actions $\mathrm{AU}55$ (tilt-left) and $\mathrm{AU}56$ (tilt-right). The rotations along the pitch, yaw and roll axes, $\angle_{\text{yaw}}$, $\angle_{\text{pitch}}$ and $\angle_{\text{roll}}$ respectively, are calculated from expression invariant points. These points are the nose tip, nose root and inner and outer eye corners.

Algorithm 5.2 describes how head actions along the yaw axis are identified. The algorithm keeps a vector $\mathbf{M} = \{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_k\}$ of k motions identified along the yaw axis so far. On each frame, the yaw angle $\angle_{\text{yaw}}$ is given by `FaceTracker` or by the ratio of the left to right eye widths. To be considered as a valid head action, the angle has to meet the threshold ε. The direction of $\angle_{\text{yaw}}$, in addition to the direction and duration of the most recent motion M_k, determines whether a symbol $Z_k[t] \in \{\mathrm{AU}51, \mathrm{AU}53, \mathit{null}\}$ is output or not. The intensity of the motion is encoded as low, medium or high depending on the magnitude of the angle.

The same algorithm is used to track the head pitch and head roll actions. The pitch and roll angles, $\angle_{\text{pitch}}$ and $\angle_{\text{roll}}$ are computed by `FaceTracker` as euler angles. They can also be implemented as the vertical displacement of the nose tip between frames $[t-1, t]$ for head pitch, and the image-plane rotation angle calculated using the two inner eye corners for head roll.

5.5 Extraction of facial actions

Facial actions are identified from motion, shape and colour descriptors. Motion and shape-based analysis meet the constraints imposed by real time systems in which mo-

tion is inherent. While shape descriptors capture the deformation of face components such as the lips and eyebrows, they fall short of accurately representing some facial AUs such as those of the mouth. Colour-based analysis complements shape-based analysis and is also computationally efficient. It is invariant to the scale or viewpoint of the face, especially when combined with feature localization that limits the analysis to regions already defined by the tracker.

5.5.1 Lip actions

The lips are described by the eight feature points and the anchor point in the 2D face model (Figure 5.1). The eight feature points also define the bounds of the mouth polygon. I use shape-based analysis to identify the following lip AUs: lip corner pull, lip stretch and lip pucker (Table 5.2).

Algorithm 5.3 Extraction of lip actions

Objective: Extract $t = 6$ consecutive lip shape symbols $Z_k[1:6]$

Given: The set of 25 localized feature points $\{P_i\}_{i=1}^{25}$ at frame t, and $\mathbf{M} = \{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_k\}$ vector of k lip AUs identified so far, where M_k comprises the following set of variables: the description ($\mathrm{AU12}$, $\mathrm{AU20}$, $\mathrm{AU18}$ or *null*), start frame, end frame and energy of the action

Track lips

for all frames i **do**
$\delta = (\overline{AP_7} + \overline{AP_8})[0, t]$
if $\delta > \alpha$, where α is a neutral range threshold **then**
currentLipAction = $\mathrm{AU12}$ or $\mathrm{AU20}$ (lip corner pull, or lip stretch)
else
if $\delta < -\alpha$ **then**
currentLipAction = $\mathrm{AU18}$ (lip pucker)
else
currentLipAction = null
ProcessLipAU

ProcessLipAU

if currentLipAction and M_k (most recent stored lip AU), are equal **then**
if duration of M_k is less than 166 ms **then**
AppendLipAction
else
OutputLipAction
Start a new lip action M_{k+1}
else
OutputLipAction
Start a new lip action M_{k+1}

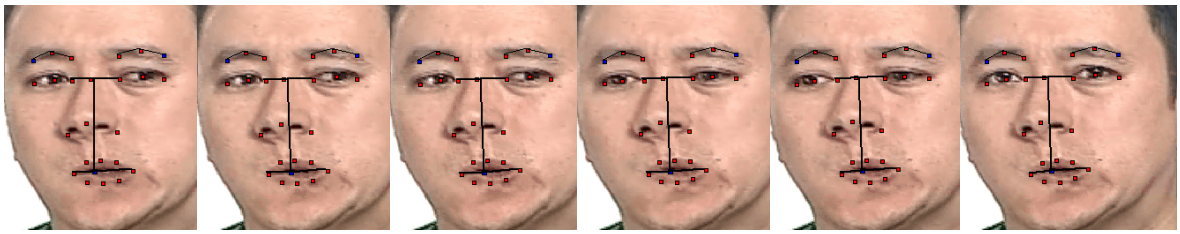
AppendLipAction: update the end frame of M_k

OutputLipAction: output a symbol $Z_k[t]$ of description equal to that of M_k

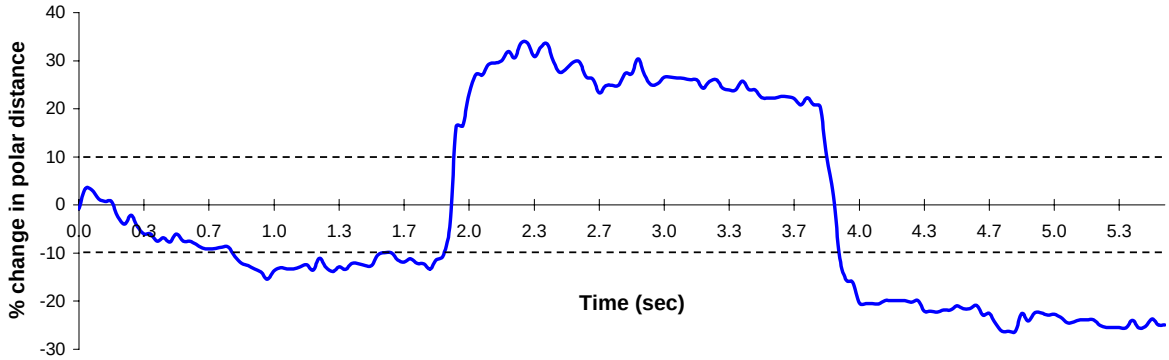
---

I define the polar distance as the distance between each of the two mouth corners and the anchor point. The use of polar distances to analyse lip shape has several advantages over the width and height parameters that are typically used, such as in Tian *et al.* [TKC00b] and Oliver *et al.* [OPB97]. In particular, the use of polar distances has

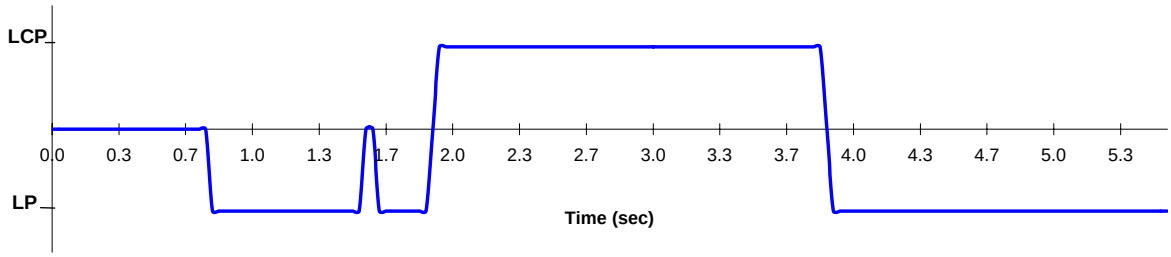
been suggested by Schmidt *et al.* [SC01] for their discriminative ability in expression classification. Measuring the lip corner motion with respect to a stable anchor point works well in spite of rigid head motion, and is more robust to inaccurate feature point tracking, compared with geometric mouth width and height parameters. Polar distances can also be used to describe facial action asymmetry.



(a) Lip pucker actions at 1.9 seconds in a video showing the mental state *unsure* from the Mind Reading DVD (167 frames at 30 fps).



(b) Percentage change in polar distance. Two segments of lip puckers can be seen between 0.8-1.9 and 3.8-5.6 seconds, and a lip corner pull between 1.9-3.8. The threshold α is shown as a dotted line.

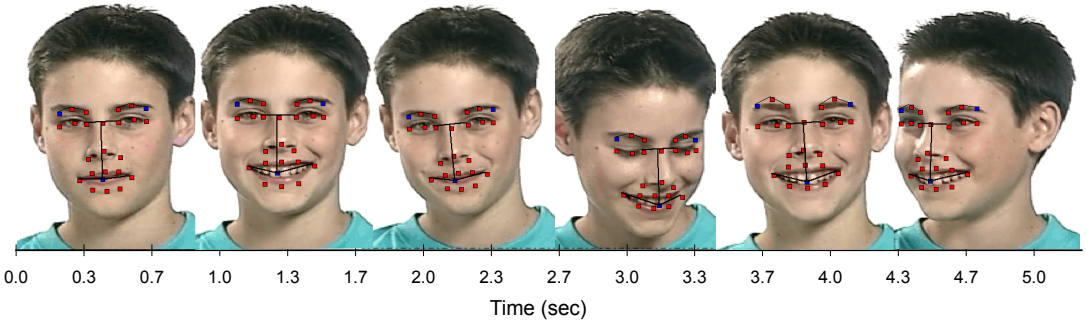


(c) The output symbol sequence for lip actions. LCP: lip corner pull, LP: lip pucker.

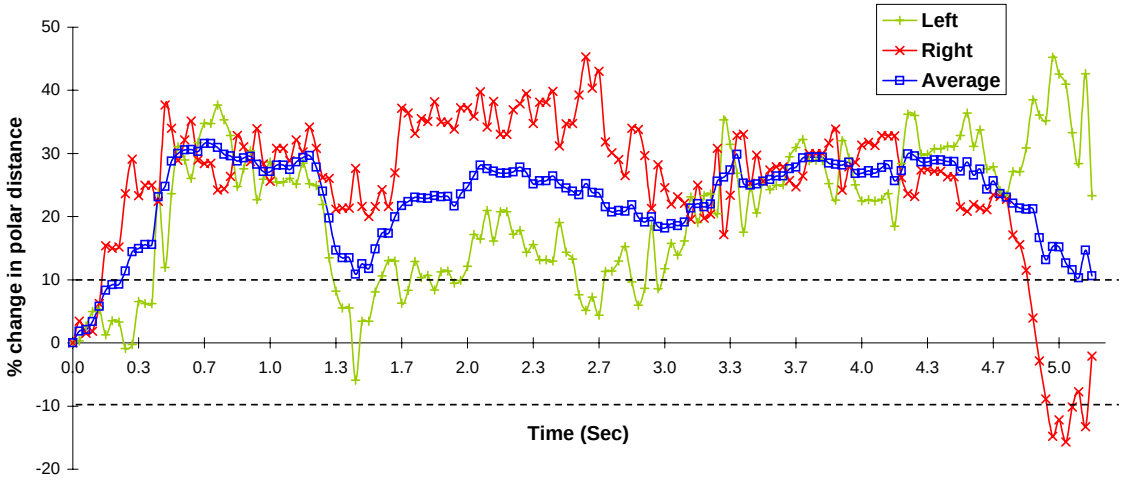
Figure 5.7: Classification of lip pull and lip pucker actions.

Lip action extraction proceeds as described in Algorithm 5.3. On each frame of the video, the average change in the distance between the anchor point and each of the two mouth corners is calculated. This is normalized to the corresponding polar distance on the initial frame to minimize the effects of variation in face size between image sequences. An increase of $\alpha = 10\%$ or more, depicts a lip corner pull or lip stretch. A decrease of 10% or more is classified as a lip pucker. The sign of the change indicates whether the action is in its onset, apex or offset.

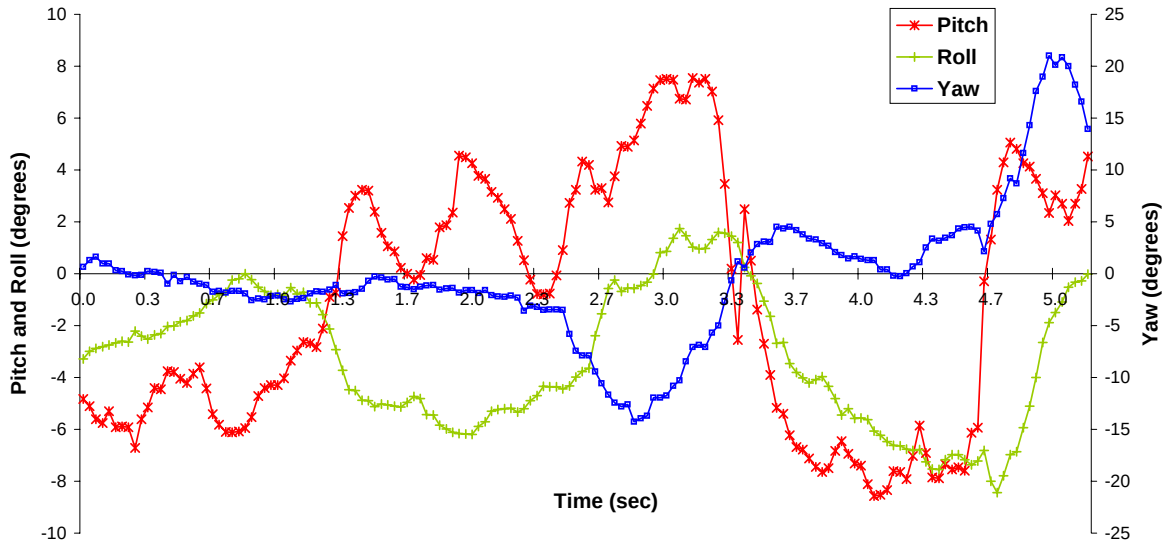
Figure 5.7 shows the result of analyzing the polar distances in a video containing both lip corner pull and lip pucker actions. The facial actions occur alongside a head yaw. Figure 5.8 demonstrates the resilience of lip action extraction to rigid head motion. In that figure, the lip corner pull is correctly detected in spite of a head yaw of up to 22° and a head roll and pitch of almost 10°.



(a) The video is of the mental state *amused* from the Mind Reading DVD (157 frames at 30 fps).

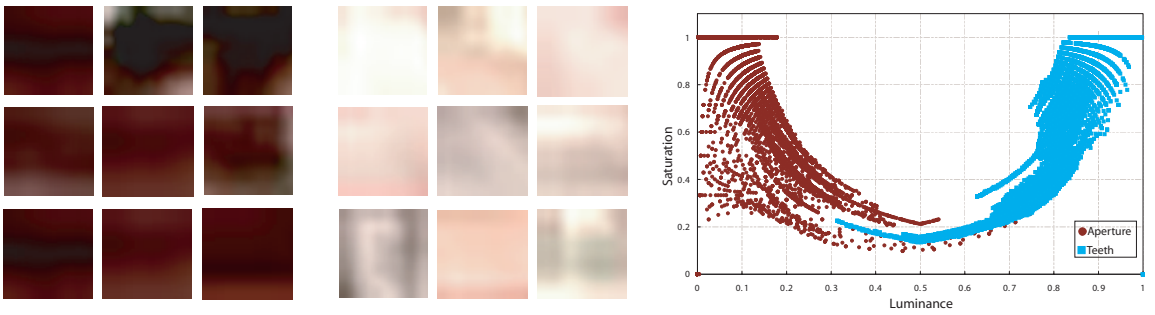


(b) Polar distance of the two mouth corners, and their average. Note that the smile has several peaks. The threshold α is shown as a dotted line.



(c) Head pitch, roll and yaw.

Figure 5.8: The resilience of lip action extraction to out-of-plane rigid head motion.



(a) Aperture regions. (b) Teeth regions. (c) Aperture and teeth in luminance-saturation space.

Figure 5.9: Aperture and teeth samples used to train the Gaussian models, and a plot of their values in saturation-luminance space.

The principal cause of falsely detected or undetected lip actions is that shape parameters are normalized to an initial frame that is assumed to be a neutral one. A video that starts with a non-neutral frame may result in a misinterpretation of the action. This depends on the intensity of the lip action on the initial frame and how the lip actions progress over the course of the video. For instance, if a sequence starts with a lip pucker that persists over time, then the pucker action will be undetected.

Static-based analysis of the initial frame using the polar angle and colour analysis would alleviate this problem by approximating the initial lip state. Another common cause of misclassifications is that of erroneous tracking, which happens in some cases of a lip corner pull: a drift in tracking, results in the feature points getting fixated in the lip-pull position.

5.5.2 Mouth actions

In addition to lip actions, mouth actions also play an important role in conveying facial displays. The mouth AUs that are listed in Table 5.2 are divided into three groups depending on the extent of aperture and teeth present:

- **Mouth-open**: a significant extent of aperture is present such as in a jaw drop (AU26), and the more exaggerated form of that action, a mouth stretch (AU27), and combinations of these AUs, such as $\mathrm{AU12} + 26$.
- **Teeth-present**: although aperture may still be present, it is the significant presence of teeth that characterizes these actions. Examples include combinations of mouth and lip actions such as lip corner pull and lips part as in a smile ($\mathrm{AU12} + 25$), lip stretch and lips part ($\mathrm{AU20} + 25$), and lip stretch and jaw drop ($\mathrm{AU20} + 26$). Note how the lip actions change the appearance of the lips part (AU25).
- **Closed or other**: no significant presence of aperture or teeth is detected such as when the mouth is closed or lips parted (AU25).

I use region-based colour tracking to discern the aperture and teeth regions inside a mouth. Region-based colour tracking has been used to solve a variety of machine vision problems including face detection [TFAS00, HAMJ00, Hoe04], facial expression analysis [TKC00b] and mouth tracking [OPB97]. The idea is to measure the colour of pixels inside some region of interest over local neighbourhoods, and use these colour values

to determine whether the pixels belong to some class. Note that colour analysis of the bounding polygon of the mouth area, determined by the eight lip feature points, is more efficient and defines a better search space than the rectangular bounding box typically used in colour analysis of the mouth as in the systems described in Oliver *et al.* [OPB97] and Tian *et al.* [TKC00b].

Figure 5.9 shows a sample of aperture and teeth regions, and a plot of the samples in luminance-saturation space. Luminance, given by the relative lightness or darkness of the colour, acts as a good classifier of the two types of mouth regions. A sample of $n = 125000$ pixels was used to determine the probability distribution functions of aperture and teeth. A lookup table defining the probability of a pixel being aperture given its luminance is computed for the range of possible luminance values (0% for black to 100% for white). A similar lookup table is computed for teeth. The luminance value at each pixel in the mouth polygon is used as an index to obtain the probability of that pixel being aperture or teeth.

Algorithm 5.4 Region-based colour analysis of the mouth

Objective: Extract $t = 6$ consecutive mouth action symbols $Z_k[1:6]$

Given: The probability density functions P_α and P_τ for aperture and teeth respectively, and the mouth polygon containing a total of N pixels

1. **Initialization:** $n_\alpha = n_\tau = 0$, where n_α and n_τ denote the number of aperture and teeth pixels found in the mouth respectively

2. **Scan mouth for teeth and aperture:**
   for all pixels i in N **do**
   Calculate the luminance l_i luminance of pixel i
   Compute probability that i is aperture $p_\alpha(i) = P_\alpha(l_i)$
   Increment n_α if $p_\alpha(i) \geq c^*$
   Compute probability that i is teeth $p_\tau(i) = P_\tau(l_i)$
   Increment n_τ if $p_\tau(i) \geq c^*$

3. **Classification:**
   Normalize $n_\alpha = n_\alpha/N$ and $n_\tau = n_\tau/N$
   if $(n_\alpha \geq n_\tau) \wedge (n_\alpha \geq c^*_\alpha)$ **then**
   Mouth-open: jaw drop (AU26) or mouth stretch (AU27)
   else
   if $n_\tau \geq c^*_\tau$ **then**
   Teeth-present: lip corner pull and lips parted (AU12 + 25), lip stretch and lips parted (AU20 + 25), and lip stretch and mouth stretch (AU20 + 27)
   else
   Closed or other: mouth closed or lips parted (AU25)

*Thresholds c, c_α and c_τ are determined empirically.

Online classification is summarized in Algorithm 5.4 and proceeds as follows: on every frame in the sequence, the luminance value of each pixel in the mouth polygon is computed. The luminance value is then looked up to determine the probability of the pixel being aperture or teeth. Depending on empirically determined likelihood thresholds, the pixel is classified as aperture or teeth or neither. The total number of aperture and teeth pixels in the polygon are used to classify the mouth region into one of the three groups described earlier: mouth-open, teeth-present or closed. Figure 5.11

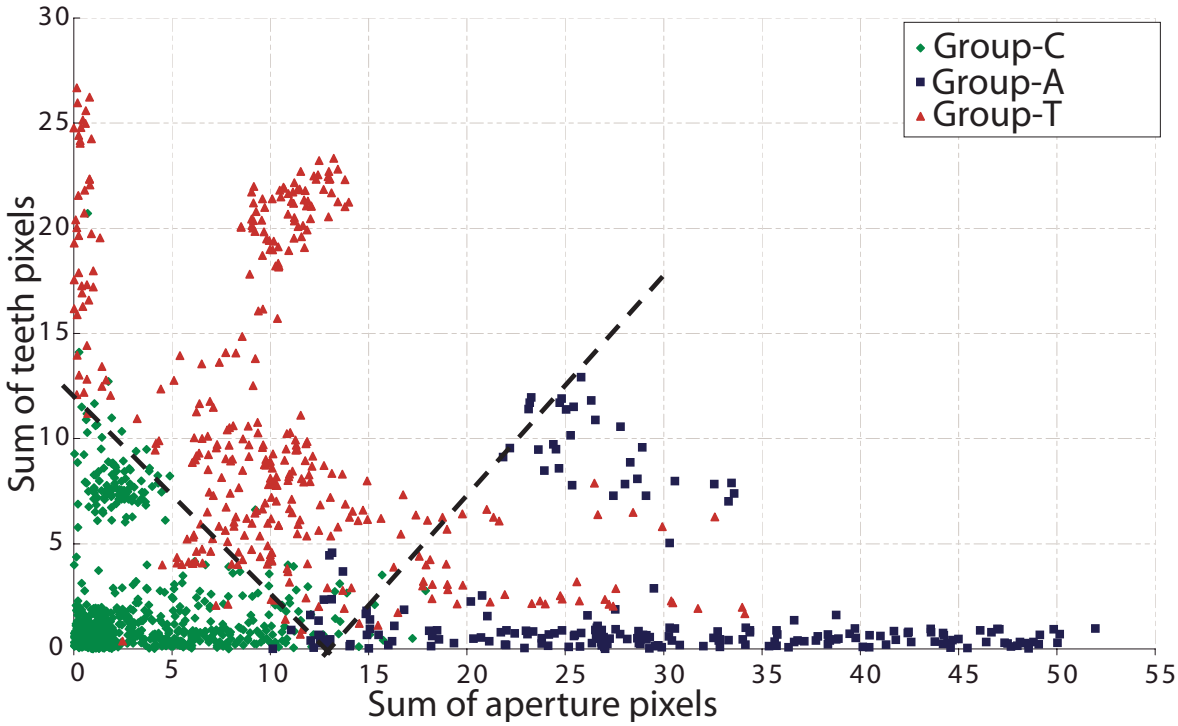


Figure 5.10: Classifying 1312 mouth regions into actions that belong to mouth-open, teeth-present, or mouth-closed.

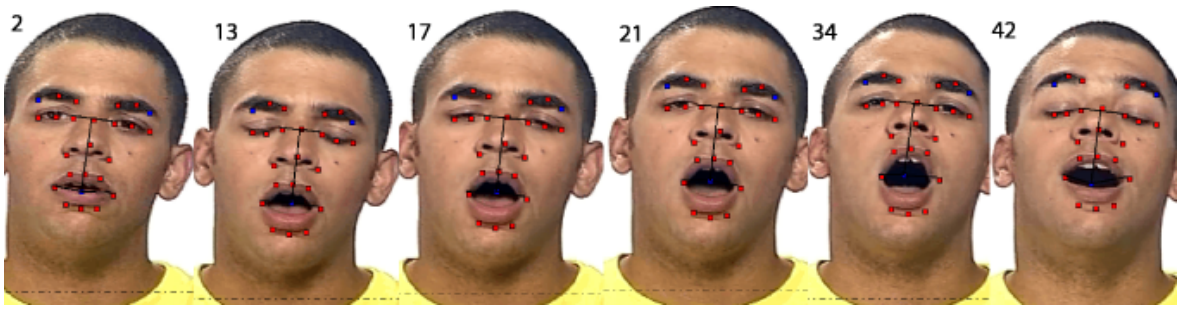
shows the aperture and teeth masks in the mouth polygon in an example of mouth-open actions. Figure 5.12 shows similar results in an example of teeth-present actions.

The main problem with an approach that relies on luminance is that features that are extracted from frames with lighting effects may be projected to an incorrect region in the luminance-saturation space, resulting in a misclassification. Figure 5.10 shows classification results of 1312 frames into mouth-open, teeth-present and mouth-closed actions. These results show that the detection of aperture is more robust to these changes in illumination than the detection of teeth. The false positive cases of teeth-present actions are caused by specular highlights around the lips, while the undetected cases are due to the wide variation of possible teeth colours that are not accounted for in the examples used to train the teeth-model. To improve the extraction of teeth, it is possible to extend the colour-based analysis to account for overall brightness changes, have different models for each possible lighting condition, or use adaptive modelling where the probability density functions are updated with each frame [OPB97].

5.5.3 Eyebrow actions

The eyebrow raise is controlled by one large muscle in the scalp and forehead. This muscle, shown in Figure 5.13, runs vertically from the top of the head to the eyebrows and covers virtually the entire forehead. The medial or inner portion of this muscle can act separately from its lateral or outer portion. The inner eyebrow raise (AU1) pulls the inner corners of the brow and center of the forehead upwards. The outer eyebrow raise (AU2) pulls the outer corner of the brow upwards. The combination (AU1 + 2) is more common than either AUs, mainly because raising the inner corners of the eyebrows is a difficult movement for most people to make voluntarily without adding AU2.

An eyebrow raise is computed using the increase in distance between the inner eye corner and inner eyebrow point for both the left and right eyebrows, measured with

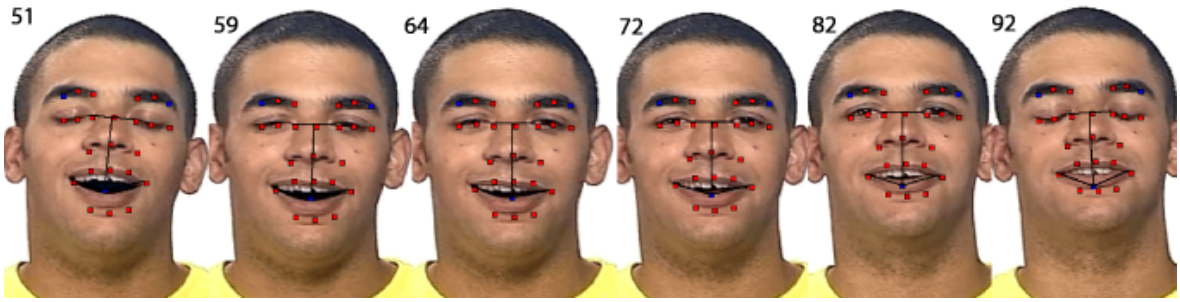


(a) Selected frames between 1 and 50 of the localized mouth polygon.

(b) Aperture mask (aperture pixels are highlighted in white).

(c) Teeth mask (teeth pixels are highlighted in white).

Figure 5.11: Tracking the mouth aperture and teeth in a video of the mental state *comprehending* from of the Mind Reading DVD. These frames are classified as mouth-open.



(a) Selected frames between 51 and 100 of the localized mouth polygon.

(b) Aperture mask. Aperture pixels are highlighted in white.

(c) Teeth mask. Teeth pixels are highlighted in white.

Figure 5.12: Continuation of Figure 5.11, showing the tracking of the mouth aperture and teeth. Note that there is no neutral expression in the transition between the AUs in Figure 5.11 and this one. These frames are classified as teeth-present.

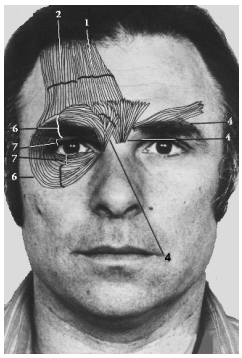


Figure 5.13: The eyebrow raise is controlled by one large muscle in the scalp and forehead. From Ekman and Friesan [EF78].

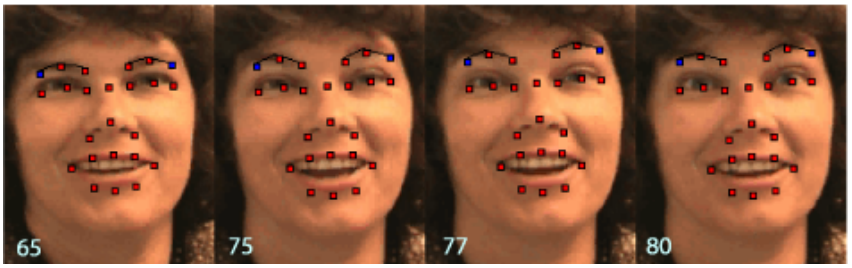


Figure 5.14: Selected frames showing an eyebrow raise ($\mathrm{AU}1+2$) from a video labelled as *agreeing* from the CVPR 2004 corpus. The eyebrow raise is described by the increase in distance between the inner eye corner and inner eyebrow point for both the left and right eyebrows.

respect to the distance on the initial frame. This calculation, though robust to head rotation, is not invariant to scale variations. An example of frames that are classified as an eyebrow raise is shown in Figure 5.14.

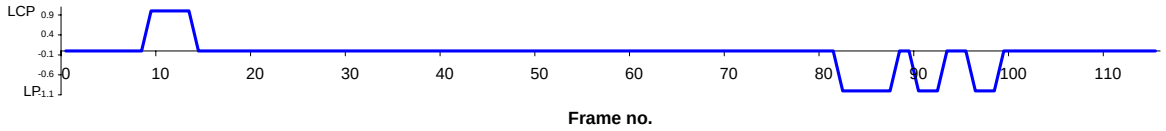
The frown ($\mathrm{AU}4$) posed a particular challenge. While correctly identifying this AU would have certainly improved recognition results for mental states such as *thinking*, the drawing-in movement of the eyebrows is too subtle to pick from feature point movements alone. Gradient analysis of the wrinkles in the forehead formed by a frown could be used to detect this action, such as in Lien *et al.* [LZCK98].

5.5.4 Asymmetry in facial actions

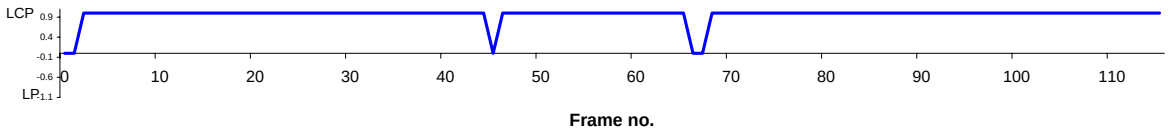
The detection of asymmetry can be useful in the inference of complex mental states since asymmetric facial actions occur frequently in expressions of cognitive mental states. In particular, an asymmetric eyebrow raise and asymmetric lip pull are frequent in *confusion*, while asymmetric lip bites are frequent in expressions of *worry* [RC03]. Recently, Mita and Liu [ML04] have shown that facial asymmetry has sufficient discriminating power to improve the performance of an expression classifier significantly. With a few exceptions, very little research has gone into the automated recognition of asymmetric facial actions. Fasel and Luettin [FL00] use image differences to identify asymmetry between the left and right face regions. Their approach however, has two limitations: it is not robust to rigid head motion, and it does not provide sufficient detail on the specific features where asymmetry occurs.

The shape and colour-based methods that have been described so far can be extended to describe facial action asymmetry. Instead of the one action symbol output per feature, two sets of symbols are output per feature: one for the left and one for the right region of the face. This is straightforward for the eyebrows, and equally so for the lips because the use of the anchor point to measure shape deformations means that two different measurements are obtained one for each half of the face. For the mouth, the line passing through the upper and lower lip centers defines the left and right polygons. All the shape and colour parameters are normalized against head turns to remove the effect of one region falsely appearing smaller than the other. The variance in motion, shape and colour between the two regions indicates asymmetry. Figure 5.15 shows the output symbols of the left and right lip shape actions. A lip pull is evident in the right lip corner point, as opposed to none in the left one.

(a) Selected frames from a video of the mental state *amused* from the Mind Reading DVD. The frames show an asymmetric (right) lip corner pull.



(b) The output symbol sequence for the left lip corner actions. LCP=lip corner pull, Lp=lip pucker.



(c) The output symbol sequence for the right lip corner actions. LCP=lip corner pull, Lp=lip pucker. Contrast that to the left lip corner actions.

Figure 5.15: Symbol sequence for the left and right lip corners showing an asymmetric lip corner pull action.

5.6 Discussion

The output at this level of the system is a sequence of symbols along each of the head rotation axes (pitch, yaw and roll) and facial actions (lip, mouth and eyebrows). Figure 5.16 illustrates the output for a video of the mental state *comprehending* from the Mind Reading DVD. As shown in the figure, this video has the following actions: head up, head down, head turn-left, head-turn-right, lip corner pull, eyebrow raise, mouth open and teeth-present actions. The action sequences represent a head nod, a smile and an eyebrow flash. The objective of the following chapter is to analyse the action sequences in order to recognize these displays.

Recall from Chapter 2 that FER systems have, so far, only been tested with carefully pre-segmented facial expression sequences that start and end with a neutral frame. Compared with these sequences, the videos on the Mind Reading DVD, such as that shown in Figure 5.16, are challenging for a number of reasons. First, the different displays are asynchronous; they overlap but start and end at different times. Second, the transition from one expression to another does not necessarily involve a neutral state. Third, head actions along the three rotation axes often co-occur. For instance, in the video in Figure 5.16 the head up and head down actions also involve a head tilt orientation.

The approach I have adopted builds on state-of-the-art facial expression analysis based on feature point tracking. It has several key characteristics that make the automated mind-reading system suited for use in HCI contexts. First, being computationally efficient, action extraction runs in real time. Second, the system runs without the need for any manual initialization, calibration or pre-processing. These two characteristics are a prerequisite of HCI systems. Third, facial action analysis is resilient to the substantial rigid head motion that occurs in natural human motion. Finally, the system

does not need to be calibrated or trained with every new user, an important factor in designing ubiquitous systems.

The main limitation of this feature-based approach, however, is that with each new action added to the system, a model or rule needs to be hand-coded. In addition, given that the principal focus of this dissertation is the recognition and inference of complex mental states, rather than the recognition of actions, the methods described in this chapter have not been optimized for variations in recording conditions such as illumination or for speed.

5.7 Summary

In this chapter I have described an implementation of the automated extraction of head and facial actions from video. Actions are essentially a temporal abstraction of FACS AUs and constitute the bottom-level of the computational model of mind-reading. A 2D face model is defined as a set of feature points that are located and tracked by `FaceTracker` over successive frames in a real time or recorded video stream. Head AUs are described by the magnitude and direction of three rotation angles. Motion, shape and colour analysis of the lips, mouth and eyebrow actions identify facial AUs of each of these face-components respectively. Action recognition runs in real time, without the need for any manual initialization, calibration or pre-processing, and in the presence of rigid head motion that occurs in natural human motion.

The output at this level of the system is a temporal sequence of head and facial actions. Together, these actions form the input to HMM classifiers of head and facial displays.

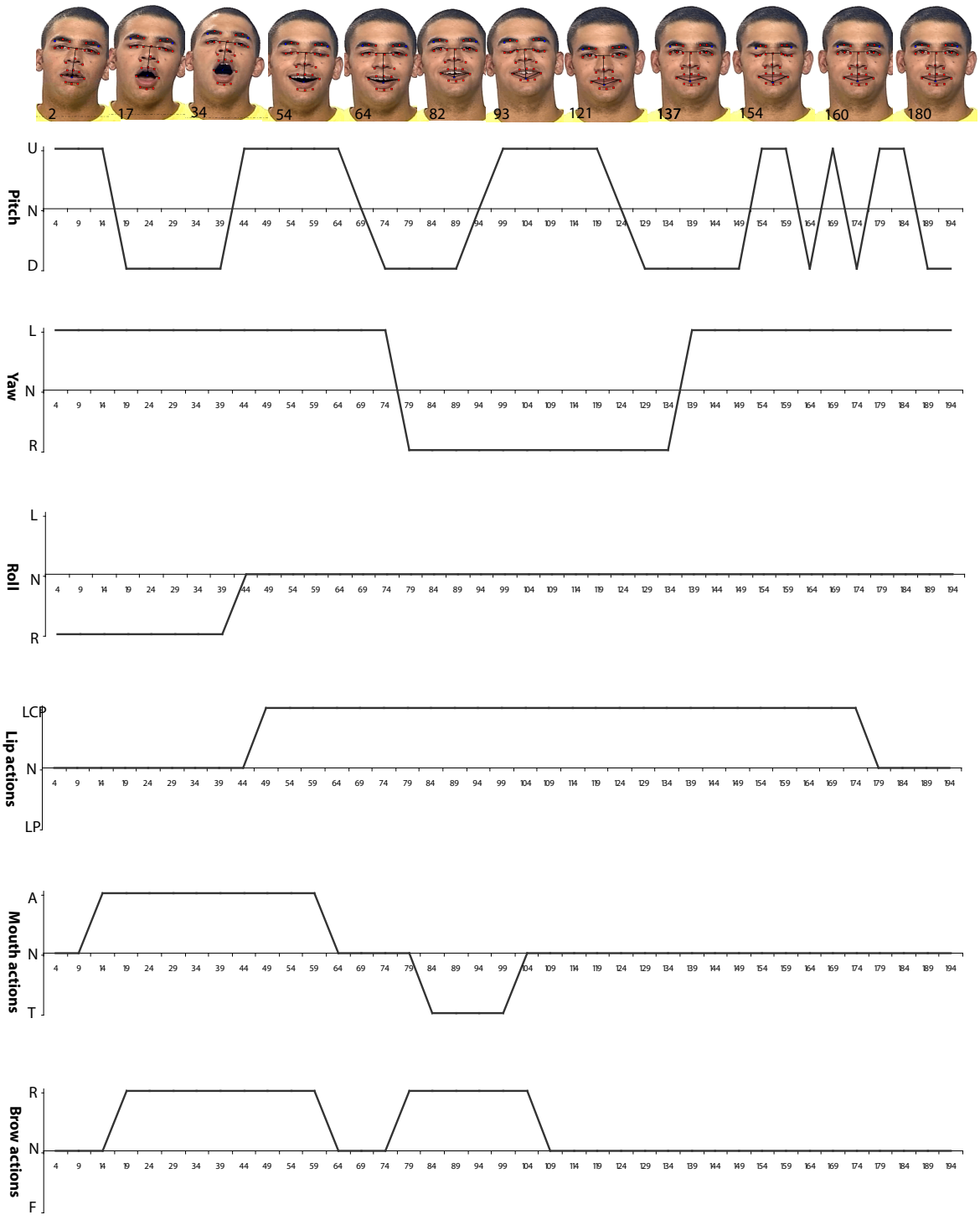


Figure 5.16: The output head and facial actions across a video showing *comprehending* from the Mind Reading DVD. From top to bottom: pitch actions (U=up, N=null, D=down), Yaw actions (L=left, N=null, R=right), Roll actions (L=left, N=null, R=right), lip actions (LCP=lip corner pull, N=null, LP=lip pucker), mouth action (A=aperture, N=null, T=teeth), brow actions (R=raise, N=null, F=frown). The plots show that this video has the following asynchronous displays: head nod, smile and eyebrow flash. The recognition of these displays is discussed in the next chapter.

Chapter 6

Recognition of Head and Facial Displays

In Chapter 3, I have shown the importance of considering inter-expression as well as within-expression facial dynamics in mental state recognition. This chapter describes how consecutive actions are analysed spatio-temporally to recognize high-level, communicative, head and facial displays. I demonstrate the suitability of Hidden Markov Models (HMMs) in representing the dynamics of displays and describe a classification framework that enables their recognition very soon after their onset, so that there is no marked delay between a user's expressions and the system recognizing it. Experimental results show reliable, real time recognition of displays in a range of complex mental states.

6.1 The dynamics of displays

Displays are defined as head or facial events that have meaning potential in the contexts of communication [Bir70]. They are the logical unit that people use to describe facial expressions and to link these expressions to mental states. For instance, when interacting with a person whose head movement alternates between a head-up and a head-down action, most people would abstract this movement into a single event—a head nod—and would presume that person is in agreement, comprehending or attending.

In the computational model of mind-reading, displays serve as an intermediate step between tracked AUs and inferred mental states. The input to this level is a running sequence of head and facial actions $\mathbf{Z}[1 : t]$, that have been extracted as described in Chapter 5. Table 6.1 lists the nine head and facial displays currently supported by the automated mind-reading system and their underlying actions. It is important to distinguish displays, which are the classes in this recognition task, from their component AUs, which are the features used for classification, both of which might be described using the same term. For instance, a tilt orientation *display* is a sequence of head tilt *actions* (AU55 and/or AU56).

This level of the automated mind-reading system has two objectives. The first is to analyse action sequences and classify them into one of the display classes in Table 6.1. The second is to do that without requiring any manual intervention and in real time.

Table 6.1: List of head and facial displays and their component AUs.

Display	Description	Dynamics
Head nod	Alternating head up (AU53) and head down (AU54) actions	Periodic
Head shake	Alternating head turn left (AU51) and head turn right (AU52)	
Tilt orientation	Persistent tilt in one direction (sequence of AU55 or AU56)	Episodic
Turn orientation	Persistent pose of turned head (sequence of AU51 or AU52)	
Lip corner pull	Onset, apex, offset of lip corner pull or lip stretch (AU12 or AU15 or AU12 + 25)	
Lip pucker	Onset, apex, offset of lip pucker (AU18)	
Mouth open	Onset, apex, offset of mouth open (AU26 or AU27)	
Teeth present	Onset, apex, offset of mouth stretch (e.g., AU12 + 25)	
Eyebrow flash	Onset, apex, offset of eyebrow raise (AU1 + 2)	

Displays differ in total duration, overall intensity (the maximum motion or total amount of kinetic energy incurred during the display), and in the time scale of the underlying movements. Such variations often signify different user intents. For instance, a strong head nod indicates more agreement than a weaker or slower one. Despite these variations, displays follow a pattern of temporal regularity that can be exploited when analyzing these displays [Bir70, DV01, SJ97]. By modelling the temporal progression of actions across an image sequence, one can infer the underlying display. The displays in Table 6.1 are grouped by their dynamic properties as either periodic or episodic.

6.1.1 Periodic displays

Periodic motion such as that seen in a head nod or a head shake recurs at regular intervals. Figure 6.1 shows the temporal structure of a natural head shake, which is characterized by alternating head-turn-right, head-turn-left motion cycles, and variations thereof. The cycle time is the interval of time during which a sequence of recurring motions is completed. In Figure 6.1, the head-turn-right, head-turn-left cycle is repeated seven times. Each of the cycles lasts for a different amount of time. The action sequence describing periodic displays may therefore vary in length, in the ordering of the actions, and in the duration of each of the actions.

6.1.2 Episodic displays

The rest of the head and facial displays listed in Table 6.1 are grouped as episodic displays. The dynamic properties of episodic displays are characterized by three stages: a fast beginning, or **onset** of the display, contrasted by a period of standstill at the point of maximum intensity or **apex**, followed by the slow disappearance, or **offset** of the display [GSS$^+$88, SC01, YD96]. The head orientation displays such as the tilt and turn, are episodic in that they are characterized by the unidirectional rotation of the head along the roll and yaw axes. Langton *et al.* [LWB00] explain how the orientation of the head signals social attention, while Baron-Cohen [Bar94] highlights the role of the head orientation as an indicator of cognitive mental states such as *thinking*.

The lip displays include the lip-corner pull display as in a smile and the lip pucker display. Figure 6.2 shows the dynamics of two “episodes” of a smile (lip corner pull) from

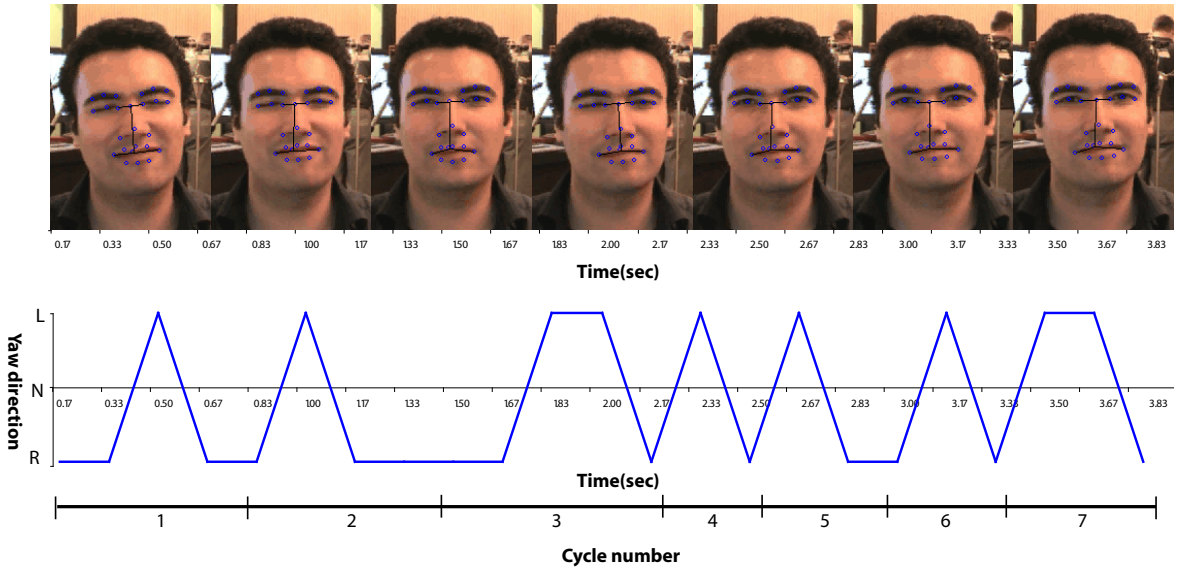


Figure 6.1: The dynamics of a periodic display: (top) selected frames from a video labelled as *disagreeing* from the CVPR 2004 corpus; (bottom) a plot of the head-yaw actions (L=left, N=null, R=right) that were extracted by the action recognition system. The symbol sequence is {R,R,L,R,R,L,R,R,R,R,L,L,R,L,R,L,R,R,L,R,L,L,R}, where a symbol is output every 166 ms. The symbols convey a head shake, which is characterized by 7 cycles of alternating head-turn-right, head-turn-left motions. Note the subtlety of the head shake.

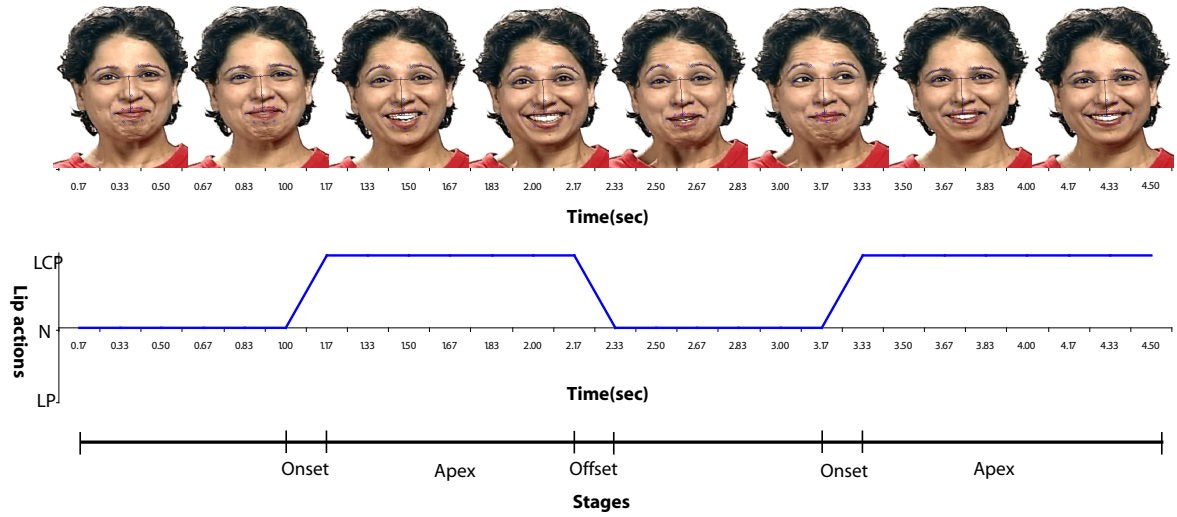


Figure 6.2: The dynamics of an episodic display: (top) selected frames from a video labelled as from a video labelled as *lucky* from the Mind Reading DVD; (bottom) a plot of the lip corner pull actions (LCP=lip corner pull, N=null, LP=lip pucker) that were extracted by the action recognition system. The symbol sequence is {N,N,N,N,N,N,LCP,LCP,LCP,LCP,LCP,LCP,LCP,N,N,N,N,N,N,LCP,LCP, LCP,LCP,LCP,LCP,LCP,LCP}, where a symbol is output every 166 ms. The symbols convey two smile "episodes" which are characterized by at least one of an onset, apex and offset.

a video labelled with the mental state *lucky*. An eye-brow flash is defined as the onset, apex and offset of an eye-brow raise that communicates social greeting if it coincides with a smile, mutual agreement if it coincides with a head nod or head dip, surprise if it coincides with jaw drop, and interest if it co-occurs with a lip stretch [GSS+88].

Like periodic displays, the action sequences describing episodic displays vary in total length and in the duration of each of the actions. The classification methodology of choice has to take into account these variations in dynamics.

6.2 Representing displays as Hidden Markov Models

To account for, and indeed exploit, the dynamics of periodic and episodic displays, the system employs HMMs for the classification of temporal sequences of actions into a corresponding head or facial display. The nine displays listed in Table 6.1 are each implemented as a separate model.

6.2.1 Hidden Markov Models

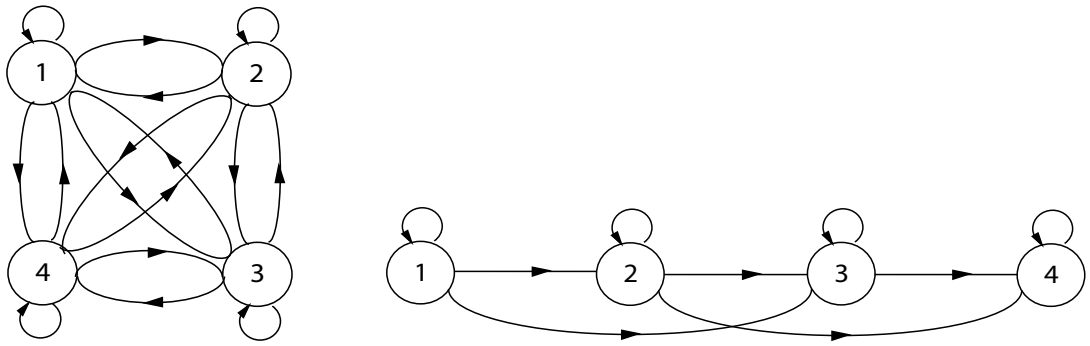
In contrast to static classifiers that classify single frames into an emotion class, dynamic classifiers model the temporal information inherent in facial events. HMMs are one of the basic (and perhaps best known) probabilistic tools used for time series modelling. A comprehensive review of HMMs can be found in Rabiner's article [Rab89]. HMMs have been successfully used in a number of applications including speech recognition [Rab89], handwriting recognition [PS00], head gesture recognition [KP01, MYD96, TR03], and automated facial expression recognition to classify facial AUs [BLB+03, LZCK98, OPB97] and basic emotions [CSGH02, CSC+03b, CSC+03a].

HMMs are particularly suited for the recognition of head and facial displays from action sequences. They provide a sound probabilistic framework for modelling time-varying sequences, and the convergence of recognition computation runs in real time, which is essential to FER systems for HCI.

6.2.2 Representation

An HMM is a generative model that represents the statistical behaviour of an observable symbol sequence in terms of a network of hidden states. The discrete states are assumed to be temporally connected in a Markov chain, that is, future states depend on the present state but are independent of the past. For each observable symbol, the process being modelled occupies one of the states of the HMM. With each symbol, the HMM either stays in the same state or moves to another state based on a set of state transition probabilities associated with that state. The complete parameter set λ_j = (π,A,B) for a discrete HMM of display j where $1 \leq j \leq y$, can be described in terms of:

- **N**, the number of states in the model $S = \{S_1, \ldots, S_N\}$. Though hidden, there is often physical significance attached to each of the states. The state at time t is denoted as q_t.
- **M**, the number of different observation symbols, and corresponds to the physical output of the process being modelled. The symbols are denoted as $V = \{v_1, \ldots, v_M\}$
- **A**= $\{a_{ij}\}$, an N x N matrix that specifies the probability that the model's state will change from state i to state j, where $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$, $1 \leq i, j \leq N$



(a) 4-state erogdic HMM

(b) 4-state left-right HMM with a maximum of two jumps between states.

Figure 6.3: Choice of Hidden Markov Model

- **B**= $\{b_i(k)\}$, an N x M matrix, the observation symbol probability matrix depicts the output observation given that the HMM is in a particular state i, where $b_i(k) = P(v_k|q_t = S_i)$, $1 \leq i \leq N$ and $1 \leq j \leq M$
- π= $\{\pi_i\}$ is an N-element vector that indicates the probability of initially being in state i, where $pi_i = P(q_0 = S_i)$ $1 \leq i \leq N$

To determine the parameter set λ_j for display j, one has to decide on the type of HMM model to adopt, define the topology, that is, select N and M, and compute the vector π and matrices A and B.

6.2.3 Choice of Hidden Markov Models and topology

The Ergodic model [Rab89] and the left-right or Bakis model [Bak91], are the two most popular types of HMMs. A review of other HMM variants can be found in Rabiner [Rab89]. In the ergodic, or fully connected HMM, every state can be reached in one single step from every other state in the model (Figure 6.3(a)). In the left-right or Bakis model, the state sequence must begin from the left at state 1 and end on the right at the final state N. As time increases, the observable symbols in each sequence either stay at the same state or increase in a progressive manner. An example of a 4-state left-right model is shown in Figure 6.3(b). Ergodic HMMs subsume Bakis models, but Bakis are generally more efficient as they require less parameters. For small HMMs, efficiency is not a concern, so I use the ergodic model.

Within a particular model, one has to pick a suitable HMM topology by defining N and M. There is no simple theoretical method for determining the optimum topology for an HMM [Rab89]. Although essentially a trial-and-error process, the number of states is estimated by considering the complexity of the various patterns that the HMMs need to distinguish. For example, to represent a head nod, I use its underlying dynamics to estimate how many distinguishable segments the display contains. Since a head nod consists of two motion types: head up and head down motions that recur throughout the display, each motion was mapped to a single state, resulting in a 2-state ergodic HMM.

The observable symbols or features of an HMM should be as simple as possible to allow fast computation. They must also be sufficiently detailed to indicate differences between patterns. The number of symbols is determined by the number of possible actions the action extraction level is able to identify for each display. For example, the action extraction system is able to distinguish among three actions along the pitch axis: head-up, head-down and null. Accordingly, the feature space of the head nod HMM consists of these three symbols.

6.3 Training

Training attempts to estimate the model parameters λ_j in order to maximize the probability of an observation sequence of actions $\mathbf{Z}[1 : t]$ given the model. The training algorithm is data-driven: for each HMM representing a display, the input is a set of example action sequences of that display. The set of training sequences must be chosen to span the number of ways different people express a particular display. To train the HMMs, I compiled between 20 and 120 examples of action sequences for each display, which were extracted from videos from the Mind Reading DVD. Note that any videos used throughout the training process were *not* included in the test set. Even though the process is linear in the length of the observation sequence t, the entire training activity is currently done off-line.

The training process associates each state of the HMM with a particular region of the feature space using the iterative Baum-Welch algorithm, also known as the forward-backward algorithm [Rab89]. In summary, the algorithm runs forward and backward through the observed output of each training example using the actual transitions to successively refine an initial estimate of the HMM parameters λ_j. This is repeated until convergence, that is, until there is no significant change in λ_j compared to previous iterations. The converged models provide an insight into how each display is represented in the system. Figure 6.4 summarizes the transition probabilities, observation probability distribution, and initial probability distributions for each display.

6.3.1 HMMs of periodic displays

As one might expect, the HMMs of periodic displays encode a cyclic transition between the states. That is, there is a higher probability of transition than recurrence among the states of the HMM. Each state corresponds to a motion segment of the cycle.

Figure 6.4 (a) shows the result of training a 2-state ergodic HMM with three observable symbols on 37 examples of head nods picked from the Mind Reading DVD. The symbols are null, head-up, and head-down. The first state of the HMM corresponds to a head-down segment of the nod cycle, while the second corresponds to a head-up segment. The transition probabilities encode a cyclic transition between the two states as is typical of a head nod display. In this HMM, it is more likely that the head nod starts with a head-up than head-down movement.

Similarly, Figure 6.4 (b) shows the result of training a 2-state ergodic HMM with three observable symbols (null, head-turn-left, head-turn-right) on 30 examples of head shakes. The first state of the head shake HMM corresponds to a head-turn-right segment of the shake cycle, while the second state corresponds to a head-turn-left segment. As in the head nod, the transition probabilities represent a cyclic motion. In this HMM, it is almost equally likely that the head shake starts with a head-turn-left or head-turn-right movement.

6.3.2 HMMs of episodic displays

While HMMs of periodic displays encode a high probability of transition between states, the HMMs of episodic displays are characterized by a high probability of recurrence within states.

The head tilt and turn orientation displays are represented as a 2-state, ergodic HMM with 7 symbols each, to account for the intensity of these displays. Figure 6.4 (c) shows

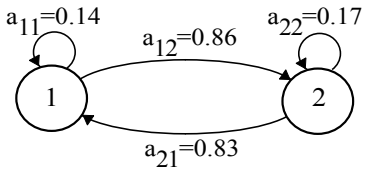
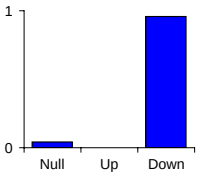
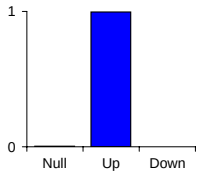
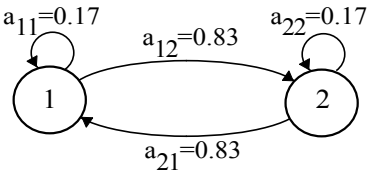
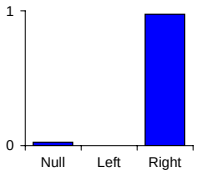
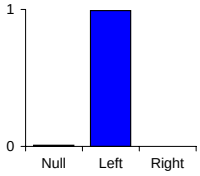
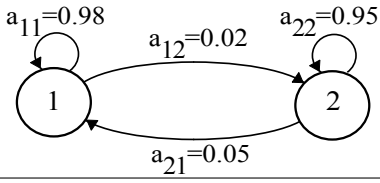
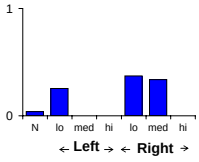
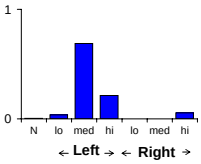
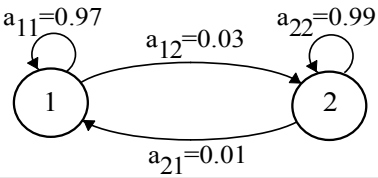
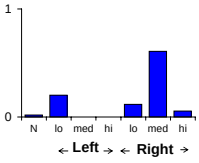
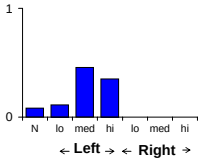
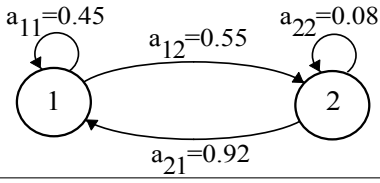
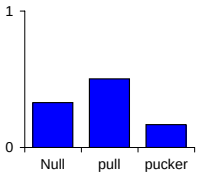
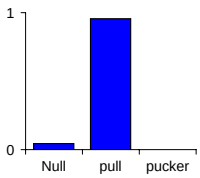
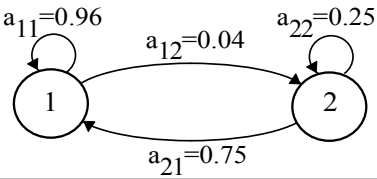
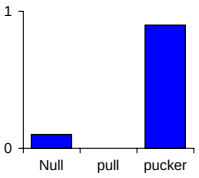
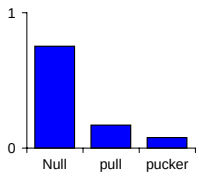
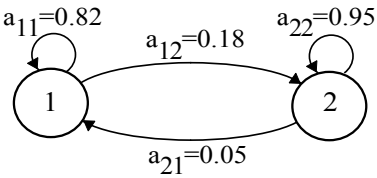
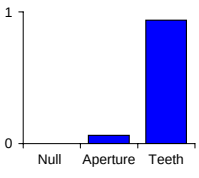
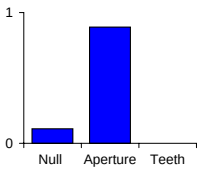
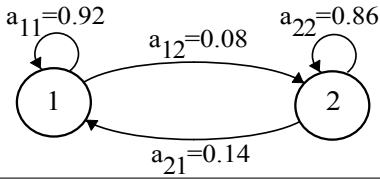
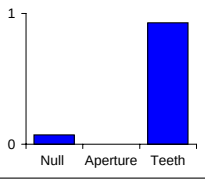
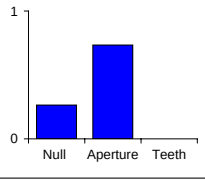
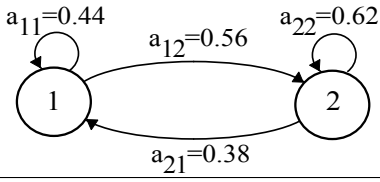
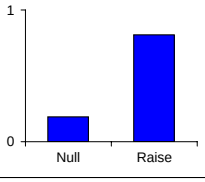
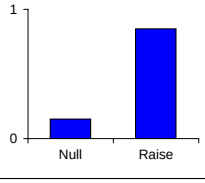
Display	Transition	State 1	State 2	Prior
(a) Head nod				$\begin{bmatrix}0.45\\0.55\end{bmatrix}$
(b) Head shake				$\begin{bmatrix}0.53\\0.47\end{bmatrix}$
(c) Tilt				$\begin{bmatrix}0.73\\0.27\end{bmatrix}$
(d) Turn				$\begin{bmatrix}0.32\\0.68\end{bmatrix}$
(e) Lip pull				$\begin{bmatrix}0.41\\0.59\end{bmatrix}$
(f) Pucker				$\begin{bmatrix}0.72\\0.28\end{bmatrix}$
(g) Mouth open				$\begin{bmatrix}0.07\\0.93\end{bmatrix}$
(h) Teeth-present				$\begin{bmatrix}0.95\\0.05\end{bmatrix}$
(i) Eyebrow flash				$\begin{bmatrix}0.45\\0.55\end{bmatrix}$

Figure 6.4: The parameters of the nine head/ facial display HMMs: the transition probabilities, the observation probability distributions per state, and the initial state probabilities.

the result of training a 2-state ergodic HMM with seven observable symbols on 93 examples of head tilts. The symbols represent a left and right head tilt with three levels of intensities each: low, medium and high. The first state of the HMM corresponds to a tilt-right orientation; the second corresponds to a tilt-left orientation. As expected in an episodic display, the states persist.

Along the same lines, Figure 6.4 (d) shows the result of training a 2-state ergodic HMM with seven observable symbols on 126 examples of head turns. The symbols represent a left or right head turn with three levels of intensity: low, medium and high. The first state of the HMM corresponds to a turn-right orientation, while the second state corresponds to a turn-left orientation. As depicted by the initial state probabilities, the turn-left orientation display is more likely to occur than the right one.

The lip and mouth displays are each represented as a 2-state, 3-symbol ergodic HMM trained using between 20 and 50 examples. In the lip-corner pull display (Figure 6.4 (e)), the first state corresponds to the onset of the display; the second state corresponds to a lip corner pull action. The HMM encodes a high probability of transition from the first to the second state, a high probability of recurrence of the second state. This corresponds to the onset and apex phases of a smile display. As depicted by the initial state probabilities, both states are equally likely to occur. Similarly, in the lip pucker display (Figure 6.4 (f)), the first state corresponds to a lip pucker, while the second state corresponds to a neutral state. The state transition probabilities represent the two phases typical of a pucker display, which are the transition from the second to the first state and the recurrence of the first state.

Figure 6.4(g) shows the mouth-open display HMM represented as a 2-state ergodic HMM with three observable symbols (null, aperture, teeth). The first state corresponds to teeth, while the second corresponds to aperture. The HMM shows a high occurrence of aperture actions. Figure 6.4 (h) shows the teeth displays as a 2-state ergodic HMM with three observable symbols (null, aperture, teeth). The HMM encodes a high occurrence of teeth actions. The eyebrow flash (Figure 6.4 (i)) is represented as a 2-state ergodic HMM with two observable symbols (null, raise eyebrow). The HMM encodes the onset and apex of an eyebrow flash.

6.4 Classification framework

Once the parameters are estimated for each HMM—a one time off-line process—the HMMs can act as classifiers for the online recognition of displays. The problem of classification can be stated as follows: given an HMM model λ = (π,A,B) and a running sequence of head and facial actions $\mathbf{Z}[1:t]$, the objective is to find the probability that the observations are generated by the model $P(\mathbf{Z}[1:t]|\lambda)$. This is computed using the forward-backward algorithm. The output of each classifier is a probability that the vector was generated by that HMM.

Figure 6.5 shows a procedural description of how real time display recognition proceeds in a video of arbitrary length. The classification framework is implemented as a sliding window of six actions that progresses one action at a time. Recall from Chapter 5 that an action spans 166 ms or five frames at 30 fps. Since the observation vector moves one action at a time, the HMM classifiers are invoked every 166 ms. The first invocation is an exception: it occurs when six consecutive actions are available.

The output probabilities of each of the HMMs over time constitute the input observation vector of the mental state level of the mind-reading system (Chapter 7). If a *soft decision*

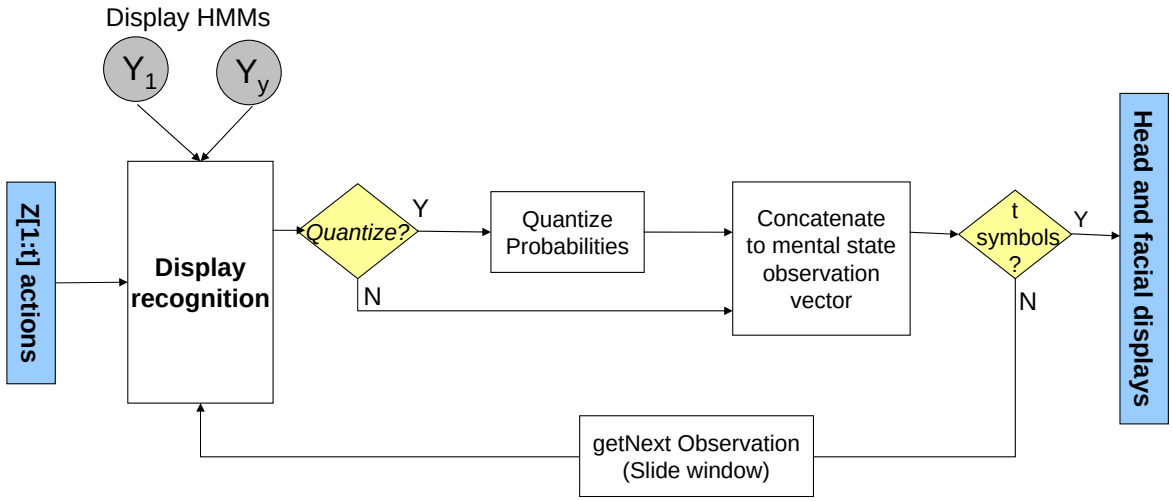


Figure 6.5: Procedural description of head and facial display recognition. The input is a vector of sequences of consecutive head and facial actions $\mathbf{Z}[1:t]$. Each sequence $Z_k[1:t]$ is input to a corresponding HMM classifier that outputs a probability that the observation vector was generated by the HMM model. After being quantized, the results are appended to the observation vector of the mental state classifiers.

approach is adopted, the output probability is directly appended to the observation vector. Figure 6.6 shows several snapshots from `AutoMR` over the course of a video of *encouraging* from the Mind Reading DVD. The video has the following co-occurring, asynchronous displays: a head nod and a lip corner pull and teeth that correspond to a smile. The vertical bars represent the output probabilities of the HMMs.

Alternatively, in a *hard decision* approach, the probability of each display is quantized to zero or one depending on a likelihood threshold before it is appended to the observation vector. Figure 6.7 illustrates the relationship between an observed action sequence, and the corresponding quantized output of the HMM (null, onset, apex). Using this classification framework, displays are recognized within approximately 166 ms of the occurrence of a corresponding action sequence, and well within its total duration.

Even though quantizing the output probabilities of the HMMs results in a loss of detail, training and inference of mental state classifiers that use these quantized probabilities are more efficient [ZGPB+04, LBF+04a]. In the next section, I evaluate the recognition accuracy for each of the displays at different likelihood thresholds.

6.5 Experimental evaluation

This level of the system has been tested live by a public audience on two occasions. The first was at the Royal Institution of Great Britain (Friday Evening Discourse Series), London, UK. The second occasion was at a demonstration at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington D.C., US. On both occasions the system ran successfully, detecting the head and facial displays of a diverse audience in real time as they performed the displays. Interactions from the two demonstrations were not recorded, and so were not evaluated quantitatively. Instead, I conducted the experimental evaluation using a subset of the videos on the Mind Reading DVD.

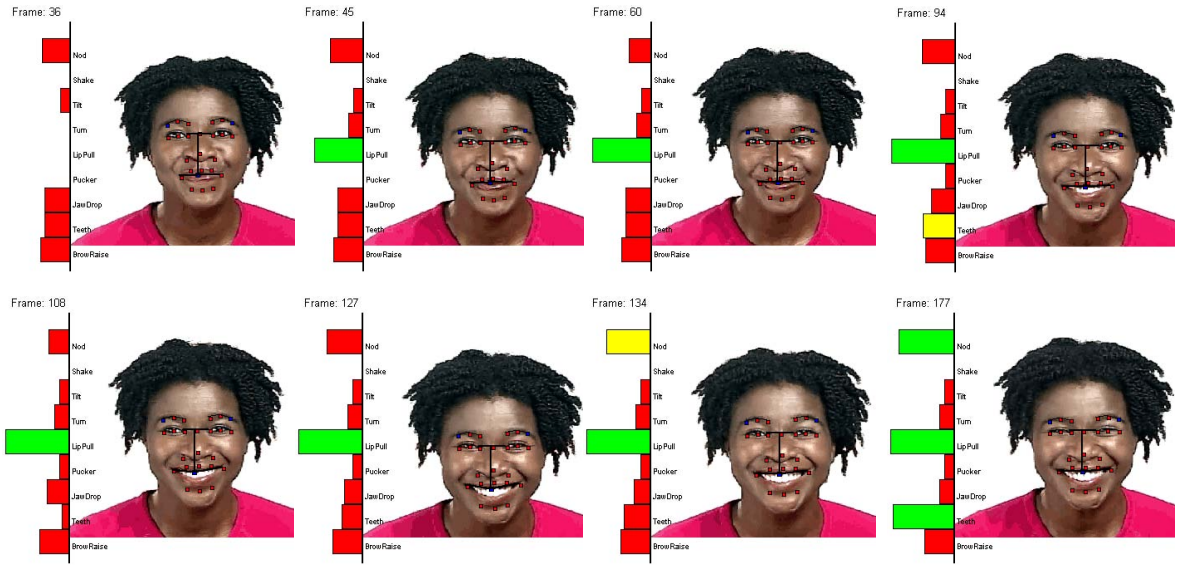


Figure 6.6: Snapshots from display recognition for a video of *encouraging* from the Mind Reading DVD. The histograms on the left of each frame show the output probability of each of the display HMMs. The bars are colour-coded to show the different stages of that display: null ■ if the probability is less than 0.3, onset ■ if the probability is between 0.3 and 0.6 or apex ■ if the probability is greater than 0.6.

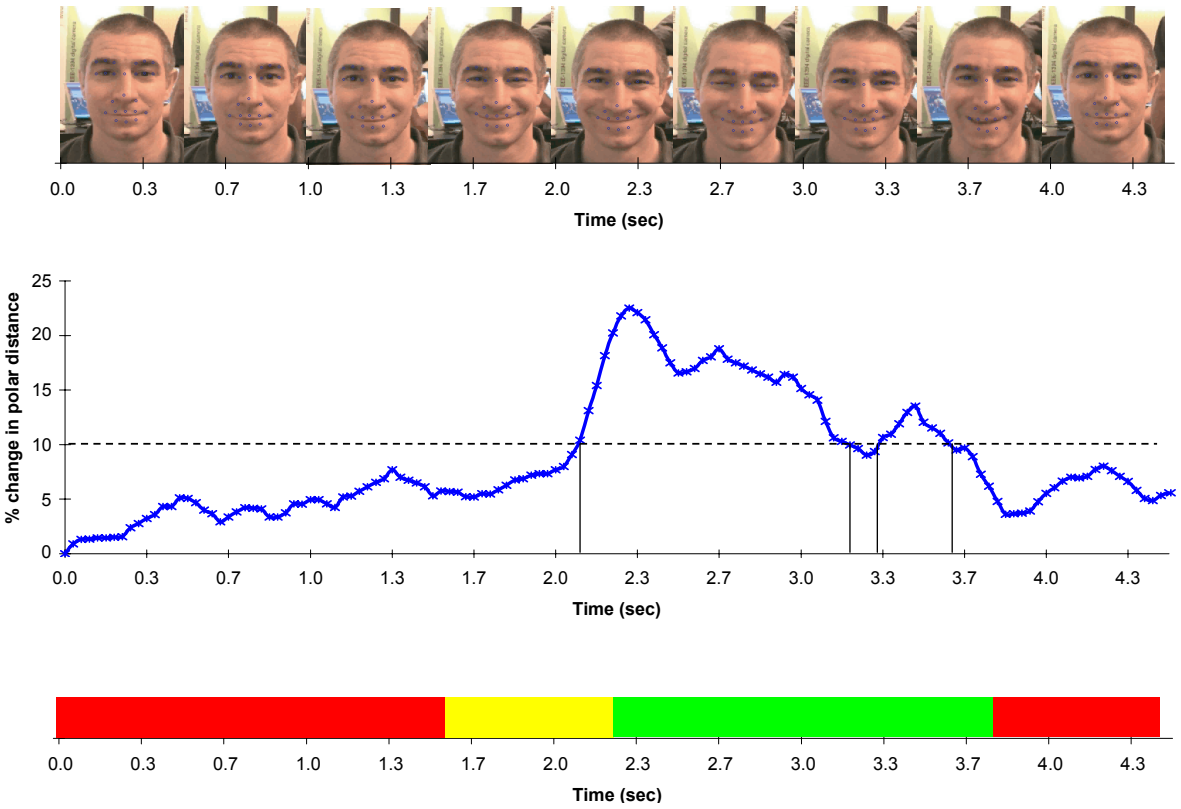


Figure 6.7: Display recognition in the case of a smile. Top: selected frames from a 4.5 second video showing *agreeing* from the CVPR 2004 corpus. Middle: change in polar distance in real time. The lip-pull threshold is shown as a dotted line. Bottom: the output of the lip-pull HMM quantized to three levels that represent the different stages of that display: null ■ if the probability is less than 0.3, onset ■ if the probability is between 0.3 and 0.6 or apex ■ if the probability is greater than 0.6. Note that the smile is recognized within approximately 166 ms of the occurrence of the corresponding action sequence, and well within its total duration.

6.5.1 Objectives

For a quantitative analysis of this level of the system, I evaluated the accuracy of recognition for the following nine displays: head nod, head shake, head tilt, head turn, lip-pull, lip pucker, mouth open, teeth present and eyebrow flash. This analysis has two objectives. The first objective is to determine the recognition rate and false positive rate of each of these displays as the likelihood threshold is varied. The threshold value that yields the best recognition results for each display is the one adopted at the next level of the system. The second objective is to gain an insight into the performance of the display classifiers, and determine the reasons for undetected or false detections.

A total of 174 videos from the Mind Reading DVD were chosen for the test. The videos represented the following six groups of complex mental states, and the 29 mental state concepts they encompass: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The videos were recorded at 30 fps with durations between five to eight seconds.

6.5.2 Results

Out of the 174 videos chosen for the test, ten were discarded because `FaceTracker` failed to locate the non-frontal face on the initial frames of the videos. Thus, the remaining 164 videos (25645 frames, 855 seconds) were used to generate the action sequences on which we tested the display classifiers. The 164 videos yielded a total of 6294 instances of head and facial displays in Table 6.1, and 30947 instances of non-displays.

Figure 6.8 shows the results of display recognition in a 6-second long video labelled as *undecided* from the Mind Reading DVD. The system correctly identifies a head shake, a head tilt, a head turn, a lip pucker and an eyebrow flash. A classification instance is correct if the likelihood of the HMM is above an empirically determined threshold ϵ, and the corresponding video contains that head or facial display (determined visually). The classification result for a display as the likelihood threshold is varied is shown using a Receiver Operator Characteristic (ROC) curve. ROC curves depict the relationship between the percentage of true positives (TP) and the percentage of false positives (FP).

The true positives or classification rate of display Y_j is computed as the ratio of correct detections to that of all occurrences of Y_j in the sampled videos. Correct detections are those that score above likelihood the threshold and match the ground truth. In addition, 1-TP can be thought of as a measure of **undetected displays**, that is, displays that do not meet the likelihood threshold, and are hence undetected by the system. The FP rate for Y_j is given by the ratio of samples falsely classified as j to that of all $\overline{Y_j}$ occurrences. FP is a measure of **falsely detected displays**: displays that meet the likelihood threshold, but are in fact not present in the video. Hence, the lower the FP rate the better. Figures 6.9-6.12 shows the ROC curves for each of the nine displays.

Table 6.2 lists the combination of TP and FP rate that is adopted for each display. The best accuracy is that of the lip pucker display, at a detection rate of 99.0% and a FP of only 0.9%. The worst is that of the teeth display, at a classification rate of 91.6% and high false positive rate of 14.4%. The overall average classification rate is 96.5% at an average false positive rate of 7.1%.

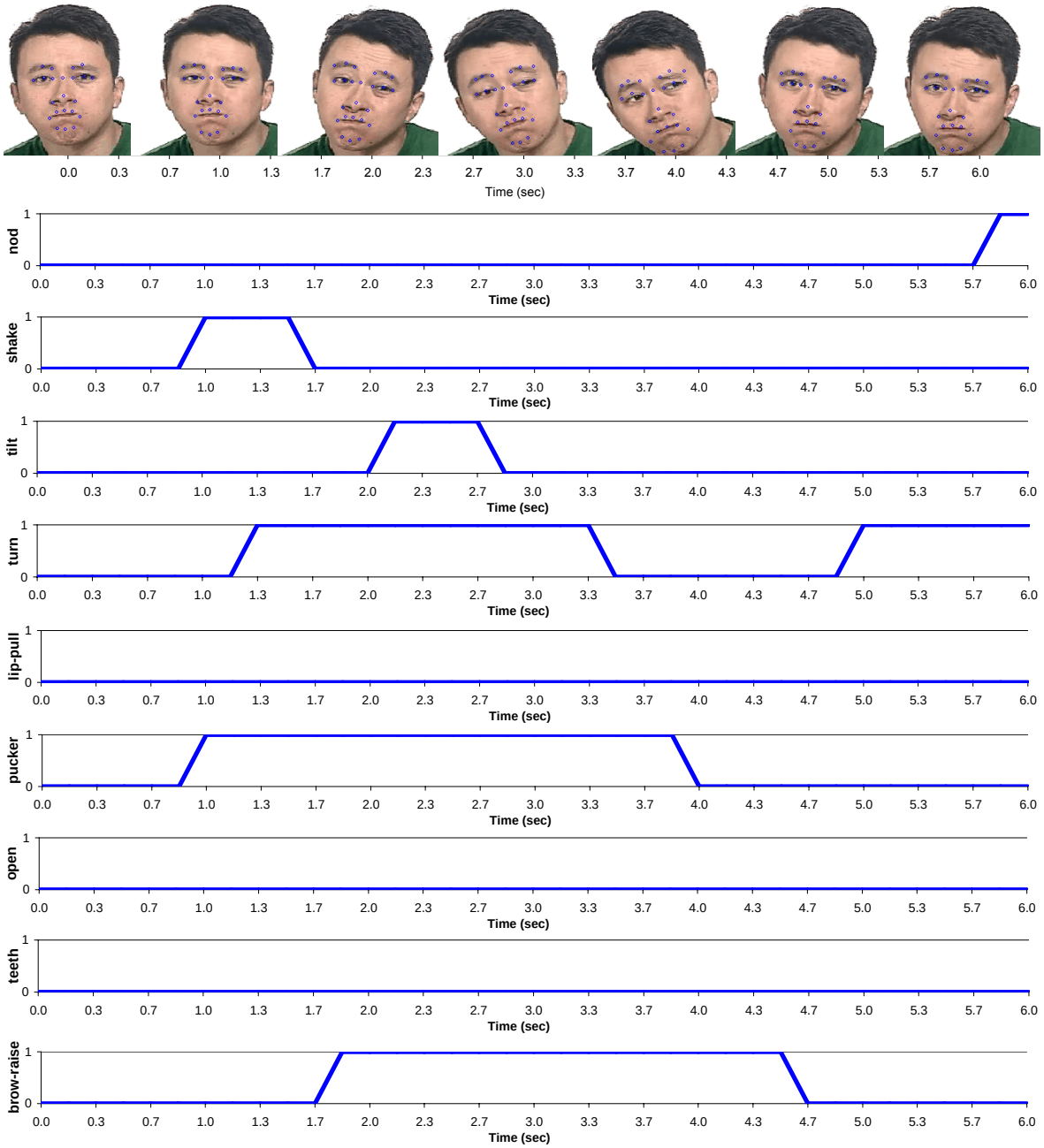


Figure 6.8: Trace of display recognition in a video labelled as *undecided* from the Mind Reading DVD. Row 1: selected frames from the video sampled every one second/ Rows [2-10]: head and facial displays. Throughout the video, a head shake, a head tilt, a head turn, a lip pucker and an eyebrow flash are observed.

Table 6.2: Adopted true positive (TP) and false positive (FP) rates for head and facial displays.

Display	#Train.	#Displays	#non-Displays	TP (%)	FP (%)
Head nod	37	338	3811	**98.0**	**7.8**
Head shake	30	411	3707	**97.1**	**8.1**
Head tilt	93	1581	2542	**96.5**	**4.6**
Head turn	126	1784	2358	**97.2**	**10.2**
Lip pull	20	844	3305	**95.7**	**7.1**
Lip pucker	49	312	3837	**99.0**	**0.9**
Mouth open	44	458	3673	**94.5**	**10.0**
Teeth-present	44	225	3906	**91.6**	**14.4**
Eyebrow raise	20	341	3808	**99.0**	**1.2**
Average	51	699	3438	**96.5**	**7.1**

6.5.3 Discussion

A closer analysis of the results explains why undetected and false positive displays occurred. The specific reasons for misclassifications for each display are summarized in Table 6.3 and are discussed in the sections that follow. These reasons fall into one of the following three cases:

1. **Error at the action recognition level:** An incorrect action sequence, which also persists, results in an incorrect output by the corresponding HMM classifier.
2. **Error at the HMM level:** Failure at the HMM level is mainly due to under-represented patterns of a particular display in the training examples.
3. **Noise in coding:** Some noise is introduced when coding the videos because of difficulties in determining the precise time of onset and offset of a display.

Head displays

The majority of undetected head nod and head shake displays were weak ones that were too slow or too subtle. For instance, a head nod is too slow if the rate of cyclic head-up, head-down motion is lower than the threshold of the system. It is too subtle if the maximum head pitch motion incurred throughout the display is lower than the threshold defined by the system. This was also the case for undetected head shakes. The false positive cases of head nods were head dips that were misidentified as the onset of a head nod. In the case of head shakes, the false positives were mostly the result of a combination of a head turn when the head pitched downward then upward. Also, the system currently does not make a distinction between a head turn and translation of the body sideways (although this type of body motion is uncommon).

The main reason for undetected head orientation displays was that the high intensity displays were under-represented at the HMM level. The converged models of both displays, presented in Figure 6.4, explains why this is the case. For instance, the observation function of the right and left tilt in Figure 6.4 (c) shows that high intensity tilts are not accounted for. The false positive rate of head orientation displays is lower than that reported for periodic head displays. These false positives occur due to errors introduced in coding the ground truth of the videos: it is difficult to exactly determine the onset and offset of a head orientation display.

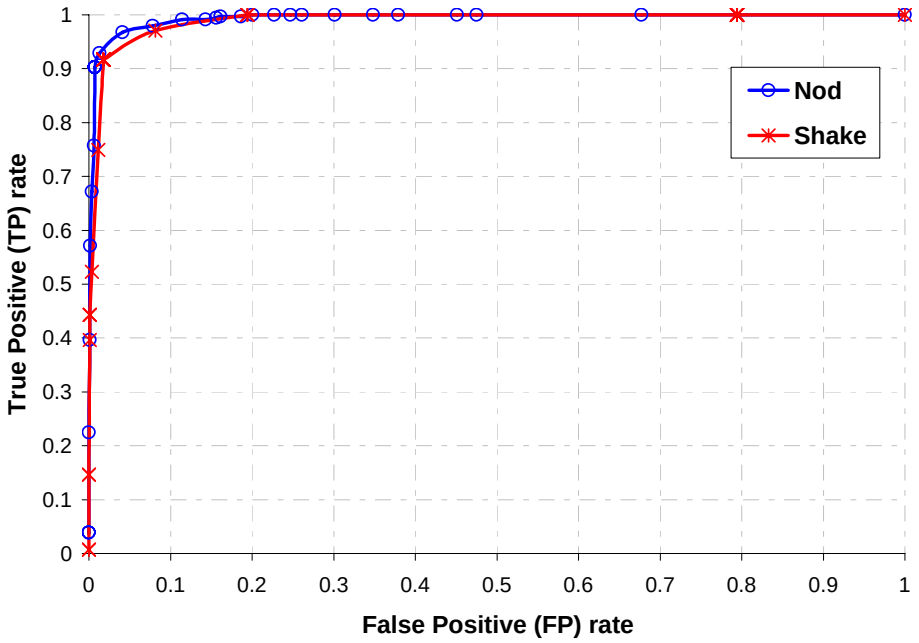


Figure 6.9: Head nod and shake ROC curves.

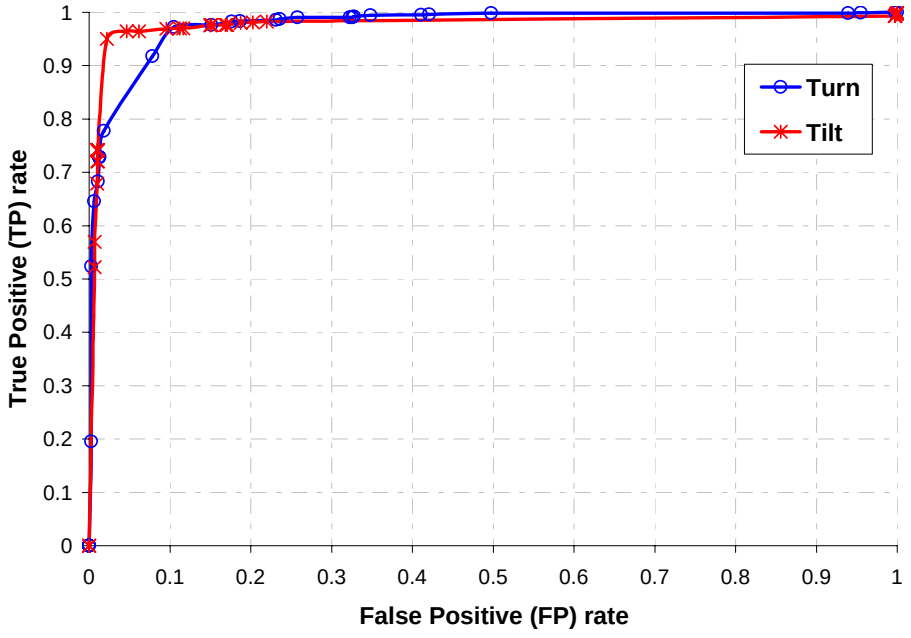


Figure 6.10: Head tilt and turn ROC curves.

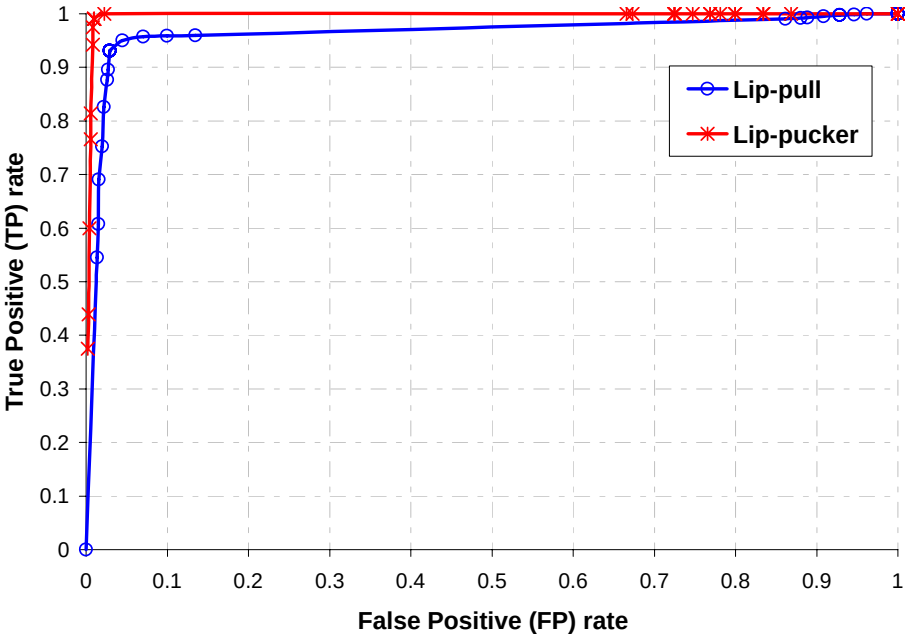


Figure 6.11: Lip pull and pucker ROC curves.

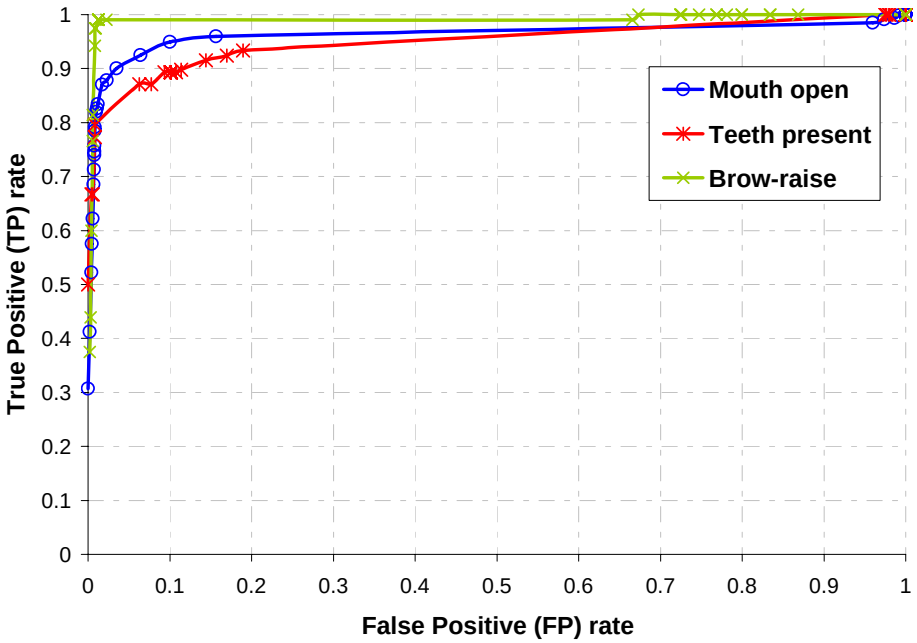


Figure 6.12: Mouth-open, teeth-present and eyebrow raise ROC curves.

Table 6.3: Reasons for undetected and false detections of displays.

Display	Undetected displays	False positives
Head nod	Weak head nod	Head dip
Head shake	Weak head shake	Turn+pitch or translation
Head Tilt	Not represented by HMM	Noise in coding
Head Turn	Not represented by HMM	Noise in coding
Lip pull	Video starts/persists lip-pull	Video starts with pucker
Pucker	Video starts/persists pucker	Video starts with lip-pull
Mouth open	Noise in coding	Noise in coding
Teeth-present	Dark teeth colour	Specular reflection
Eyebrow raise	Too subtle	z-plane motion

Facial displays

To explain the results of lip actions, recall from Chapter 5 that the calculation of the lip actions compares the percentage change in polar distance—the distance between each of the two mouth corners and the anchor point—compared to an initial neutral frame. Some of the cases where the initial frame is non-neutral may result in undetected or falsely detected displays. The main reason accounting for misclassified mouth displays is that of inconsistent illumination. The effects on feature extraction are clear: features from frames with lighting effects might be projected in a different area in luminance-saturation space, resulting in an erroneous interpretation of whether aperture or teeth are present. Finally, most of the undetected cases of eyebrow raises were ones that were subtle, while z-plane motion, moving toward the camera in particular, was the main reason for false eyebrow raises. To minimize the number of false positive cases of an eyebrow raise, scale invariance can be achieved by normalizing against the distance between the two inner eye corners.

6.6 Summary

In this chapter I have defined the notion of head and facial displays, which are facial signals with communicative intent, and their role as an intermediate abstraction in the process of recognizing mental states from head and facial action units. I then presented a system that for the real time recognition of these displays given a sequence of actions.

Displays are described as periodic or episodic depending on the dynamics of the input action sequences. Each display is represented as an HMM that is trained as a classifier for that display. HMMs are particularly suited for representing and classifying displays because they incorporate the dynamics of the actions that constitute these displays, while accounting for variations in these dynamics.

Action sequences from a video of arbitrary length are analysed spatio-temporally. Classification, which implements a sliding window of observations, executes in real time, recognizing a display well within its total duration. The experiments demonstrated reliable recognition of displays that were sampled from a range of complex mental states. The output of the HMM classifiers is concatenated to form the input to the topmost level of the computational model of mind-reading: mental state inference.

Chapter 7

Inference of Complex Mental States

This chapter describes the top-most level of the automated mind-reading system: the inference of complex mental states from head and facial displays in video. I use Dynamic Bayesian Networks (DBNs) to represent complex mental states, and show how the parameters and structures of the networks are determined from training videos. A post-hoc analysis of the resulting models yields an insight into the relevance of specific head and facial signals in discriminating six groups of 29 concepts of complex mental states. The chapter then presents how the DBNs are used to infer the probability of an incoming video sequence being "caused" by each of the states. The framework is optimized so that the latency of the system, defined as the time elapsed between the onset of a mental state and the system recognizing it, is minimal.

7.1 The uncertainty of mind-reading

Mental states constitute the top level of the computational model of mind-reading, and denote the affective and cognitive states of the mind. A person's mental state is not directly available to an observer. Instead it is communicated through nonverbal cues of which the face is arguably the most important. Chapters 5 and 6 described a system for the recognition of observed head gestures and facial expressions from a continuous video stream in real time. The quantized output probabilities of the display HMMs in Chapter 6 are the input to this level of the automated mind-reading system.

The process of reading the mind in the face is inherently uncertain. People express the same mental state using different facial expressions, at varying intensities and durations. In addition, the recognition of head and facial displays is in itself a noisy process. Bayesian probability theory accounts for the uncertainty in models by combining domain knowledge with observational evidence. It is especially powerful when represented as a graph structure, resulting in a probabilistic graphical model (PGM).

A PGM is a graph that represents the causal probability and conditional independence relations among events. It is more intuitive and is computationally more efficient than the respective probability model. The graph θ is learned from training videos of facial expressions of mental state classes. The observational evidence consists of the head and facial displays that were recognized up to the current time $\mathbf{Y}[1:t]$, along with the

previous mental state inferences $P(\mathbf{X}[1:t-1])$. The objective is to compute and update the belief state of a hidden mental state over time: $P(X_i[t]|\mathbf{Y}[1:t], X_i[1:t-1], \theta)$ where $1 \leq i \leq x$. The belief state is conditioned on the head and facial displays that are recognized throughout a video, their dynamics (duration, relationship to each other, and when in the video they occur), the previous mental state inferences, and domain knowledge.

7.2 Complex mental states as Dynamic Bayesian Networks

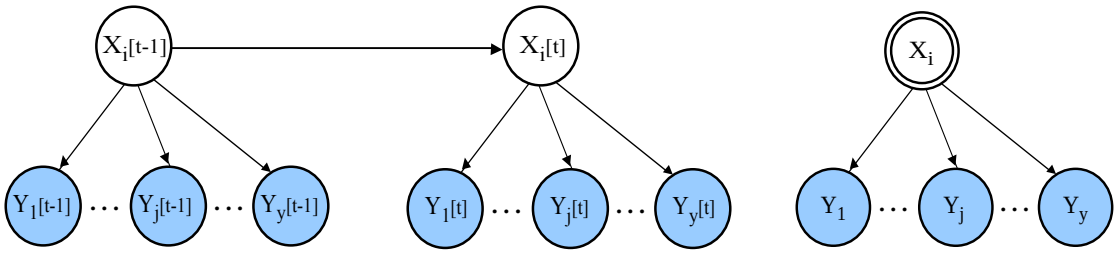
The belief state of a hidden mental state is conditioned on head and facial displays that progress over time. A dynamic classifier was needed to take into account this temporal information. The automated mind-reading system uses DBNs to model the unfolding of hidden mental states over time. With top-down reasoning, the models specify how mental states give rise to head and facial displays. With bottom-up inference, mental states are classified with a certainty level given observations of displays.

7.2.1 Dynamic Bayesian Networks

DBNs are a class of probabilistic graphical models in which nodes represent random variables or events, and the arcs (or lack of arcs) represent conditional independence assumptions. As I have explained earlier, PGMs provide an intuitive visual representation of the corresponding probabilistic models, and provide a compact parameterization of the underlying process. This in turn results in efficient inference and learning [Hec95]. DBNs encode the relationship among dynamic variables that evolve in time. Bayesian inference is used to update the belief state of the hidden variables based on the observation of external evidence over time. A comprehensive review of DBNs can be found in Murphy's thesis [Mur02].

DBNs are successfully used in applications such as activity recognition, facial event analysis, and user-modelling. In activity recognition, Garg *et al.* [GPR03] and Choudhury *et al.* [CRPP02] fuse asynchronous audio, visual and contextual cues in a DBN framework for speaker detection. Park and Aggarwal [PA04] present a DBN framework for abstracting human actions and interactions in video into three levels. In facial event analysis, Hoey and Little [HL03, HL04] use DBNs in the unsupervised learning and clustering of facial displays. Zhang and Ji [ZJ03] apply DBNs in the active fusion of facial actions to recognize facial expressions of basic emotions in image sequences. Gu and Ji [GJ04] present a task-oriented DBN to monitor driver vigilance using facial event classification. In user-modelling, DBNs are used in educational games to track student game actions. These actions provide evidence to assess student knowledge as the game proceeds [BC03, CGV02, ZC03].

DBNs enjoy several characteristics that make them an appealing framework for vision-based inference problems. First, they may function as an ensemble of classifiers, where the combined classifier often performs better than any individual one in the set [GPH02]. This lends itself nicely to vision problems that typically involve multiple cues, such as the colour and shape of objects, or information about the objects in a scene. Second, they incorporate multiple asynchronous cues within a coherent framework, which is a characteristic of many vision problems that deal with the motion of objects in a scene. Third, DBNs can model data at multiple temporal scales. This makes them well suited to modelling human behaviour, which is often hierarchically structured and is perceived at multiple abstraction levels. Finally, it is possible to compose large DBNs by reusing and extending existing ones.



(a) DBN unrolled for two consecutive slices (b) Rolled-up DBN

Figure 7.1: Generic DBN representing the hidden (unshaded) mental state event X_i and the observed displays $\{Y_1, \ldots, Y_y\}$. Note that for the purposes of this level, the displays are observed (shaded) even though they are inferred from HMM classifiers.

7.2.2 Representation

One approach to represent x mental state classes entails defining a single DBN with one hidden state that can assume one of x values. Instead, I decided to represent each mental state as a separate DBN where the hidden mental state of each network is a mental state event. The event has two possible outcomes: true whenever the user is experiencing that mental state, and false otherwise. Having a model for each class means that the hidden state of more than one DBN can be true. Hence, mental states that are not mutually exclusive or may co-occur can be represented by the system. Having a model for each class also means that the parameters of each can be learned independently of the rest of the classes, boosting the accuracy of the results.

Figure 7.1 illustrates a DBN for representing the interaction between a hidden mental state and the set of observable displays. Like all graphical models, a DBN is depicted by its structure and a set of parameters. The structure of the model consists of the specification of a set of conditional independence relations for the probability model, or a set of (missing) edges in the graphical model. The structure of a DBN, being dynamic, is described in terms of an intra-slice and an inter-slice topology.

The intra-slice topology describes the dependencies of the variables within one slice, that is, at a single point in time. In Figure 7.1, the hidden (unshaded) node X_i represents a mental state event. It influences y observation nodes (shaded), which describe what an observer would see another person doing: $\{Y_1, \ldots, Y_y\}$. The assumption is that displays are independent of each other For example, a head nod and a smile are independent given the mental state *agreeing*.

The resulting static model is known as a Naive Bayes model, where each of the display classifiers are trained independently of each other and combined based on their joint performance on some training data. The conditional independence assumption yields surprisingly good results in many classification problems, even though this assumption usually does not reflect the true underlying model generating the data [Pea88]. Domingos and Pazzani [DP97] explore the conditions for the optimality of the Naive Bayes classifier and show that under zero-one loss (misclassification rate) the classifier's region of optimal performance is far greater than that implied by the independence assumption. Variations on the Naive Bayes model include the Tree-Augmented BNs and the Semi-naive Bayes [CSC+03a].

The inter-slice topology depicts the temporal dependency between variables across two consecutive time slices. In terms of graphical representation, DBNs can be shown as unrolled or rolled-up. In the unrolled representation, two or more time-slices are shown:

Algorithm 7.1 Specifying a 2T-DBN

Objective: Specify DBN for mental state i with $y+1$ nodes per time slice (Figure 7.1)

Define DBN structure
  $n_i = 2(y+1)$ nodes over the two slices
  for all j in n_i **do**
    if j is a mental state node **then**
      Number of neighbours=$y+1$
    else
      Number of neighbours=1
    Specify the neighbours
    Specify the type of neighbour: Parent or Child
    Bind j to binary variable

Instantiate DBN with random θ_i
Find θ_i as described in Section 7.3
Update DBN with θ_i

in Figure 7.1(a), an additional arc between $X_i[t-1]$ and $X_i[t]$ encodes the temporal dependency between that mental state class in consecutive slices of the network. In the rolled-up case, only a single time-slice is shown, and dynamics nodes such as X_i are indicated by a double circle (Figure 7.1(b)).

The total number of nodes in each model is $y+1$, where y is the number of observable displays. All the nodes denote events that have two possible outcomes, true or false. The parameter set θ_i for mental state i is described in terms of an observation function, a state-transition function, and a prior as follows:

- The **observation matrix** $B_\phi = \{b_{ij}\}$ denotes the conditional probability distribution tables for each two connected nodes in the graph, where for $1 \leq i \leq x$ and $1 \leq j \leq y$,
$$b_{ij} = \begin{bmatrix} P(Y_j|X_i) & P(Y_j|\overline{X}_i) \\ P(\overline{Y}_j|X_i) & P(\overline{Y}_j|\overline{X}_i) \end{bmatrix}$$
- The **transition matrix** $A = \{a_{ii}\}$ is the conditional probability distribution table for the temporal transition between two variables that are connected across slices where for $1 \leq i \leq x$,
$$a_{ii} = \begin{bmatrix} P(X_i[t]|X_i[t-1]) & P(X_i[t]|\overline{X}_i[t-1]) \\ P(\overline{X}_i[t]|X_i[t-1]) & P(\overline{X}_i[t]|\overline{X}_i[t-1]) \end{bmatrix}$$
- The **prior** $\pi = \{\pi_i\}$, where $\pi_i = P(X_i[0])$ for $1 \leq i \leq x$ represents our prior belief about mental state i.

The model is given by its joint probability distribution:

$$P(X_i, \mathbf{Y}, \theta) = P(\mathbf{Y}|X_i, B_\phi)P(X_i|A, \pi)$$

Programmatically, specifying a DBN in Matlab's BNT [Mur01] or its equivalent C++ version, Intel's Probabilistic Networks Library (PNL) [PNL03], involves defining the structure of a DBN over two time slices (2T-DBN). The steps are shown in Algorithm 7.1. Essentially, for each of the $2(y+1)$ nodes over the two slices of a DBN, the neighbour at each node and its type are specified. The DBN is instantiated with random parameters and then updated once the actual parameters are determined.

7.3 Learning

Learning is the problem of determining the parameters and defining the structure (model selection) of a DBN. This process is data-driven: as long as training examples are available for any mental state class, then it is possible to learn a model of that class. Currently, learning is implemented as a one time off-line process.

Table 7.1: Structure and parameter learning in DBNs. From Murphy's thesis [Mur02].

Structure	Observability	Method	Section number
Known	Full	Maximum Likelihood Estimation	Section 7.3.1
Known	Partial	Expectation Maximization (EM)	not applicable
Unknown	Full	Search through model space	Section 7.3.4
Unknown	Partial	EM + search through model space	not applicable

Table 7.1 summarizes four possible cases of learning based on whether the network structure is known or not, and whether the training data is fully observed or not. The training data is fully observed when the values or labels of all the nodes—including the hidden ones—are available. The training data is partially observed when there is missing data and/or latent variables. Learning in the partially observed case is much harder as it requires the use of approximation inference methods that are computationally more expensive [Hec95, HL04].

The videos from the Mind Reading DVD constitute the training data. They are fully observed since they have already been labelled with visible head/facial displays and a mental state class. A Naive Bayes model structure is assumed as shown in Figure 7.1. Thus, to estimate the parameters of the DBNs I use Maximum Likelihood Estimation (first row in Table 7.1). Later, the network structure is challenged, and a search through the model space is carried out to find a (locally) optimal network structure (third row).

7.3.1 Parameter estimation

When the data is fully observed and the network structure is known, Maximum Likelihood Estimation (MLE) can be used to estimate the parameters of a DBN. When the nodes are discrete, MLE amounts to counting how often particular combinations of hidden state and observation values occur, and the parameters are estimated as follows:

- **Observation matrix** B_ϕ: the conditional probability distribution table for display j and mental state i in B_ϕ is determined by $P(Y_j|X_i)$ and $P(Y_j|\overline{X}_i)$. The probability of observing each head/facial display given a mental state, $P(Y_j|X_i)$, is computed by counting the number of occurrences of Y_j in X_i. The probability of observing a display given all other mental states in a set of training examples, $P(Y_j|\overline{X}_i)$, is given by the number of occurrences of Y_i in all mental states except X_i.
- **Transition matrix** A: the transition function of the system can be learned independently of the observation matrix. It is given by the number of transitions between hidden state values over time. Since each video in the training set maps to a single mental state, transitions from one mental state to another are not available. Hence, the statistical truth of the data is such that the probability of moving from $X_i[t-1]$ to $X_i[t]$ is one, and moving from $X_i[t-1]$ to $\overline{X}_i[t]$ is zero. Likewise, the probability of $X_i[t]$ given $\overline{X}_i[t-1]$ is zero.
- **The priors** π: the priors are set to the frequency of each class in the training set.

MLE provides a closed-form solution to estimating the parameters of the DBNs, and the resulting model is amenable to exact inference [Hec95]. By contrast, Bayesian learning tries to learn a distribution over the parameters, and even though it is more elegant, inference algorithms are computationally intensive.

7.3.2 Discriminative power heuristic

In addition to the DBN parameters θ, I define a heuristic $H = P(Y_j|X_i) - P(Y_j|\overline{X}_i)$, which is later used in the parameter estimation process. The magnitude of H quantifies the discriminative power of a display for a mental state; the sign depicts whether a display increases or decreases the probability of a mental state. To explain how the heuristic works, consider the following hypothetical cases of the discriminative ability of a head nod in identifying the mental state *agreeing*:

1. Assume that a head nod is always present in *agreeing* $P(Y_j|X_i) = 1$, but never appears in any of the other mental states $P(Y_j|\overline{X}_i) = 0$. The heuristic is at its maximum value of one, and its sign is positive. The presence of a head nod is a perfect discriminator of *agreeing*.

2. Now, consider that a nod never shows up in *agreeing*, $P(Y_j|X_i) = 0$, but always shows up in all other mental states $P(Y_j|\overline{X}_i) = 1$. The magnitude of H would still be one, but its sign would be negative. In other words, the head nod would still be a perfect discriminator of *agreeing*.

3. Finally, if a head nod is always observed in *agreeing*, and is also always observed in all other mental states, then $P(Y_j|X_i) = P(Y_j|\overline{X}_i) = 1$. In this case, H has a value of 0, and the head nod is an irrelevant feature in the classification of *agreeing*.

7.3.3 Results of parameter estimation

DBNs are data-driven and hence are not fixed to particular mental states. Nonetheless, to validate the DBN framework I chose six complex mental state groups and the mental state concepts they encompass, which were derived from the taxonomy of Baron-Cohen *et al.* [BGW+04]. A tree diagram of the mental state groups and concepts was presented in Chapter 3. The six groups are *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*.

The parameters are summarized in Figures 7.2–7.7. They depict the conditional probability distribution tables and discriminative-power heuristic for each display and mental state combination. It is possible to think of the parameters of each mental state as its "signature". Note that the probabilities reach a maximum of 0.5. This confirms the findings of the studies presented in Chapter 3: on their own, none of the displays are particularly strong classifiers of any of the mental states.

Effect of choice and size of training set on the parameters

When MLE is used for parameter estimation, the resulting models are a factor of the size and choice of the examples used for training. To account for this variability, the parameters reported in this section are the results of several runs of MLE for each class. Table 7.2 summarizes the total number of runs carried out for each class. For instance, the parameters for *agreeing* were computed from 64 runs. These runs were generated from the 34 videos representing *agreeing* as follows:

Table 7.2: Number of runs of Maximum Likelihood Estimation for each mental state class

agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
64	48	51	60	61	60

- Using a leave-one-out methodology, one video is removed and MLE is run on the remaining 33 videos. This results in 34 different runs of MLE with varying training examples. Note that each video contributes to more than one training example.

- Using a leave-four-out methodology, four videos are randomly removed from the set and MLE is run on the remaining 30 videos. I have chosen to do 30 runs of MLE to test the effect of different combinations of training examples.

- The mean, max, and min values over all 64 runs are computed for $P(Y_j|X_i)$, $1 \leq j \leq y$. The max and min values depict the range of $P(Y_j|X_i)$, that is the result of varying the size and choice of training examples.

- Note that in estimating the parameters of *agreeing*, leave-one-out and leave-four-out are applied to the positive examples of that class. The training examples of all other classes are fixed. That is, $P(Y_j|\overline{X}_i)$ for $1 \leq j \leq y$ is fixed across all the MLE runs of class i, and there is no range reported.

- The mean, max, and min values of $|H| = |P(Y_j|X_i) - P(Y_j|\overline{X}_i)|$ are computed given the mean, max, and min values of $P(Y_j|X_i)$ and the fixed value of $P(Y_j|\overline{X}_i)$.

- The range of values for $P(Y_j|X_i)$ and $|H|$ is indicated by the error bars in Figures 7.2-7.7. It shows that MLE is robust to the choice of size and videos used for training.

Agreeing

The *agreeing* class is a refinement of the *sure* group in the Mind Reading DVD that encompasses the mental states that communicate agreeing, granting consent or reaching a mutual understanding about something. It is associated with a *feeling of knowing*, which is a cognitive state [Hes03]. The class includes *assertive*, *committed*, *convinced*, *knowing*, *persuaded* and *sure*, each represented by six videos. Out of the 36 videos, `FaceTracker` fails on two of them; the resulting 34 videos generate a maximum of 855 training samples, since every video in the training set contributes to more than one inference instance.

Figure 7.2 shows the parameters for the *agreeing* DBN. The lip corner pull, head nod and head turn are the most likely displays to occur. Of the three, the head nod is the strongest discriminator of *agreeing*. This is because even though the probability of a lip corner pull occurring in *agreeing* is higher than that of a head nod, the latter seldom occurs in any of the other mental states, while the lip corner pull does. By the same token, the head turn has no discriminative power because it has an equal probability of incidence in *agreeing* as in all other mental states.

The prevalence of the head nod as a strong identifier of *agreeing* is an important finding because it confirms the few studies in psychology that observe people who are in this mental state. For example, in one study parents are reported to give a vertical nod with a smile to their children when approving of their conduct [Dar65].

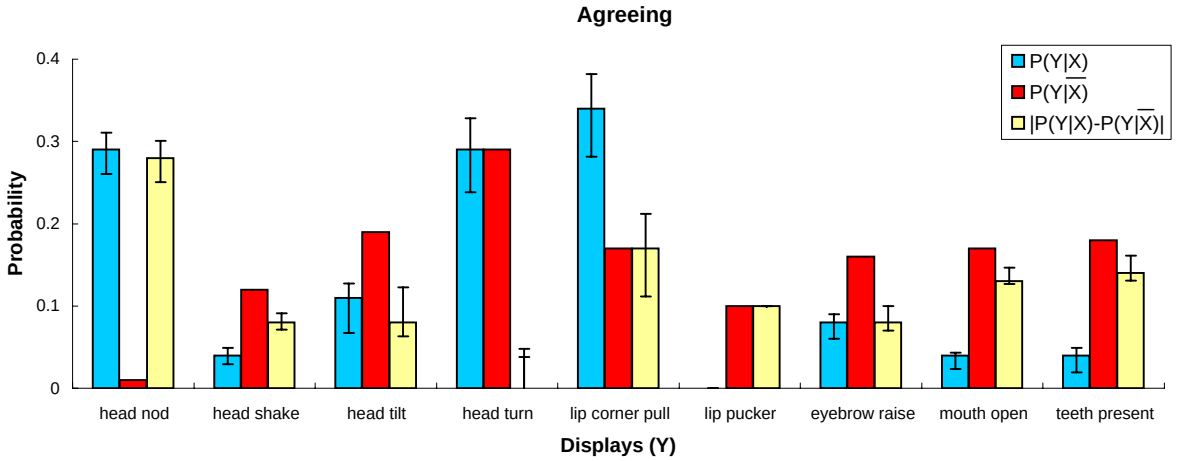


Figure 7.2: The results of parameter estimation for *agreeing*. While the lip corner pull is the most likely display to occur, the head nod is the most discriminative. The error bars depict the effect of the size and choice of training examples on the parameters.

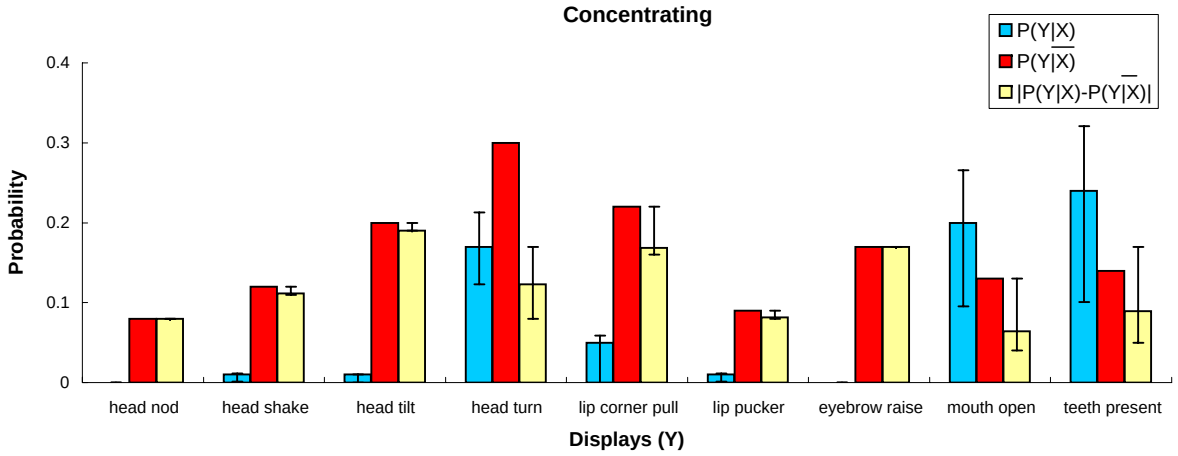


Figure 7.3: The results of parameter estimation for *concentrating*. The presence of teeth is the most likely display to occur; the head tilt is the most discriminative.

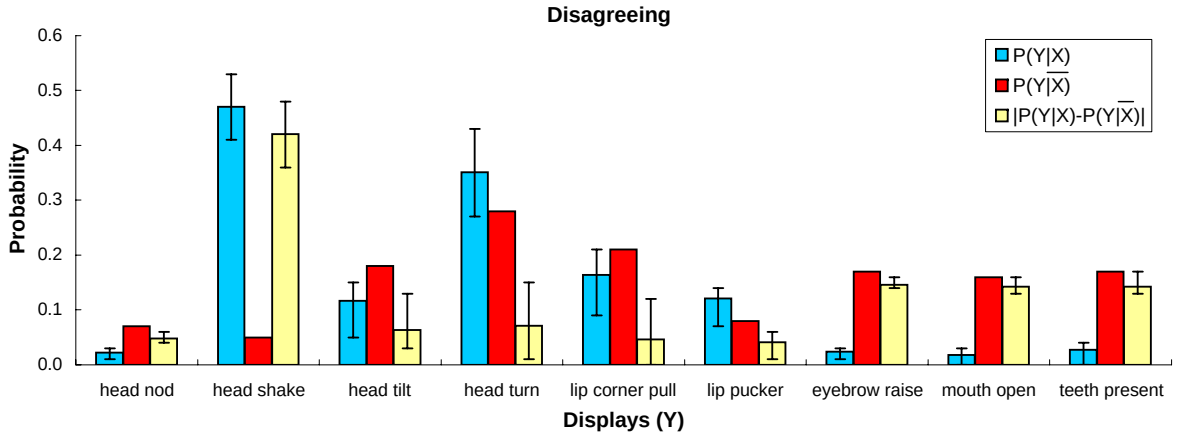


Figure 7.4: The results of parameter estimation for *disagreeing*. The head shake is the most likely display to occur and it is also the most discriminative.

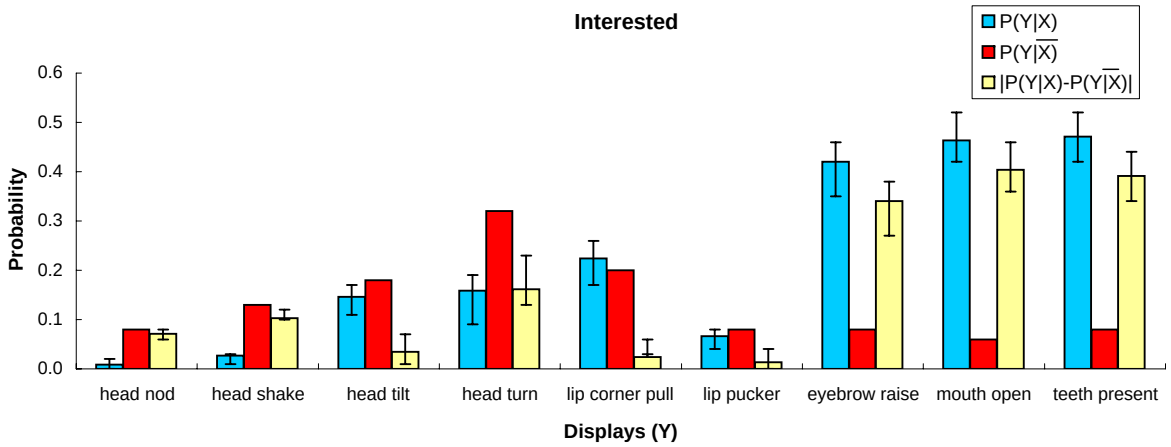


Figure 7.5: The results of parameter estimation for *interested*. The eyebrow raise, mouth open and presence of teeth are the most likely displays and the most discriminative.

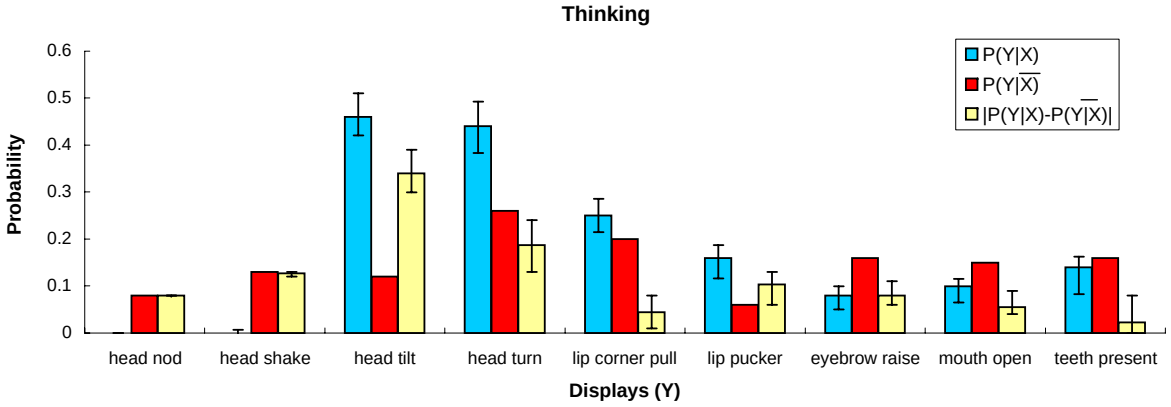


Figure 7.6: The results of parameter estimation for *thinking*. The head tilt and the head turn are the most likely displays to occur and are also the most discriminative.

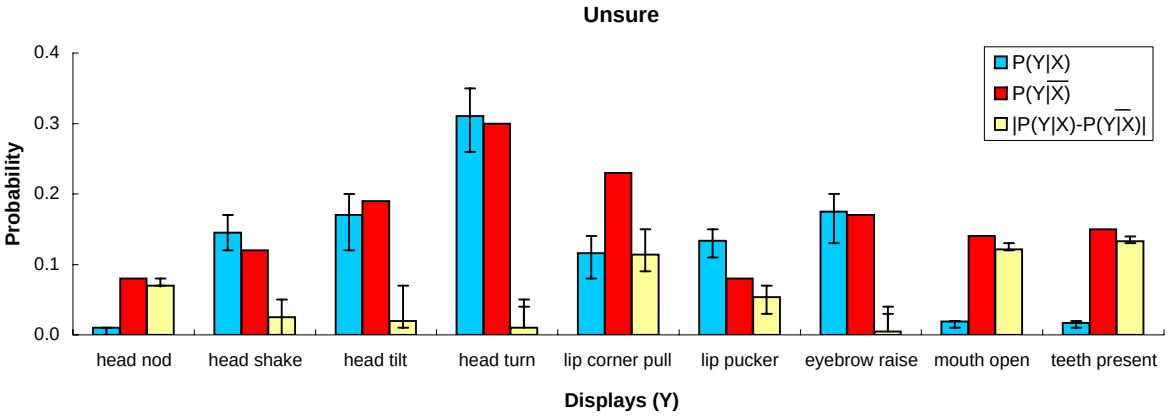


Figure 7.7: The results of parameter estimation for *unsure*. While the head turn is the most likely display to occur, the absence of teeth and closed mouth are the most discriminative.

Concentrating

The *Concentrating* class includes the following three mental state concepts from the Mind Reading DVD: *absorbed*, *concentrating* and *vigilant*. Each concept is represented by six videos for a total of 18 videos, or 465 training samples. Figure 7.3 shows the learned model parameters for the *concentrating* DBN. An open mouth and the presence of teeth are the two most likely displays to occur. However, they are not strong discriminators of *concentrating*. Rather, it is the (absence of) head tilt, lip corner pull and eyebrow raise that are the strongest indicators of *concentrating*, although individually they would have been weak discriminators.

In general, the parameters indicate that there is a low incidence of displays occurring in *concentrating*, an effect that is evident in the little head or facial movement in these 18 videos. This observation is interesting since, in the literature, concentration is referred to as a state of "absorbed meditation" [Dar65]. Indeed, under the behavioural ecology view of facial expressions [Fri92], the low incidence of facial displays in concentration indicates inaction. This inaction is designed to discourage interaction with others ("Do not bug me now, I am in the middle of something") [RC03].

Disagreeing

The *disagreeing* class communicates that one has a differing opinion on or is disputing something. It includes: *contradictory*, *disapproving*, *discouraging*, *disinclined*, each represented by six videos. Out of the 24 videos, `FaceTracker` fails on three; the remaining 21 videos generate 553 training examples. Figure 7.2 shows the parameters for the *disagreeing* DBN. The head shake is the most likely display to occur. This confirms the literature on *disagreeing* where people are reportedly seen to move their heads several times from side to side, or shake their heads in negation [Dar65]. Finally, as in *agreeing*, the head turn has little discriminative power in *disagreeing* despite a high probability of occurring.

Interested

Being interested indicates that one's attention is directed to an object or class of objects. The ability to recognize if a person is interested is especially relevant in intelligent tutoring systems. The class includes: *asking*, *curious*, *fascinated*, *impressed* and *interested*, each represented by six videos. The 30 videos generate 782 training samples. Figure 7.5 summarizes the learned model parameters for the *interested* DBN. The figure shows that the eyebrow raise, mouth open and presence of teeth are the most likely displays to occur, and they are also the most discriminative since $P(Y_j|\overline{X}_i)$ is low for all three displays. The eyebrow flash is often linked to interest [Ekm79].

Thinking

Thinking communicates that one is reasoning about, or reflecting on, some object: *brooding*, *choosing*, *fantasizing*, *judging*, *thinking* and *thoughtful*. The 36 videos generate 681 training samples.

Figure 7.6 summarizes the learned model parameters for the *thinking* DBN. It shows that the head tilt and head turn orientation are the most likely displays to occur, and they are also the most discriminative. Studies by Baron-Cohen and Cross [BC92] have shown that people infer that a person is thinking when a person's head orientation and eye-gaze are directed away from the viewer, to the left or right upper quadrant, and when there is no apparent object to which their gaze is directed.

Unsure

Unsure communicates a lack of confidence about something, and is associated with a feeling of not knowing, which like the feeling of knowing is a cognitive one [Hes03]. The class encompasses the following mental states concepts: *baffled*, *confused*, *puzzled*, *undecided*, and *unsure*. Out of the six mental states, confusion is the most commonly cited in the general literature of cognition and emotion. With six videos representing each concept there is a total of 30 videos, of 809 training samples.

As shown in Figure 7.7, the head turn is the most likely display to occur in *unsure*. However, it is not a strong discriminator of *unsure* since the display occurs nearly as often in all other mental states. Even though the probability of an open mouth and presence of teeth are low, their discriminative power is higher than that of the head turn. In other words, a closed mouth and absence of teeth are indicators of *unsure*.

7.3.4 Model selection

The results of parameter estimation show that the head and facial displays that are most relevant in discriminating mental states are not by necessity the same across mental states. This observation provided the motivation to implement model selection in search for the optimal subset of head and facial displays most relevant in identifying each of the mental states. Using only the most relevant features for the DBN structure reduces the model dimensions without impeding the performance of the learning algorithm, and improves the generalization power of each class by filtering irrelevant features [Bil00].

One possible approach to picking the most relevant features is to use the discriminative-power heuristic from Section 7.3.3. The problem, however, is that the difference in the values of the heuristic were in some cases very small, such as in the *concentrating* class. The small differences in the heuristic could be an artefact of the training data. Picking features based on that may result in over-trained models. The alternative is to select models on the basis of their classification performance on some test set. The resulting models would represent only the most relevant features of each class; the features that are common across all classes or do not help in classification are not included in the models. The models are optimized for classification performance (discriminative) as opposed to generating good examples (generative).

Assuming the inter-slice topology is fixed, the problem of feature selection is an optimization one defined as follows: given the set of y displays $\mathbf{Y}$, select a subset that leads to the smallest classification error for videos in a test set of size S. Each video in the set yields T mental state inferences. The classification error of a single instance within a particular video is $1 - P(X_i[t]|\mathbf{Y}[1:t])$. Accordingly, the classification error of mental state i is given by the sum of the errors over the T instances for all S videos:

$$e_i = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} 1 - P(X_i[t]|\mathbf{Y}[1:t]) \tag{7.1}$$

The most straightforward approach to feature selection would involve examining all possible subsets and selecting the subset with the least value of e_i. However, the number of possible subsets grows exponentially, making this exhaustive search impractical for even moderate values of y. Deterministic, single-solution methods, such as the sequential forward selection (SFS), the sequential backward selection (SBS) [MM63] and

their respective floating versions [FPHK94] are commonly used methods for performing feature selection. While the floating versions yield better results, SFS and SBS are the simplest to implement and are reasonably faster [JZ97]. I use SBS to find a subset of observation nodes for each mental state such that the classification error of the DBN is minimized (Algorithm 7.2). The algorithm works by eliminating features recursively from the feature set such that the classification error of the feature subset is minimized.

Algorithm 7.2 Sequential backward selection (SBS) for DBN model selection

Objective: For each of the x mental state DBNs, find a subset of the y displays $\mathbf{Y} = \{Y_1, \ldots, Y_y\}$ that leads to the smallest classification error e_{min} on a test set of size S

```
subset SBS(F, e)
    e_min = e
    for all display Y_j in F do
      e_s = ComputeClassificationError(F-Y_j)
      if e_min > e_s then
        e_min = e_s
        J = Y_j
    if e_min < e then
      return SBS(F-J, e_min)
    else
      return F

BestSubset = SBS({Y_1, ..., Y_y}, ComputeClassificationError({Y_1, ..., Y_y}))
```

7.3.5 Results of model selection

Figure 7.8 shows a trace of the search algorithm for *agreeing*. Initially, the search starts with the full display set of size y, and evaluates the classification error e_s for a fixed test set. A display is then removed recursively from the feature subset **F** and once again is evaluated on the test set. The classification error e_s is compared to the current lowest bound e_{min}. In the case where e_s is less than or equal to e_{min}, the current bound is updated and that branch is searched further, otherwise it is pruned. This is repeated until there is no further reduction in classification error. Note that the algorithm does not guarantee a global optimum since that depends on the training and the test sets.

Table 7.3: Summary of model selection results. Column i summarizes how the probability of mental state i is affected by observing evidence on each of the displays. Row j depicts the effect of observing display j on the probability of each of the mental states.

	agreeing	concentrating	disagreeing	interested	thinking	unsure
head nod	+0.28	-0.08	-0.05	-0.07	-0.08	-0.07
head shake		-0.11	+0.42		-0.13	+0.04
head tilt		-.019	-0.06		+0.34	
head turn					+0.18	
lip corner pull	+0.17	-0.17				-0.1
lip pucker	-0.10				+0.1	+0.06
mouth open	-0.13	+0.07	-0.14	+0.40		-0.05
teeth present	-0.14		-0.14	+0.39		-0.17
eyebrow raise	-0.08	-0.17	-0.15	+0.34	-0.08	

The results of sequential backward selection for each of the mental state classes are summarized in Table 7.3. A non-blank entry at cell (j, i) implies that display j is present

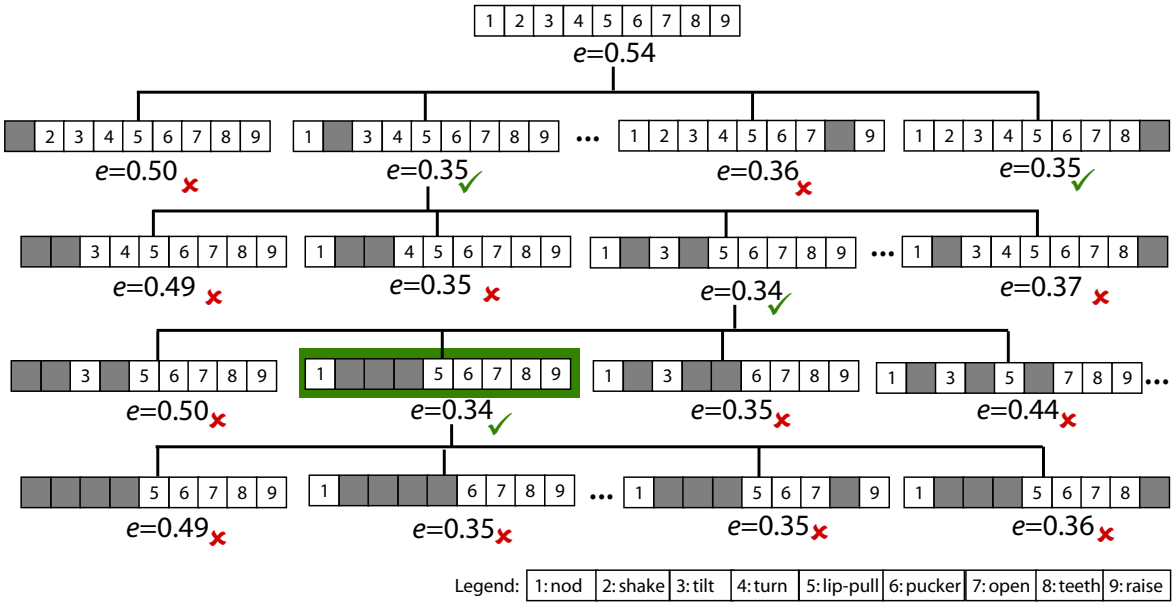


Figure 7.8: Sequential backward selection to select the set of displays in *agreeing*. The search starts at the top with the full set of displays (y=9). A feature is then removed recursively from the feature subset. The classification error e is shown below each feature subset. At each level, the feature subset with the least classification error (indicated with a ✓) is explored further, the others are pruned. The resulting DBN for *agreeing* consists of the head nod, lip corner pull, lip pucker, mouth open, teeth present, and eyebrow raise.

in the converged model of mental state i, while a blank entry implies that it is not. The results are consistent with the discriminative power heuristic presented in Section 7.3.3 in that the most discriminative displays are, in fact, the ones which were selected by the sequential backward selection algorithm.

The value of a non-blank cell (j, i) is the sign and magnitude of the discriminative-power heuristic H of display j for mental state i from Section 7.3.3. A positive value in cell (j, i) means that observing display j increases the probability of mental state i, while a negative one means that observing display j decreases that probability. The magnitude depicts the extent with which the probability changes. Hence, Table 7.3 predicts how the DBNs behave when evidence of head and facial displays is observed:

- **Column** i summarizes how the probability of mental state i is affected by observing evidence on each of the displays. For instance, the table predicts that the presence of an open mouth, teeth or eyebrow raise would increase the probability of *interested*, while a head nod would decrease it, assuming that the probability was non-zero.

- **Row** j depicts the effect of observing display j on the probability of each of the mental states. For instance, observing a head shake would have the following effect: increase the probability of *disagreeing*, and to a lower extent, increase that of *unsure*. It would decrease the probability of *concentrating* and *thinking*, and would have no effect on the probability of *agreeing* and *interested*.

Note that the table only provides a prediction; the actual behaviour of the DBNs depends on the combination of displays recognized, their dynamics, and the probability of the previous mental states. Figure 7.9 shows the resulting DBN of each mental state.

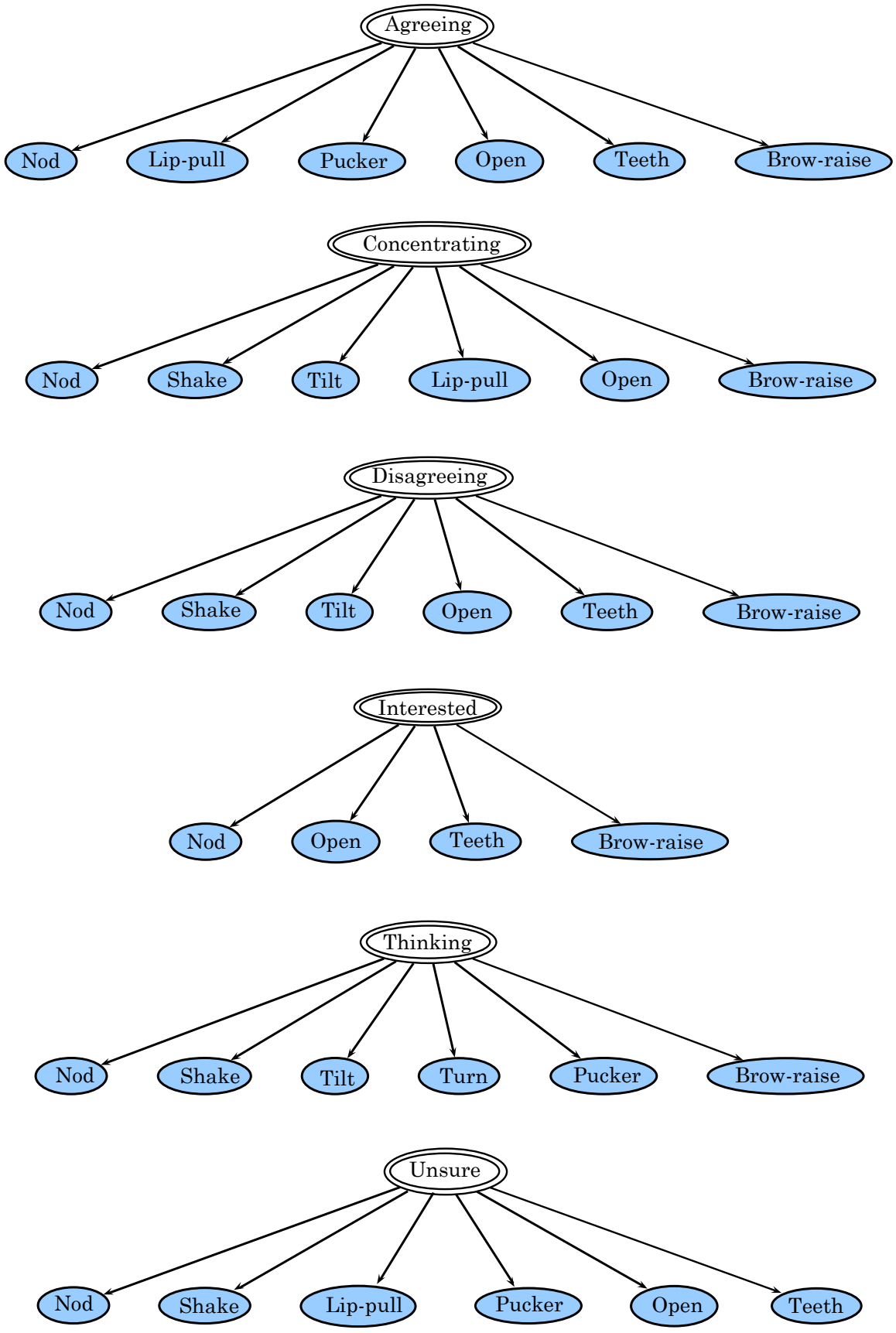


Figure 7.9: Converged mental state DBNs.

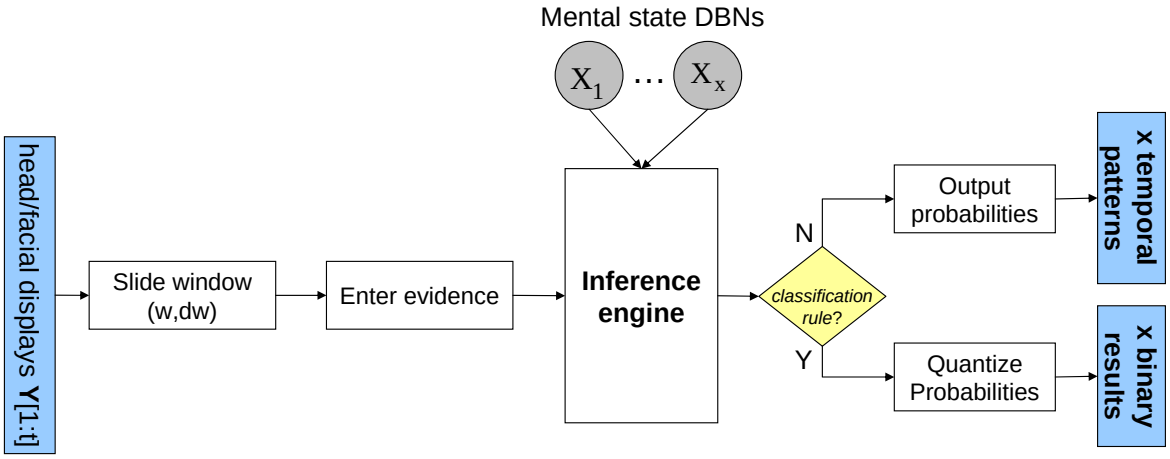


Figure 7.10: Procedural description of mental state inference. Inference is implemented as a sliding window of evidence. The evidence consists of the w most recent displays; it progresses dw actions at a time. Depending on whether or not a classification rule is used, the output of the system is either x binary results or x temporal patterns.

7.4 Inference

Once the structures and parameters of the DBNs are determined, the mental state DBNs act as classifiers for the online inference of complex mental states. Inference involves recursively updating the belief state of hidden states based upon the knowledge captured in the models and the observational evidence.

7.4.1 Inference framework

Figure 7.10 shows a procedural description of how real time mental state inference proceeds in a video of arbitrary length. The classification framework is implemented as a sliding window of evidence. The evidence size is w displays, and it progresses dw actions at a time. At any instant t, the observation vector that is input to the DBN classifiers is a vector of the w most-recent displays $\mathbf{Y}[t-w:t]$, and the corresponding most-recent mental state inferences $P(\mathbf{X}[t-w:t-1]$. The output is a probability that the observation vector was generated by each of the DBNs.

Algorithm 7.3 describes the implementation details of mental state inference. First, an inference engine is instantiated. Because in this specific model the hidden states are discrete variables and there are no loops in the structure of the DBNs, exact inference algorithms can be used. The forward-backward algorithm and the unrolled-junction-tree algorithm are two examples of exact inference algorithms [Mur02]. In the forward-backward algorithm, the DBNs are first converted to an HMM before applying the algorithm. In the unrolled-junction-tree algorithm, the DBNs are unrolled, and then any static Bayesian Network inference algorithm is used. The algorithms are simple and fast for small state spaces.

Once an inference engine is instantiated, the evidence is accumulated over w time slices for the purpose of the first inference instance, which occurs at time $t = w$. The inference involves calculating the marginal probability of X_i by integrating out the values of the observations. The window then slides dw actions, upon which the observation vector is updated and the inference engine invoked.

The choice of the sliding factor dw has an effect on the latency of the system. In the context of mind-reading, latency is the time elapsed between the onset of a mental

state and the system recognizing it. With the current hardware—a Pentium 4 processor, 3.4 GHz, 2 GB of RAM—a sliding factor of one head/facial action is implemented, that is, displays are incorporated into the observational vector for inference as they occur, and a mental state inference is output every 166 ms or five frames at 30 fps.

Algorithm 7.3 Mental state inference

Objective: Compute the belief state of a hidden mental state $P(X_i[t]|\mathbf{Y}[t-w:t], P(\mathbf{X}[t-w:t-1]))$ over time, $1 \leq i \leq x$ and $1 \leq t \leq T$

Given: x mental state DBNs as in Figure 7.1, with y observations nodes $\mathbf{Y}$; evidence length is w and sliding factor, or lag, is dw

Initialization: Instantiate inference engine such as the unrolled-junction-tree or the forward-backward algorithms

Initial inference instance

for all t in w time slices **do**
Get current observations $\mathbf{Y}[t]$;
Enter evidence so far $\mathbf{Y}[1:w]$;
Calculate $P(\mathbf{X}[w]|\mathbf{Y}[1:w])$

Inference

$t = w + dw$
for all t in T time slices **do**
Get current observations $\mathbf{Y}[t]$
Enter evidence so far: $\mathbf{Y}[t-w:t]$ and $P(\mathbf{X}[t-w:t-1])$
Calculate $P(\mathbf{X}[t]|\mathbf{Y}[t-w:t], P(\mathbf{X}[t-w:t-1]))$
Advance window $t = t + dw$

The window size or evidence length w describes the number of most recent observations to include as evidence on each DBN invocation. In physical terms, it denotes the amount of inter-expression dynamics or the amount of head and facial display history to consider when making a mental state inference. The criterion for choosing a value for w is as follows: a small value of w may result in inaccurate results as it may not be incorporating enough inter-expression dynamics during inference. It however produces results as early as 1.3 seconds for $w = 2$. For larger values of w, the system becomes more resilient to noise generated at the bottom two levels. However, the output is smoothed out so some details may be lost. Also the results are produced much later, as late as 3 seconds for $w = 12$. Recall from Chapter 3 that two seconds is the minimum time required for a human to reliably infer a mental state and that video segments of less than two seconds yield inaccurate recognition results. I chose to sample these two seconds (60 frames) using a sliding window of $w = 7$ displays that progresses one facial action at a time.

Figure 7.11 shows an example of the head and facial display evidence over a 5.5-second long video labelled as *discouraging* from the Mind Reading DVD. Note that *discouraging* belongs to the *disagreeing* group. This video should therefore be labelled by the system as *disagreeing* to be considered a correct classification. Throughout the video, a number of asynchronous displays that vary in duration are recognized: a head shake, a head turn, a lip corner pull, and an open mouth and teeth. The resulting mental state inferences are shown in Figure 7.12 for different sliding window sizes. The window size aside, the recognized displays affect the mental state inferences as predicted by Table 7.3.

A sliding window implementation offers a number of advantages. First, it provides a continuous and automatic means of segmenting the input video stream. Second, it enables the system to continuously produce inferences at a frequency determined by dw. Third, it allows the system to account for inter-expression dynamics by processing displays in the context of each other, while still being computationally efficient.

7.4.2 Output of the automated mind-reading system

The output of the automated mind-reading system consists of the probabilities of each of the mental state classes over the course of a video. The change in the probability values of a mental state over time carries important information about a person's mental state. For instance, in Figure 7.12, the increase in the probability of *unsure* between 1.7 and 2 seconds indicates that the person is moving into a state of being *unsure*, while the decrease in the probability of *unsure* during the fifth second suggests that the person is moving out of that state.

The six probabilities (for the six mental states I chose to work with here) and their development over time represent a rich modality analogous to the information humans receive in everyday interaction through mind-reading. For the purpose of measuring performance and comparing to other systems, one could use a classification rule with the output probabilities to force a binary classification result for each mental state. This would be equivalent to asking if, for example, the person in the video is *unsure* or not. It is also possible to reinforce only one correct mental state out of the possible six, although this inherently makes the assumption that mental states do not co-occur.

The classification rule that I adopt to evaluate the recognition accuracy of the system is discussed in the next chapter. For real-world applications however, it is how the mental state probabilities unfold over time that provides the richest source of information.

7.5 Summary

In this chapter I have described the inference of complex mental states from head and facial displays in video using DBNs. This class of probabilistic graphical models are a good choice for representing complex mental states because they act as an ensemble of classifiers, fusing multiple dynamic and asynchronous cues over varying temporal scales.

The results of parameter estimation and model selection for the mental state DBNs provide an insight into the relationship between observable displays and hidden mental states. In particular, the discriminative-power heuristic predicts the effect of observing some evidence on the output of the DBNs. The robustness of MLE to the size and choice of training examples is also demonstrated. The post-hoc analysis of parameter estimation and model selection is an interesting one to repeat with a wider set of displays and mental state classes. The inference framework was designed as a sliding window implementation to ensure that the classification of mental states occurs soon after the onset of observed displays.

The chapter describes several important contributions compared with existing research on facial analysis systems. First, the automated inference of complex mental states is a significant step forward from existing systems that only address the basic emotions. Second, the system supports many fine shades of a mental state, not just the prototype expression of that state as is often the approach in existing facial analysis systems.

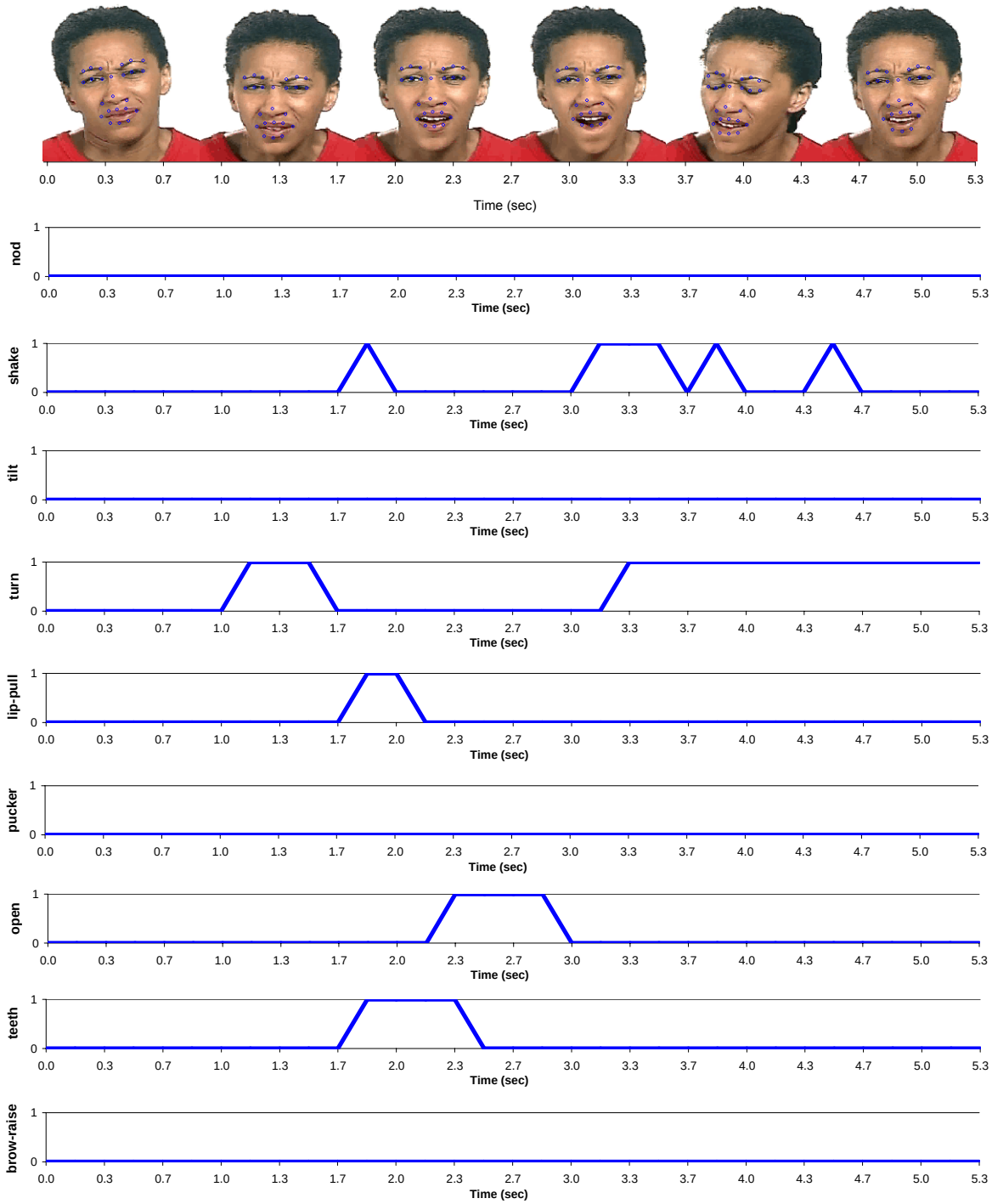
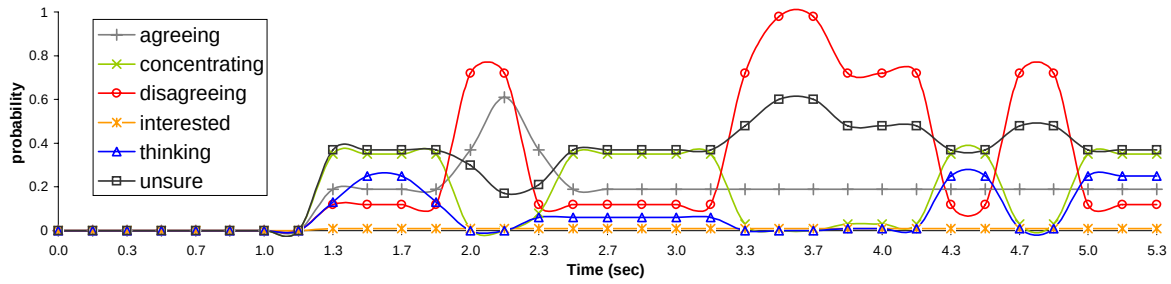
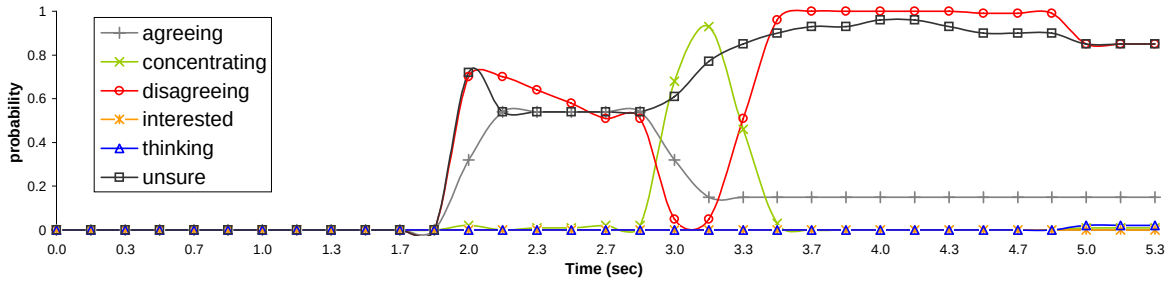


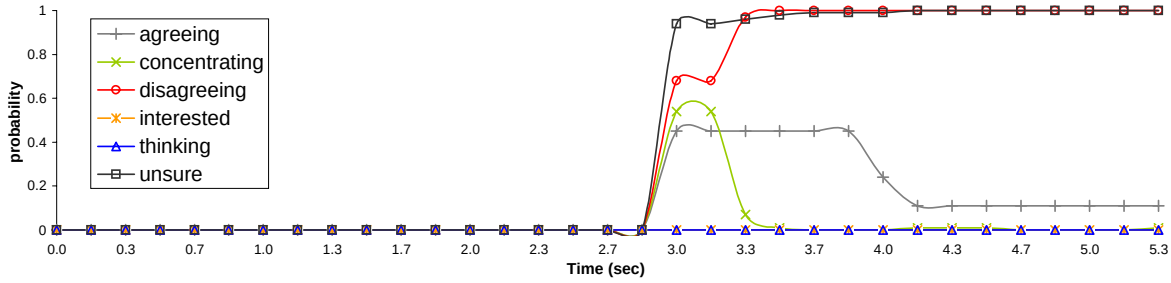
Figure 7.11: Trace of display recognition in a 5.3 second video labelled as *discouraging* from the Mind Reading DVD. *Discouraging* belongs to the *disagreeing* group. The top row shows selected frames from the video sampled every one second. The following nine rows show the output of the head and facial display recognition, which constitute the input to the DBN. The first set of head and facial display output occurs at one second. The mental state inferences for this video are shown in Figure 7.12.



(a) Results of mental state inference for the video in Figure 7.11 using as evidence the two most recent head and facial display outputs for each instance of mental state inference. The first DBN invocation occurs at 1.3 seconds. The probability of *disagreeing* integrated from 1.3 to 5.3 seconds is 0.38, an incorrect classification of the video.



(b) Results of mental state inference using as evidence the six most recent head and facial display outputs for each instance of mental state inference. The first DBN invocation occurs at two seconds. The probability of *disagreeing* integrated from 2 to 5.3 seconds is 0.75, a correct classification. This is the sliding window size that I adopt for the system.



(c) Results of mental state inference using as evidence the 12 most recent head and facial display outputs for each instance of mental state inference. The first DBN invocation occurs at three seconds. The probability of *disagreeing* integrated from 3 to 5.3 seconds is 0.96, a correct classification.

Figure 7.12: The criterion for choosing a value for the sliding window of evidence w, the number of most recent head and facial displays used for each instance of mental state inference. The smaller the value of w, the earlier the first inference results are produced, but the less accurate they are since little history is incorporated for inference. The larger the value of w, the more resilient the system is to noisy observations, but the later the output of the first inference. Also with larger values of w, some detail is lost. The system is configured to use the six most recent head and facial displays for mental state inference. The mental state inferences shown are for the 5.3 second video in Figure 7.11, labelled as *discouraging*. *Discouraging* belongs to the *disagreeing* group.

Third, the relationship between head/facial displays and mental states is explored using the statistical truth of the data. Finally, mental state inference is executed automatically without the need to manually pre-segment videos. The performance of the system in terms of accuracy, generalization and speed is evaluated in the next chapter.

Chapter 8

Experimental Evaluation

This chapter describes the performance of the automated mind-reading system in terms of accuracy, generalization and speed. The accuracy is a measure of the classification performance of a system on a pre-defined set of classes. The generalization ability of a system describes its accuracy when trained on one corpus and tested on a completely different, previously unseen corpus of videos. It is an indicator of whether it is possible to deploy the system with different users and in different settings outside of the lab once it has been trained. The speed measures the real time performance and latency of the system, and is an important consideration if the system is to be used in a natural computing environment.

8.1 Evaluation of the Mind Reading DVD

I evaluated the automated mind-reading system for the following six groups of complex mental states: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The system was trained and tested on the Mind Reading DVD videos using a leave-one-out testing methodology. The system was configured to use the DBNs resulting from model selection in Section 7.3.4. The sliding window parameters are as described in Section 7.4.1: a window that spans the duration of a head/facial display, and progresses six times, 166.7 ms at a time.

Each of the six mental state groups encompasses a number of mental state concepts. For instance, the *thinking* group includes *brooding* and *choosing*. *Brooding* signals that a person is deep in thought or meditating, while *choosing* indicates that a person is selecting from a number of possible alternatives. *Brooding* and *choosing* are two examples of being in a *thinking* state that may be expressed in slightly different ways. The more "shades" or concepts of a mental state one can recognize, the better are one's social intelligence skills [Bar03].

The objective of this experiment was to test the system's ability to correctly classify mental state concepts in each of the six groups into their respective classes. For instance, when presented with videos of *brooding* and *choosing*, the system should correctly label them as examples of *thinking*. The challenge that the test posed is that while the concepts share the semantic meaning of the group they belong to, they differ in intensity, in the underlying head and facial displays, and in the dynamics of these displays. Moreover, mental states concepts within a group are not equidistant. For

instance, within the thinking group, *choosing* and *judging* are closer to each other in meaning than *brooding* and *thoughtful*. The findings of this experiment can be used to refine the taxonomy of Baron-Cohen *et al.* [BGW+04]. For the six groups, 29 mental state concepts were included; each concept is represented by six videos from the DVD for a total of 174 videos. As of the time of this writing no other automated facial expression analysis system was evaluated against the fine shades of mental states.

8.1.1 Classification rule

Figure 8.1 shows the results of display recognition and mental state inference in a 6-second long video labelled as *undecided* from the Mind Reading DVD. *Undecided* belongs to the *unsure* group, so for the purposes of this analysis, its ground truth label is *unsure*. Throughout the video, a number of asynchronous displays that vary in duration are recognized: a head shake, a head tilt, a head turn, a lip pucker, and an eye-brow raise. The recognized displays affect the inferred mental states over time as shown in the figure. The strength s_i of a mental state i over the course of a video of T instances, is represented as the average in time of the area under the corresponding curve:

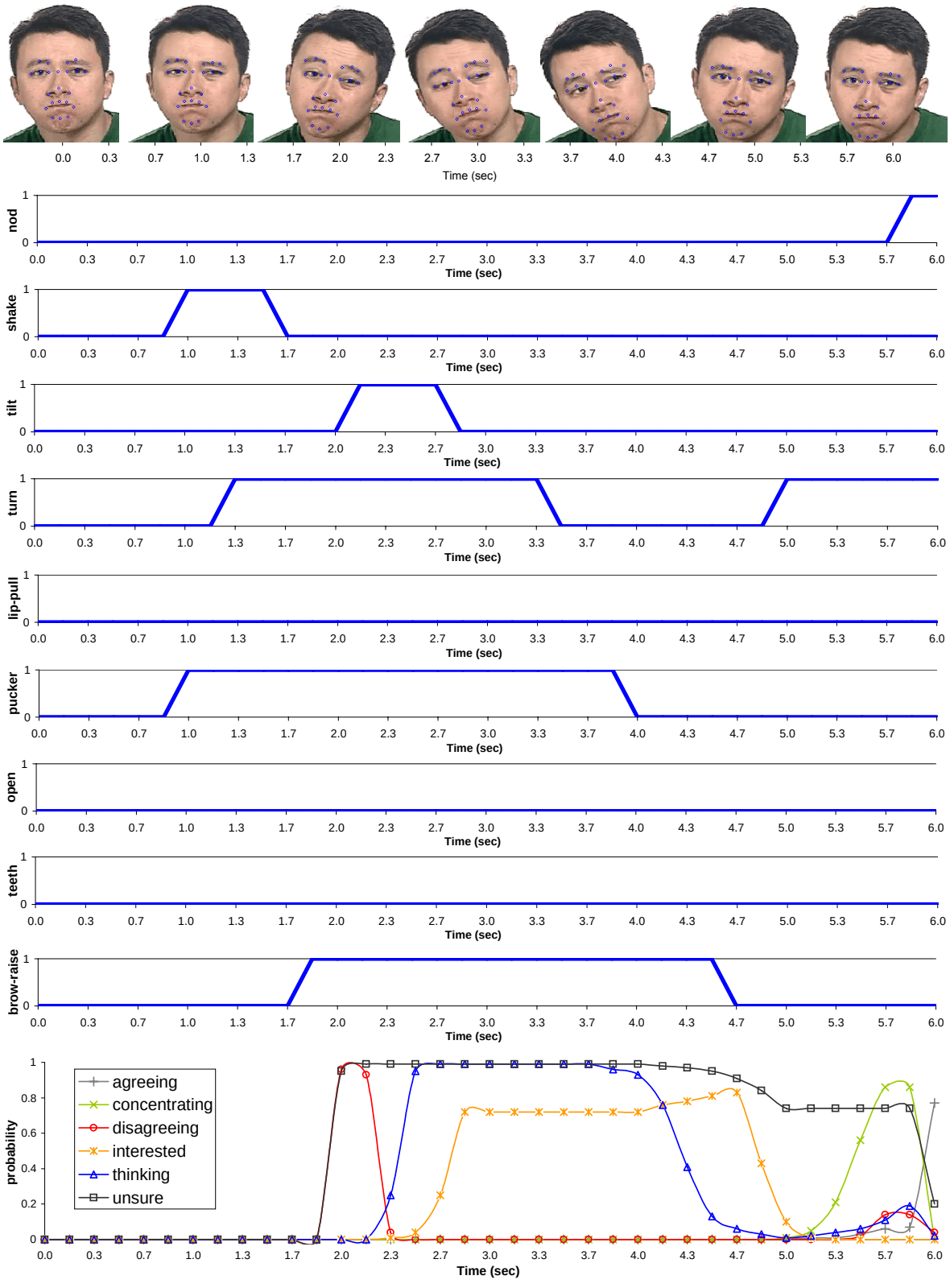
$$s_i = \frac{1}{T}\sum_{t=1}^{T} P(X_i[t]|\mathbf{Y}[1:t]) \tag{8.1}$$

Alternatively, an aggregate error value for that mental state is represented by $e_i = 1 - s_i$. This is the classification error introduced in the previous chapter. The aggregate error values for each of the six classes are shown at the bottom of Figure 8.1. The label assigned to a video by the system is that of the mental state scoring the minimum error, which, as explained earlier, also corresponds to the largest area under the curve. To account for the system's explicit representation of co-occurring mental state events, a video is assigned multiple labels if the corresponding errors are small enough, that is less than a threshold value of 0.4, determined empirically. If any of the labels that are assigned by the system match the ground truth label of the video, then this is deemed as a correct classification. If none of the labels match the ground truth, then the class with the least error is a false positive. In Figure 8.1, *unsure* is the mental state class with the least error and no other classes meet the threshold. Thus the system assigns only a single label—*unsure*—to that video. Since this label matches the ground truth label of the video, this is an example of a correct classification.

8.1.2 Results

Out of the 174 videos chosen for the test, ten were discarded because `FaceTracker` failed to locate the non-frontal face on the initial frames of the videos. I tested the system on the remaining 164 videos of 30 actors, which spanned 25645 frames or approximately 855 seconds. Using a leave-one-out methodology, 164 runs were carried out, where for each run the system was trained on all but one video, and then tested with that video. As explained above, for each run the system assigns to the video any of *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. If any of the results match the ground truth label of the test video, then this is a correct classification, otherwise it is not. Note that since this is effectively a six-way forced choice procedure, the probability of responding by chance is 16.7%. This would be the recognition rate of the system if it did not encode any useful information at all, and was merely picking a class at random.

Figure 8.2 shows the breakdown of videos in each of the six groups by mental state concept (row) and by actor (column). There is exactly one example of a person acting



	agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
error value e	0.96	0.90	0.91	0.61	0.52	0.11

Figure 8.1: Trace of display recognition and mental state inference in a 6-second video labelled as *undecided* from the Mind Reading DVD: (top) selected frames from the video sampled every one second; (middle) detected head and facial displays; (bottom) mental state inferences for each of the six mental state classes and corresponding table of errors, integrated from 2 to 6 seconds. Since the least error is *unsure* and *undecided* belongs to the *unsure* class, this is a correct classification.

any mental state concept in any class. For example, Figure 8.2(b) shows that 15 out of 16 actors pose exactly one video in *concentrating*. Y4 is the only actor contributing with more than one video; the three videos—*absorbed*, *concentrating* and *vigilant*—acted by Y4 differ considerably in their signature. About half of the actors contribute with more than one example per mental state group. Hence, in half of the test cases, this evaluation experiment is user-independent in the strict sense, that is, the person in the video will *not* have appeared in the training set of a class. In the other half of the test cases, a person will have appeared in the training set of a class, but will have acted a different mental state concept altogether. Even though this is not entirely user-independent, it is still considerably more challenging than tests in which a prototypic sequence of a person acting an emotion, say the smile of happiness, is included both in the training and in the test set.

The results are summarized in the confusion matrix and 3D bar chart in Figure 8.3. Row i of the matrix describes the classification results for mental state class i. Column i shows the number of times that mental state class i was recognized. The totals column gives the total number of videos that are labelled as i. The last column lists the true positive (TP) or classification rate for class i. It is given by the ratio of videos correctly classified as mental state i to the total number of videos that are labelled as i. The totals row yields the total number of videos that are classified as i by the system. The bottom row yields the false positive (FP) rate for class i, computed as the ratio of videos falsely classified as i to the total number of videos that are labelled as anything but i. In the 3D bar chart, the horizontal axis shows what the classification percentages are for videos that are labelled as i. The percentage that a certain mental state was recognized is described along the z-axis. Figure 8.3 shows that the classification rate is highest for *concentrating* (88.9%) and lowest for *thinking* (64.5%). The false positive rate is highest for *concentrating* (9.6%) and lowest for *disagreeing* (0.7%). For a mean false positive rate of 4.7%, the accuracy of the system is 77.4%.

For the 125 videos that were correctly classified, seven videos were assigned more than one label by the system, of which one matched the ground truth label. I analysed the multiple labels in each of these instances and found that it was plausible that these were cases of mental state co-occurrences. For example, *choosing*, which belongs to the *thinking* group, was assigned two labels: *thinking* and *unsure*. Arguably, being in a state of *choosing* means that a person is both *unsure* about a number of alternatives, and is *thinking* about which of these to select. Hence, I have scored this video as a correct classification of *thinking*. Further research on the labelling and classification of co-occurring mental states is necessary to more accurately analyse these cases.

To gain a better understanding of the distribution of errors within each class, I analyse the results at the finer level of mental state concepts. Figure 8.4 shows the classification results of the mental state concepts within each of the six groups of complex mental states. Although one should be careful about drawing conclusions out of these results because there are only six videos within each mental state concept, the results do show that some perform better than others in a class. For instance, while 86% of the videos labelled as *choosing* were correctly classified as belonging to the *thinking* class, only 25% of those labelled as *thoughtful* were classified as *thinking*. The results emphasize that mental state concepts within a class in the emotion taxonomy of Baron-Cohen *et al.* [BGW+04] are not by necessity equidistant and that further studies are needed to quantify these distances to determine how these mental state concepts and classes fit in a multi-dimensional state space.

Agreeing	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Assertive									•			•								•	•					•	•			
Committed									•			•										•					•		•	•
Convinced											•			•	•							•	•			•				
Knowing										•	•						•					•		•	•					
Persuaded												•	•				•					•				•	•			
Sure		•								•			•													•			•	•

(a) agreeing

Concentrating	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Absorbed												•	•				•	•								•			•	
Concentrating	•	•								•	•														•	•				
Vigilant														•	•				•			•	•			•				

(b) concentrating

Disagreeing	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Contradictory													•			•	•					•			•	•				
Disapproving									•	•							•			•			•					•		
Discouraging												•	•			•					•				•	•				
Disinclined									•	•									•			•				•			•	

(c) disagreeing

Interested	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Asking	•											•			•							•			•				•
Curious											•	•					•	•							•	•				
Fascinated									•			•				•					•	•					•		
Impressed		•	•						•								•			•				•						
Interested				•						•					•						•			•	•					

(d) interested

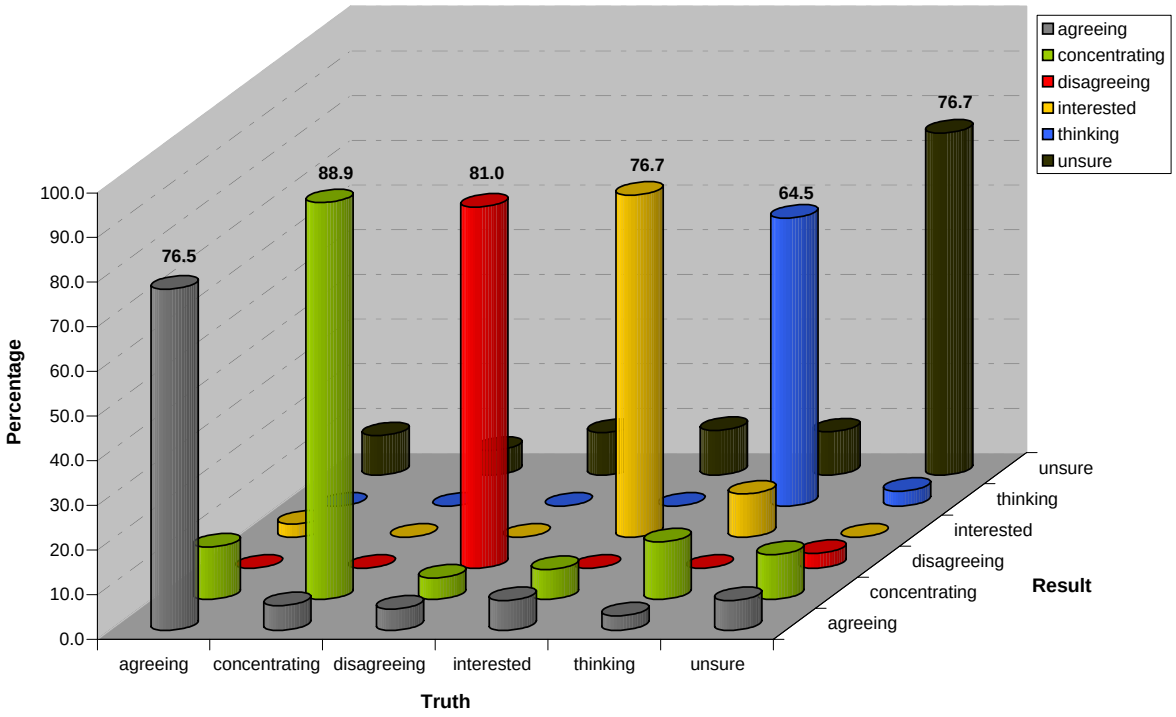
Thinking	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Brooding											•			•			•			•			•	•					
Choosing	•	•								•					•							•	•						
Fantasizing									•			•					•					•					•	•	
Judging											•			•						•	•			•	•				
Thinking		•		•								•					•						•	•					
Thoughtful	•			•						•	•										•						•		

(e) thinking

Unsure	C1	C2	C3	C4	C5	C6	C7	C8	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	S6	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Baffled									•					•		•			•										•	•
Confused							•	•				•									•				•	•				
Puzzled		•			•							•											•				•		•	
Undecided												•	•				•							•	•					•
Unsure	•	•											•	•							•						•			

(f) unsure

Figure 8.2: Breakdown of the six mental state groups by mental state concept (row) and by actor (column)



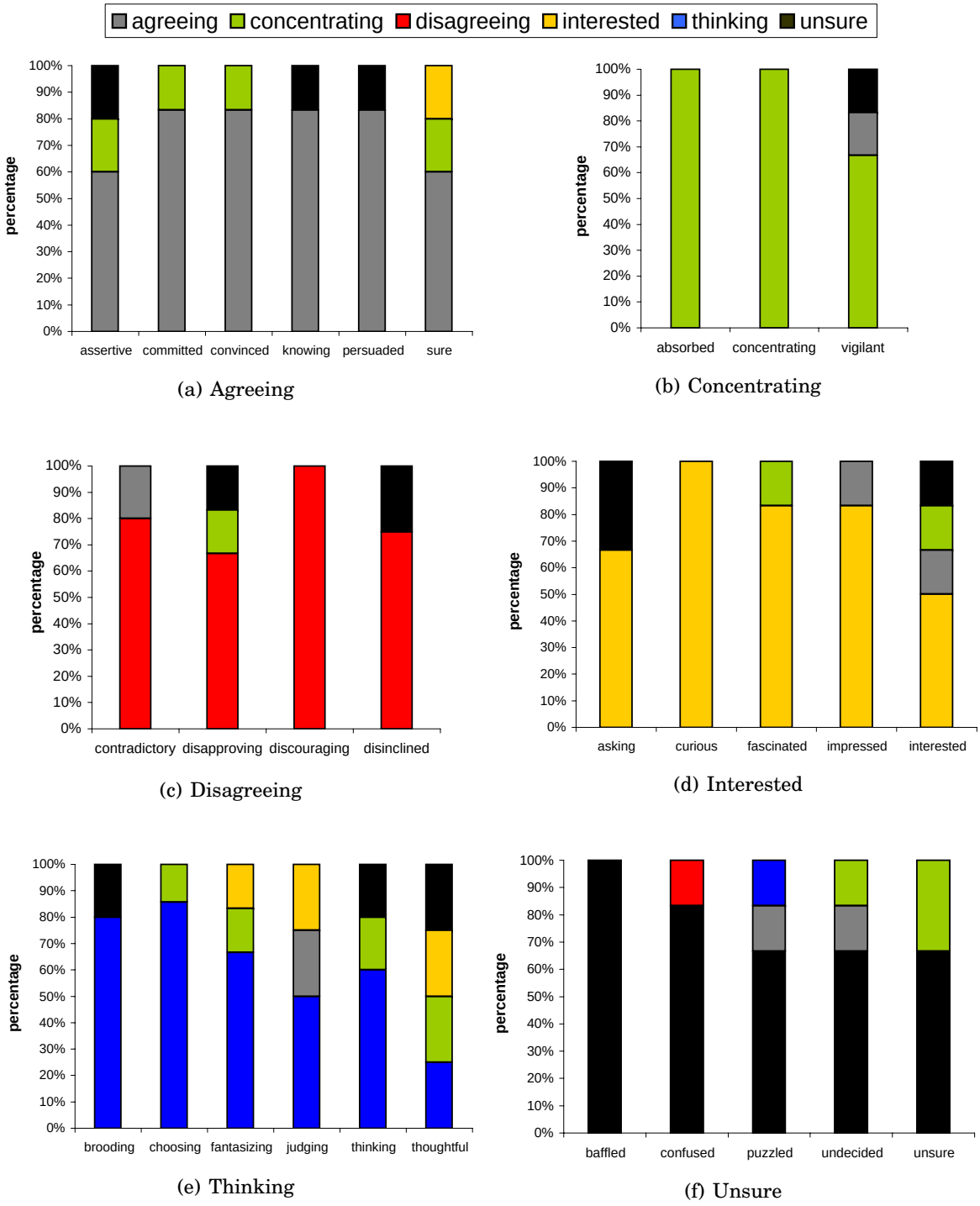
mental state	agreeing	concentr*	disagreeing	interesting	thinking	unsure	Total	TP %
agreeing	**26**	4	0	1	0	3	34	76.5
concentr*	1	**16**	0	0	0	1	18	88.9
disagreeing	1	1	**17**	0	0	2	21	81.0
interesting	2	2	0	**23**	0	3	30	76.7
thinking	1	4	0	3	**20**	3	31	64.5
unsure	2	3	1	0	1	**23**	30	76.7
Total	33	30	18	27	22	35	**164**	**77.4**
FP %	5.4	9.6	0.7	3.0	0.8	9.0	**4.7**	

Figure 8.3: Confusion matrix of **machine** recognition for the **Mind Reading DVD** shown as a 3D bar chart and corresponding table. The last column in the table shows the true positive (TP) rate for each class; the bottom row yields the false positive (FP) rate. For a false positive rate of 4.7%, the mean recognition accuracy of the system is 77.4%. * *concentrating* is abbreviated for space considerations.

8.1.3 Discussion

The mean accuracy of the system (77.4%) and the classification rates of each of the six classes are all substantially higher than chance (16.7%). It is not possible to compare the results to those of other systems since there are no prior results on the automated recognition of complex mental states. Instead I compare the results to those reported in the literature of automated recognition of basic emotions and to human recognition of complex mental states.

As shown in the survey in Chapter 2, the percentage accuracy of automated classifiers of basic emotions typically range between 80–90. Although this is higher than the results reported here, it is to be expected since the basic emotions are by definition easier to identify than complex ones. First, while basic emotions are arguably identifiable solely from facial action units, complex mental states additionally involve asynchronous information sources such as purposeful head gestures. Secondly, whereas basic emotions are identifiable from a small number of frames or even stills, complex mental states can only



(a) Agreeing

(b) Concentrating

(c) Disagreeing

(d) Interested

(e) Thinking

(f) Unsure

Figure 8.4: Evaluation of the Mind Reading DVD: classification results for the mental states within each class. Note that the probability of responding by chance is 16.7%.

be reliably discerned by analysing the temporal dependencies across consecutive facial and head displays. Thirdly, whereas basic emotions have distinct facial expressions that are exploited by automated classifiers, this is not the case with complex mental states, where single facial and head displays are only weak classifiers. Finally, in experimental settings as opposed to natural interactions, facial stimuli are limited to the facial region, and contextual cues are not available to the participants or the machine, making the classification problem more challenging.

I have already shown in the results from the experiments in Chapter 3 that human recognition of complex mental states from the Mind Reading DVD is lower than that of the classic basic emotions, and reaches an upper bound of 71% for videos from the Mind Reading DVD. Hence, at 77.4%, the results of the automated mind-reading system compares favourably to humans. Even though a direct comparison is not possible between the two experiments because the videos were not by necessity the same, they were nonetheless from the same corpora.

Aside from the difficulties inherent in recognizing complex mental states described in Chapter 3, using the Mind Reading DVD videos presents further challenges because, unlike other corpora used to evaluate such systems, the DVD was originally designed to be viewed by humans and not machines. In shooting the DVD, the actors were given the freedom to express the emotion labels in the ways they felt suitable, they were not shown video examples of the mental states being acted out and were allowed to move their heads freely.

Within-class variation

The stimuli used in training and evaluating existing automated facial analysis systems is confined to a single prototypical expression of an emotion class. In comparison, the videos from the Mind Reading DVD that represent each of the six mental state classes exhibit a lot of within-class variation. Each class includes a range of mental state concepts that differ in the way they are expressed. There is variation even within a mental state concept because the actors were not given any guidelines on *how* to act a mental state. Added to that, there is only one example of an actor posing a mental state.

The result is a set of videos within a class that vary along several dimensions including the specific mental states they communicate, the underlying displays and their dynamics and the facial physiognomies of the actors. A video that varies substantially compared to other videos in the class along any of these dimensions may end up being misclassified. For instance, as shown in Figure 8.4(a), only 60% of the videos labelled as *assertive* were correctly classified as *agreeing*. The rest were misclassified as *concentrating* or *unsure* since the underlying displays did not contain a head nod or a lip-corner pull (a smile), the most frequently observed displays in *agreeing*.

The evaluation results largely depend on the specific concepts that are picked for training and testing in each class and how different are their underlying displays. When the mental state concepts that share the underlying head/facial displays are the only ones picked for training and testing the system, the results reported are much higher. For example in el Kaliouby and Robinson [KR04c], the total number of videos used in a six-way classification was 106, and covered 24 mental state concepts; the results were higher than those reported here, reaching a mean accuracy of 89.5%.

Uncontrolled rigid head motion

Most FER systems of basic emotions rely solely on facial expressions for classification and do not support the recognition of head gestures. Accordingly, the stimuli used in evaluating these systems is typically restricted in terms of rigid head motion. In contrast, the Mind Reading DVD had no restrictions on the head or body movements of the actors. Hence, the resulting head gestures and facial expressions are natural, even if the mental state is posed, and contain in-plane and out-of-plane head motion. Rigid head motion adds complexity to head and facial action recognition; the development of automated facial analysis systems that are robust to substantial rigid head motion is the subject of interest of many vision researchers.

Noisy evidence

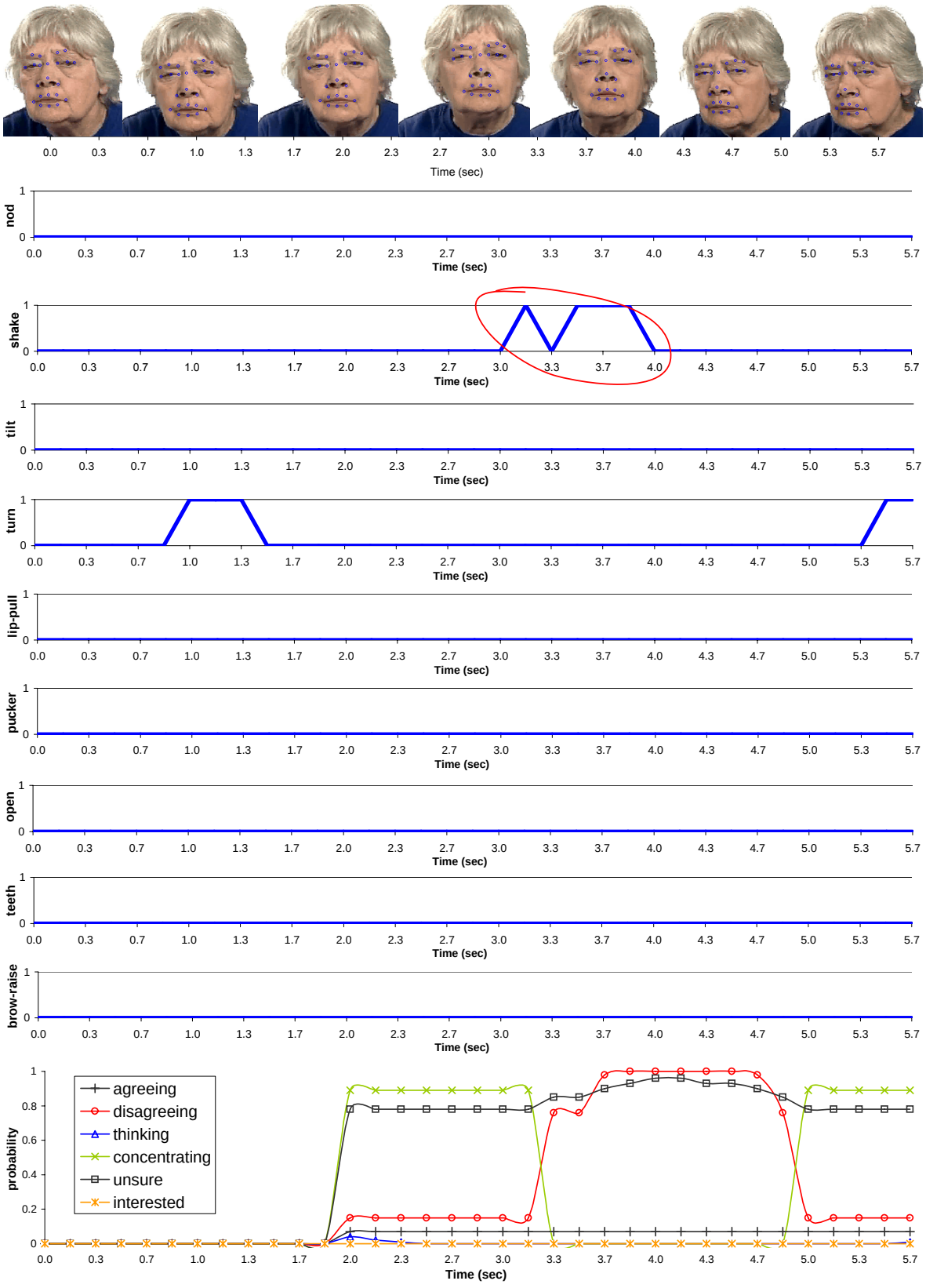
The experimental evaluation in Chapter 6 has shown that the HMM classifiers of head and facial displays are imperfect: displays may be misclassified or undetected by the system. Both cases result in incorrect evidence being presented to the mental state DBNs. Depending on the persistence of the erroneous evidence, its location within the video and its discriminative power, the resulting inferences may be incorrect.

Figure 8.5 shows an example of misclassification due to noisy evidence. The 5.7 second video is labelled as *vigilant*, which belongs to the *concentrating* class. The mental state inferences start with a high probability of *concentrating* and *unsure*. Due to a sideways movement of the woman's body, a head shake is falsely detected at 3.0 seconds and persists for one second. This causes the probability of *concentrating* to drop to zero and the probability of *unsure* to increase. The noisy evidence eventually leads to the (mis)classification of this video as *unsure* instead of *concentrating*. Note that despite the head shake, the video is not classified as *disagreeing*. This is consistent with the detailed classification results for *vigilant* in Figure 8.4: out of the six videos labelled as *vigilant* two were misclassified, one as *unsure* and the other as *agreeing*.

Consider the results of this example—*concentrating* incorrectly classified as *unsure*—in the light of how the Mind Reading DVD videos were given their labels. Recall from Chapter 3 that the process of labelling the videos on the Mind Reading DVD involved a panel of 10 judges who were asked if, for instance, the video in Figure 8.5 is a good enactment of *vigilant*. When 8 out of 10 judges agreed, a statistically significant majority, the video was included in the corpus. In a separate test, which did not involve viewing any videos, a lexical assignment of mental state words to groups was carried out, for instance assigning the mental state *vigilant* to the *concentrating* group. In other words, none of the videos on the Mind Reading DVD including the one in Figure 8.5, were directly labelled in terms of the 24 groups in the emotion taxonomy of Baron-Cohen *et al.* [BGW+04]. Hence, it is unclear if given the choice, people would classify this video as *concentrating* or *unsure*. Using the results of the automated mind-reading system as a starting point, further studies can be carried out to verify and further refine emotion taxonomies, particularly that of Baron-Cohen *et al.* [BGW+04].

8.2 Generalization to the CVPR 2004 corpus

When the training and testing are both done on the same corpus, even when a leave-x-out methodology is used as in the previous section, a bias is introduced by the similarity of recording conditions and actors of the videos in that corpus. This combined with locally optimal training algorithms, such as Maximum Likelihood Estimation, may lead



	agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
error value e	0.93	0.50	0.51	1.0	0.99	0.17

Figure 8.5: Incorrect classification due to noisy evidence: (top) selected frames—sampled every one second—from a 5.7 second video labelled as *vigilant* from the Mind Reading DVD (*vigilant* belongs to the *concentrating* class); (middle) detected head and facial displays, including a false head shake (circled in red); (bottom) mental state inferences for each of the six mental state classes and corresponding table of errors, integrated from 2 to 6 seconds. Note the effect of the false head shake on decreasing the probability of *concentrating*. The video is (mis)classified as *unsure* instead of *concentrating*.

to over-fitting. In over-fitting, the exact model recovered is partially dependent both on the size of the training set and on the randomness in the initialization procedure. The result is that the classifier is optimized for one corpus but transfers poorly to situations in which expressions, users, contexts, or image properties are more variable.

Generalization is an important predictor of the system's performance in a natural computing environment. This is because the better the generalization ability of the system, the more feasible it is to train the system on some data-set then deploy it in different interaction scenarios, with many users, without having to re-train or calibrate the system. Evaluating the generalization performance of a system however, is an expensive task, especially in vision-based systems where the collection, filtering and labelling of videos is time-consuming task. As a result, most automated facial analysis systems are evaluated on a single corpus of images or video using cross-validation or bootstrapping; and the generalization ability of these systems remains unknown.

The objective of this evaluation experiment was to assess the generalization ability of the mental state models when trained on the Mind Reading DVD videos, and tested on a previously unseen corpus—the CVPR 2004 corpus. This is a very different experiment from that presented in the previous section. In the previous section, the videos used to train and to test the system were of people trained to pose the different mental states. In this experiment, the system was still trained on "good" examples, but was tested on videos that were posed by untrained people. In the CVPR 2004 corpus, 16 volunteers from the CVPR 2004 conference acted all six mental states for a total of 96 videos[1].

8.2.1 Human baseline

Compared to the Mind Reading DVD videos used to train the system, the videos in the CVPR 2004 corpus were not posed by professional actors. As a result, the videos are likely to include incorrect or bad examples of a mental state, and on the whole are weakly labelled.

It would have been possible to use expert judges to label and filter the CVPR 2004 corpus for bad or incorrect examples. Instead, I decided to include all the videos collected at the conference in the evaluation, but tested how well a general population of amateurs would classify them. The labelling experiment was done several months after the CVPR 2004 corpus was recorded in Washington DC. None of the participants who were asked to view and classify the videos knew when or where these videos were recorded, and they did not recognize any of the actors in the CVPR 2004 corpus. The results from this human population were used as a baseline with which to compare the results of the automated mind-reading system.

A forced-choice procedure was adopted for the experiment. There were six choices on each question: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The procedure was as follows:

- Participants were first asked to go through the six mental state word choices and were encouraged to inquire about any word meanings they were unclear about. A clarification was made that *unsure* means that the actor in the video is *unsure*, not that the participant is unsure of his/her answer.

[1]This experiment is user-independent by definition because none of the volunteers in the CVPR 2004 corpus were actors in the Mind Reading DVD.

- Participants were shown a video on a standard multimedia player and projection screen, and then asked to circle the mental state word that they thought best matched what the actor in the video was feeling. They were told that there was only one correct answer for each question, and were asked to provide an answer for all the questions.
- Participants were encouraged to request as many replays as they deemed necessary to properly identify the mental state.

A total of 18 participants (50.0% male, 50.0% female) between the ages of 19 and 28 took part in the experiment. The participants were company employees, mostly software developers. All participated on a voluntary basis.

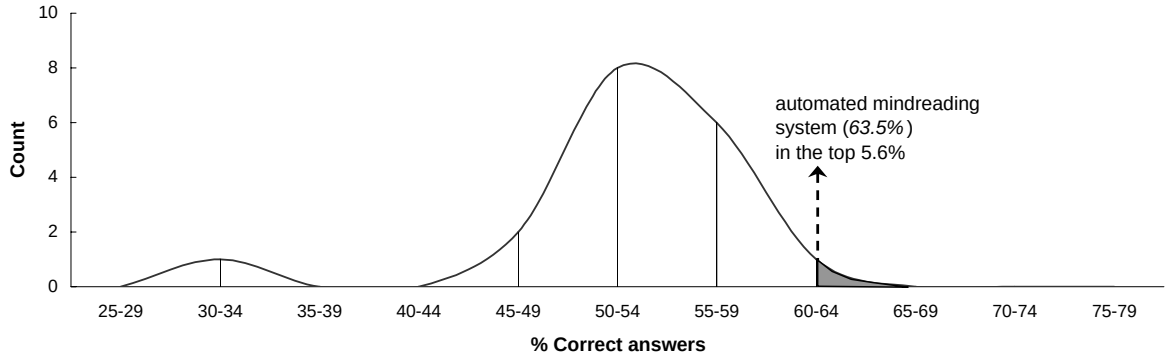


Figure 8.6: Distribution of responses in classifying the CVPR 2004 corpus videos of a general population of 18 people. The result of the automated mind-reading system is shown for comparison.

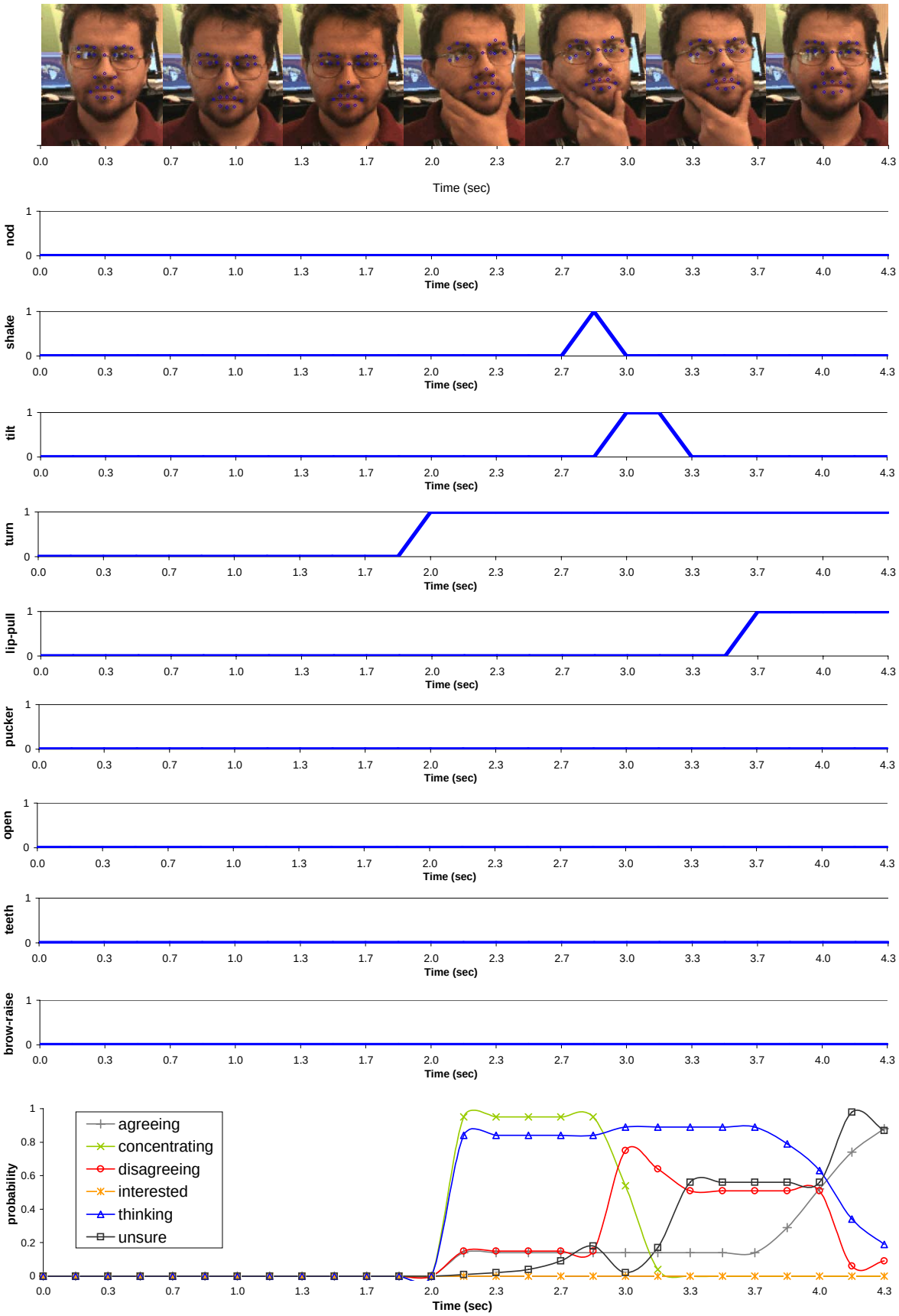
The test generated 88 trials for every participant for a total of 1584 responses. The distribution of results is shown in Figure 8.6. The responses of the participants range from 31.8% to 63.6% (mean=53.0%, SD=6.8) of correct answers. A confusion matrix of all 1584 responses is shown in Figure 8.8. The classification rate is highest for *disagreeing* (77.5%) and lowest for *thinking* (40.1%). The false positive rate is highest for *concentrating* (11.8%) and lowest for *thinking* (5.0%). For a false positive rate of 9.4%, the mean recognition accuracy of humans was 54.5%.

The total number of correct responses, where a correct response is one that matches the label of a video, ranges from 100% to 0% over the entire set. The questions on which none of the responses matched the label of the video, that is, 0% of the answers were correct, suggests that some of the videos are inaccurate enactments of a mental state.

8.2.2 Results

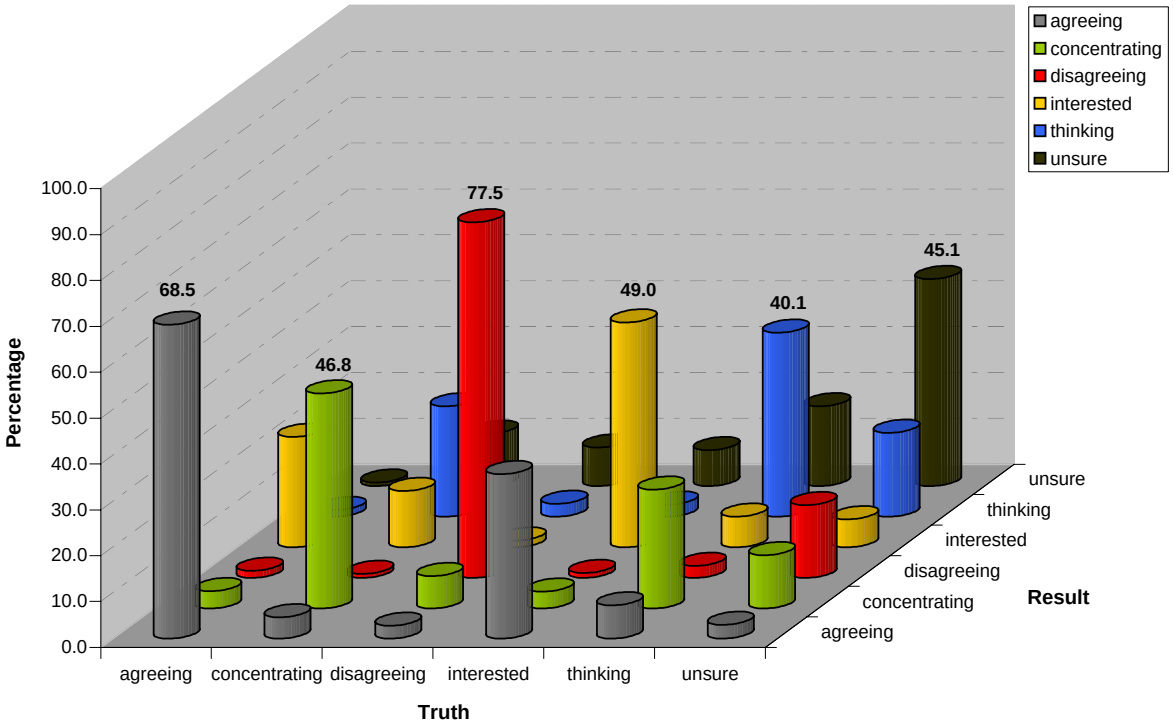
The six mental state DBNs were trained on the entire set of videos picked from the Mind Reading DVD. The system was then tested on each of the videos from the CVPR 2004 corpus. Out of the 96 videos, 8 were discarded: three videos lasted less than two second, which is when the first DBN invocation occurs, and `FaceTracker` failed to locate the face in 5 videos. I therefore tested the system on the remaining 88 videos (9380 frames or approximately 313 seconds). Figure 8.7 shows a trace of mental state inference in a 4.3-second long video labelled as *thinking*. This is an example of a correct classification because *thinking* scored the least error (and also meets the threshold).

The results are summarized in the confusion matrix and 3D bar chart in Figure 8.9. The classification rate of the system is highest for *disagreeing* (85.7%) and lowest for



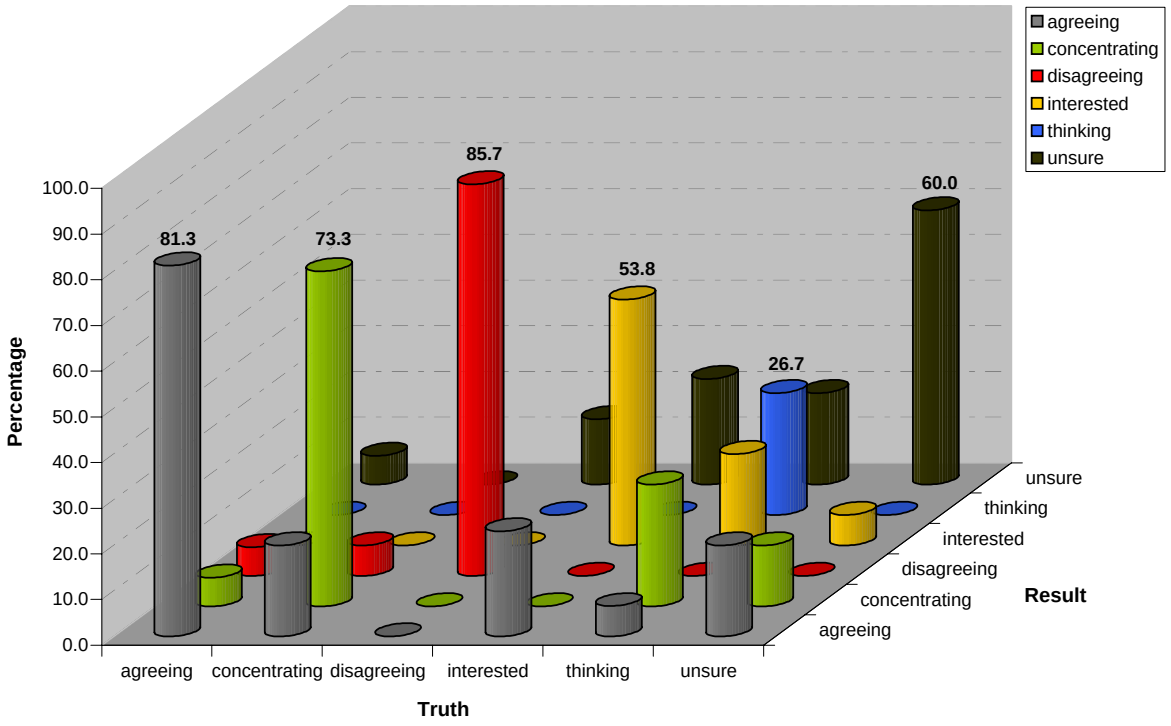
	agreeing	*concentrating*	*disagreeing*	*interested*	*thinking*	*unsure*
error value e	0.73	0.62	0.65	1.0	0.24	0.63

Figure 8.7: Correct classification: (top) selected frames—sampled every 0.7 seconds—from a 4.3 second video labelled as *thinking* from the CVPR 2004 corpus; (middle) detected head and facial displays; (bottom) mental state inferences and corresponding table of errors, integrated from 2 to 4.3 seconds. The most likely class—given by the least error—is *thinking*.



mental state	agreeing	concentr*	disagreeing	interested	thinking	unsure	Total	TP %
agreeing	**185**	10	4	65	4	2	270	68.5
concentr*	11	**111**	2	29	57	27	237	46.8
disagreeing	6	15	**165**	3	6	18	213	77.5
interested	69	7	2	**94**	5	15	192	49.0
thinking	25	89	9	23	**139**	60	345	40.1
unsure	10	38	52	20	60	**148**	328	45.1
Total	306	270	234	234	271	270	**1585**	**54.5**
FP %	9.2	11.8	5.0	10.1	10.6	9.7	9.4	

Figure 8.8: Confusion matrix of **human** recognition for the **CVPR 2004 corpus** shown as a 3D bar chart and corresponding table. For a false positive rate of 9.4%, the mean recognition accuracy of the participants is 54.5%. * *concentrating* is abbreviated for space considerations.



mental state	agreeing	concentr*	disagreeing	interested	thinking	unsure	Total	TP %
agreeing	**13**	1	1	0	0	1	16	81.3
concentr*	3	**11**	1	0	0	0	15	73.3
disagreeing	0	0	**12**	0	0	2	14	85.7
interested	3	0	0	**7**	0	3	13	53.8
thinking	1	4	0	2	**4**	4	15	26.7
unsure	3	2	0	1	0	**9**	15	60.0
Total	23	18	14	10	4	19	88	**63.5**
FP %	13.9	9.6	2.7	5.3	0.0	12.3	7.3	

Figure 8.9: Confusion matrix of **machine** recognition for the **CVPR 2004 corpus** shown as a 3D bar chart and corresponding table. For a false positive rate of 7.3%, the mean recognition accuracy of the system is 63.5%. * *concentrating* is abbreviated for space considerations.

thinking (26.7%). The false positive rate is highest for *agreeing* (13.9%) and lowest for *thinking* (0.0%). For a mean false positive rate of 7.3%, the mean accuracy of the system is 63.5%, which is higher than chance (16.7%).

Compared to the results obtained when evaluating the Mind Reading DVD (Figure 8.3), the results of the automated mind-reading system are on average 13.9% less. Notably though, *agreeing* and *disagreeing* perform better with the CVPR 2004 corpus, increasing 4.8% and 4.7% respectively. *Thinking* however, performs substantially worse, dropping 37.8%. The statistical significance of these differences has not been analysed.

Compared to the results of humans classifying the exact set of videos, the automated mind-reading system scores in the top 5.6% of humans, and 10.2% better than the mean accuracy reported in the sample of 18 people. The result of the system is superimposed on the normal distribution of human responses shown in Figure 8.6.

In addition to providing a baseline for comparison to the automated mind-reading system, the classification results reported by the panel of 18 amateurs can also be used to rigorously re-label the CVPR 2004 corpus videos. Similar to the process used to label the Mind Reading DVD, if a majority of the judges agree on a label for a video, then this label is adopted as the ground truth label for that video. A majority of 85% of this panel agreed on the label of 11% of the videos; these videos were deemed as "good" examples of mental state enactments. The system's recognition accuracy of these videos is 80%. This result emphasizes that the system's generalization performance is comparable to that of humans, and that the system generalizes well to new examples of mental states, assuming that they are reasonably well acted.

8.2.3 Discussion

When tested on the CVPR 2004 corpus videos, the automated mind-reading system scores in the top 5.6% compared to a group of people classifying the same set of videos. The accuracy (63.5%) however, is lower than that reported in evaluating the Mind Reading DVD videos (77.4%). There are a number of reasons for this drop in accuracy. The principal reason can be attributed to the unprofessional acting and weak labelling of the CVPR 2004 corpus videos, compared to the Mind Reading DVD videos on which the system was trained. Notably, the actors on the Mind Reading DVD were mostly British, while the volunteers for the CVPR 2004 corpus were from a more diverse audience of countries and ethnic backgrounds. This may have resulted in cultural differences in the expression of a mental state. In addition, the recording conditions of the CVPR 2004 corpus were much less controlled than those of the Mind Reading DVD, resulting in a number of technical challenges in processing these videos.

Figure 8.10: The recording conditions of the CVPR 2004 corpus were relatively uncontrolled in terms of background, distance from the camera and lighting setup.

To start with, the videos were recorded at the CVPR 2004 conference at a demonstration booth, which was co-located with other demonstrations. As shown in Figure 8.10, the background of the videos is dynamic and cluttered since people were moving in and out of the neighbouring demonstration booth. The distance from the camera varies across videos. It even varies over the course of a video as an "actor" moves toward/away from the camera. The videos were shot using only the lighting in the conference room. To the contrary, the videos on the Mind Reading DVD all had a uniform white background, the faces were recorded at a constant distance from the camera, and the lighting was professionally set up. These factors reduce the accuracy of shape and colour analysis used in the recognition of head and facial actions.

Figure 8.11: Some volunteers had a non-frontal pose and did not look into the camera.

In terms of pose, several volunteers maintained a non-frontal pose throughout the recording session and several were also looking down at the instructions, rather than at the camera (Figure 8.11). In terms of facial physiognomies, three volunteers had a moustache/beard, while one volunteer had his glasses on. In comparison, almost all of the actors in the Mind Reading DVD had a frontal pose, and none had a moustache/beard or wore glasses. These variation in pose and facial physiognomies make the feature point tracking less accurate.

Figure 8.12: Two volunteers were talking throughout the videos. Notice the difference in the size of the face within the image across subjects.

In the instructions handed out to the CVPR volunteers, there was no mention of speaking. However, three volunteers asked if they could speak and were told they could (Figure 8.12). In comparison, none of the actors on the training videos from the Mind Reading DVD were talking. This meant that mouth actions were interpreted by the system as affective displays even when they were speech-related, resulting in a number of misclassifications. The current implementation of the automated mind-reading system—like most existing automated facial expression recognition systems—assumes that no speech is present, since speech introduces an additional complexity of deciding whether the deformation of the mouth is speech-related or emotion-related or both. Since speech is a natural component of social interaction, research into features that are robust to speech but are sensitive to emotional expressions is necessary.

Figure 8.13: Expression through modalities other than the face (from left to right): volunteer is scratching the chin in *thinking*; scratching the head in *unsure*; shoulder shrug in *unsure*.

Finally, in some videos, the volunteers expressed a mental state through modalities in addition to the face. For example, Figure 8.13 shows an example of a chin scratch in *thinking*, head scratch in *unsure* and shoulder shrug in *unsure*. While those non-facial expressions may be important indicators of the underlying mental states that people readily use when mind-reading, the current version of the automated mind-reading system does not support these modalities.

8.3 Real time performance

Most user interfaces require real time responses from the computer: for feedback to the user, to immediately execute commands, or both. Although there is no universal definition for real time, in an HCI context real time pertains to a system's ability to respond to an event without a noticeable delay [TK04]. The opposite phenomena, referred to as lag, is when there is a delayed response to some input from the user [Mac95]. Of course, whether or not some delay is noticeable or acceptable is dependent on the application context in which the delay is measured. For instance, in a mouse-based target acquisition task, a lag of 75 ms is easily noticeable by the users of the system; at a lag of 225 ms, the error rate of humans finding the correct target increases by 214% [MW93].

Operating in real time is a pre-requisite to an automated mind-reading system that is expected to be used with applications that adapt their responses depending on the inferred mental state of the user. For instance, it is pointless for an application to respond to a confused user long after she is no longer experiencing this mental state. In order to support intelligent HCI, mental state inferences need to be made early enough after their onset to ensure that the resulting knowledge is current.

8.3.1 Objectives

The objective of this analysis is to quantify the real time performance of the automated mind-reading system, and gain an insight into how it scales as more actions, displays and mental state classes are implemented. The throughput and the latency are typically used to quantify the real time performance of a vision-based system [TK04].

The **throughput** is the number of events that are processed per unit time. For the automated mind-reading system, the throughput translates to the number of mental state inferences made in a second. The **latency** is defined as the time elapsed, or delay, between the onset of an event and when the system recognizes it. For the automated mind-reading system, the latency is the difference between the instant a frame is captured and the time when the system infers the mental state.

Table 8.1: The processing time at each level of the automated mind-reading system measured on an Intel Pentium® 4, 3.4 GHz processor.

level	tracking	action-level	display-level	mental state-level	**total**
time (ms)	3.00	0.09	0.14	41.10	**44.33**

8.3.2 Results

The system was tested on an Intel Pentium® 4, 3.4 GHz processor with 2 GB of memory. The processing time at each of the levels of the system is summarized in Table 8.1. The details are as follows (note that the code has not been optimized for speed):

- **Live tracking**: `FaceTracker` runs at an average of 3.0 ms per frame using a Video for Windows (VfW) or DirectX compatible video capture device supporting the capture of 320x240 interlaced frames at 30 fps.
- **Head and facial actions**: The time taken to extract head pitch, yaw and roll actions was averaged over 180 calls to each of these functions. On average, each function call took 0.022 ms per frame depending on the amount of head motion in the frame. The time taken to extract mouth, lip and eyebrow actions was also averaged over 180 calls to each of the functions. On average, each function call lasted 0.010 ms per frame, subject to the amount of feature motion in the frame. In total, this level of the system executes at 0.096 ms per frame.
- **Head and facial displays**: The time taken to compute the probability of a head/facial display given a vector of head/facial action symbols was averaged over 180 invocations of the HMM inference. On average, a call to the HMM inference lasts 0.016 ms, and is incurred once every five frames. Since there are currently nine displays implemented so far, this level of the system executes at 0.140 ms every five frames.
- **Mental states**: I tested the performance of DBN classification on Intel's recently released Probabilistic Network Library (PNL) [PNL03]. The implementation of fixed lag smoothing of the six previous inferences using unrolled junction tree inference for a DBN with an average of seven nodes takes 6.85 ms per slice. This varies with the size of lag. For instance, if only the three previous inferences are considered this number drops to 4.46 ms. Hence, this level executes in 41.1 ms for the six classes of complex mental states.

8.3.3 Discussion

To be deemed as real time the throughput of the automated mind-reading system has to be at least equal to the input video. For video captured at 30 fps, and using a sliding window that moves 5 frames per inference, the system needs to be able to perform six inferences per second. On the machine configuration tested, the system would theoretically be able to run at 22 inferences per second. At six inferences per second this roughly translates into 27% of the processing power of an Intel Pentium® 4 3.4 GHz processor.

The extraction of head and facial actions and displays all run in linear time. The processing time increases by 0.016 ms with each head or facial action or display added to the system. At the mental state level, inference algorithms such as the unrolled junction tree algorithms run in polynomial time in the number of nodes [GH02].

8.4 Summary

In this chapter, I presented the results of evaluating the automated mind-reading system in terms of accuracy, generalization and speed. It was not possible to compare the results of this work to those of other researchers because there are no prior results on the automated recognition of complex mental states. Instead, the results were compared to human recognition on similar recognition tasks.

In terms of accuracy, the system achieves an upper bound of 88.9%, and a mean accuracy of 77.4% when tested on videos from the Mind Reading DVD that included 29 different mental state concepts. I decided to include many mental state concepts in the test because the recognition of fine shades of a mental state is an indication of social intelligence skills in humans. To evaluate the generalization ability of the system, the system was trained on the Mind Reading DVD and tested on videos from the CVPR 2004 corpus. The results compare favourably to human classification of the same set of videos. In terms of real time performance, the processing times of each level was measured to determine the overall throughput and latency of the system.

On the whole, the results show that the automated mind-reading system is successfully able to classify a pre-defined set of mental state classes, and generalize to new examples of these classes with an accuracy and speed that is comparable to that of human recognition. The experimental results also suggest that, similar to humans, the accuracy and reliability of the system can be improved through the addition of context cues, a direction for future work.

Chapter 9

Conclusion

This chapter first describes the principle contributions of this research, followed by several directions for future work, before summarizing the dissertation by summing up the progress it has made towards the development of socially and emotionally intelligent interfaces.

9.1 Contributions

This dissertation addresses the problem of automated inference of complex mental states—the group of affective and cognitive states of the mind that are not part of the basic emotions set—from the face. This is a challenging endeavour because of the uncertainty inherent in the inference of hidden mental states from the face, because the automated analysis of the face is an open machine-vision problem, and because there is a lack of knowledge about the facial expression composition of complex mental states.

In the light of these challenges, the dissertation makes four principal contributions. First, the dissertation describes a computational model of mind-reading as a novel framework for machine perception and mental state inference. The second contribution is that the research undertaken here particularly focuses on complex mental states, which advances the state-of-the-art in affective computing beyond the basic emotions. Third, this dissertation has emphasized a working prototype of a mental state inference system that runs in real time and is thus suited for application to human-computer interaction. Finally, the dissertation presents a body of knowledge about the configurations and dynamics of facial expressions of complex mental states, which have the potential to inform future work in face perception and emotions. Each of these contributions are further discussed in this section.

9.1.1 Novel framework for mental state inference

The computational model of mind-reading is a novel framework for machine perception and mental state recognition. The model describes a coherent framework for fusing low-level behaviour in order to recognize high-level mental state concepts. It is defined as a multi-level probabilistic graphical model that represents a raw video stream at three levels of abstraction: actions, displays and mental states. Each level of the model captures a different degree of spatial and temporal detail that is determined by the physical properties of the facial event at that level. The framework works well because

its hierarchical, probabilistic architecture is a good match to the attributes of complex mental states. The multi-level representation mimics the hierarchically structured way with which people perceive facial behaviour. It also accounts for the inter-expression dynamics that, through several studies, I have found to improve human's recognition of complex mental states. The top-most level of the model is implemented using DBNs. These classifiers allow multiple asynchronous observations of head and facial displays to be combined within a coherent framework, and provide a principled approach to handling the uncertainty inherent in mental state inference. The output probabilities and their development over time represent a rich modality analogous to the information humans receive in everyday interaction through mind-reading.

9.1.2 Beyond the basic emotions

The application of automated facial expression analysis to human-computer interaction is limited to primitive scenarios where the system responds with simple positive or negative reactions depending on which basic emotion the user is expressing. This is because basic emotions are of limited utility in understanding the user's cognitive state of mind and intentions. The automated inference of complex mental states is a significant step forward from existing facial analysis systems that only address the basic emotions. Recognizing mental states beyond the basic emotions widens the scope of applications in which automated facial expressions analysis can be integrated, since complex mental states are indicators of the user's goals and intentions.

9.1.3 Real time mental state inference system

The automated mind-reading system combines bottom-up vision-based processing of the face with top-down predictions of mental state models to interpret the meaning underlying head and facial signals. The system executes in real time, does not require any manual preprocessing, is user independent, and supports natural rigid head motion. These characteristics make it suitable for application to HCI. The classification accuracy, generalization ability, and real time performance of the system were evaluated for six groups of complex mental states—*agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The videos of these mental states were sampled from two different corpora—the Mind Reading DVD and the CVPR 2004 corpus. The results show that the automated mind-reading system successfully classifies the six mental state groups (and corresponding 29 concepts), generalizes well to new examples of these classes, and executes automatically with an accuracy and speed that are comparable to that of human recognition.

9.1.4 Facial expressions of complex mental states

The research described throughout this dissertation provides an insight into the configuration and dynamics of facial expressions in complex mental states, which is lacking in the literature. The findings from the studies in Chapter 3 have shown that complex mental states are often expressed through multiple head gestures and facial expressions, which may occur asynchronously. On their own, these displays are weak classifiers of the corresponding mental states. The studies have also shown that the incorporation of inter-expression dynamics, or previous facial events, boosts the recognition results of complex mental states, and that only a relatively small amount of facial event history accounts for most of the improvement. In Chapter 7, the converged parameters of the DBNs were used to explore the statistical truth of videos from the Mind Reading DVD. The findings of this analysis confirmed the weak discriminative power of individual displays, and identified the facial configurations of complex mental states.

9.2 Future directions

The ultimate goal of this research is to integrate automated mind-reading systems with human-computer interfaces. This objective motivates four exciting areas for future work that are extensions of the research presented throughout this dissertation. The first is boosting the accuracy and robustness of the mind-reading system. This means extending the computational model of mind-reading, and its implementation, to incorporate more evidence from the user's behaviour and surrounding context in a powerful learning and inference framework. The second direction is generalizing the system to a natural corpus of mental states collected from naturally evoked scenarios of human-machine interactions. The third direction deals with the conceptualization, implementation and validation of applications that use the mental state inferences produced by the automated mind-reading system to adapt their responses to the user. Finally, the fourth direction discusses implications of this work for research in psychology and sociology that is concerned with the facial expressions of complex mental states and with emotion taxonomies. Each of these directions are further discussed in this section.

9.2.1 Extend the computational model of mind-reading

People express their emotions and mental states through many nonverbal communication channels, of which the face is only one. Other modalities that carry nonverbal information include voice nuances, changes in posture, and affect-related hand-gestures such as head or chin scratching. The messages communicated by the different modalities often complement each other; may substitute for each other when only partial input is available; and occasionally contradict one another as in deception [SKR04]. Compared with unimodal systems that assume a one-to-one mapping between an emotion and a modality, multi-modal systems yield a more faithful representation of the intricate relationship between internal mental states and external behaviour.

Figure 9.1: Asymmetry in mouth displays selected from the CVPR 2004 corpus videos.

The computational model of mind-reading that I have presented in this dissertation currently supports a subset of head and facial displays as signals of affective and cognitive mental states. One natural extension is to support more modalities and context cues—in the face and beyond—to improve the recognition power of the system. Facial displays that are of immediate interest include the movement of the head towards or away from an object, a head dip, a lip bite, a lip purse and a frown, as well as asymmetric facial actions. Figure 9.1 shows examples of asymmetry in several mouth displays from the CVPR 2004 corpus.

Besides head and facial displays, the eyes and the direction of gaze play a crucial role in the communication of one's mental state and intention. Figure 9.2 shows a selection of frames from the Mind Reading DVD and the CVPR 2004 corpus in which the direction of eye gaze conveys a state of *thinking*. By integrating eye-gaze tracking with mental state inference, it will be possible to link mental states to intentions, making better sense of the user's current and future behaviour.

In Sobol-Shikler *et al.* [SKR04] we draw attention to the various issues inherent in building a multi-modal system for the recognition of a range of user mental states. Integrating different modalities in the computational model of mind-reading poses many research challenges with respect to building sensors and classifiers of the individual modalities, and developing a coherent model that integrates these modalities efficiently. The explicit representation of missing/partial data due to the absence of a modality-sensor or due to occlusion of the face or other parts of the body is also an interesting direction for future research.

Figure 9.2: Eye gaze in complex mental states from CVPR 2004 corpus videos.

An equally important information channel to integrate within the computational model of mind-reading is that of context. Numerous studies show how humans make considerable use of the contexts in which expressions occur to assist interpretation [ABD00, BY98, FDWS91, Fri97, Rat89, Wal91]. Context may include the time and location of an interaction, personal information about the user, the history of an interaction, information about the current application, the active task, the connected networks and available resources. Being probabilistic, hierarchical and modular, the model is particularly suited to fusing these modalities and context cues.

In terms of learning and inference, online Bayesian learning can be explored with the objective of increasing the predictive power of the model as additional evidence nodes are incorporated. The challenge in using online Bayesian learning is that dealing with a distribution over the parameters is not as simple as finding a single "best" value for the parameters; approximate methods such as Markov Chain Monte Carlo methods [Gam97], expectation propagation [Min01] and variational approximations [WB05] may have to be used. These methods are often computationally expensive, and typically do not run in real time. Finding simplified solutions to these methods that execute in real time is a challenging endeavour. For the structure of the DBNs, feature selection methods besides sequential backward selection can be explored and their results compared with different discriminative heuristics.

9.2.2 Generalize to naturally evoked mental states

The videos from the Mind Reading DVD and the CVPR 2004 corpus, like the ones on expression databases of basic emotions, are posed. However, deliberate facial displays typically differ in appearance and timing from the natural facial expressions induced through events in the normal environment of the subject. A video corpus of naturally evoked facial expressions in complex mental states is needed for the reliable detection of a user's mental state in real-world applications. A study that elicits these mental states would need to be designed, the resulting videos would have to be segmented, labelled and verified with the participants. This task is made even more complex if multiple modalities are to be measured.

Once a natural corpus is constructed and labelled, the performance of the system would have to be tested when trained on posed videos and tested on a natural one, or when

trained and tested on the same natural corpus, and finally when trained on one natural corpus and tested on another natural one. The generalization ability of systems trained on posed data and tested on natural ones is an interesting problem that has implications for the mainstream application of this technology.

9.2.3 Applications of automated mind-reading

The conceptualization, development and validation of applications of automated mind-reading in traditional, as well as novel, computing scenarios is a challenging and exciting research endeavour. Specific (vertical) application areas include assistive technologies, learning, security, e-commerce and the automobile industry. In assistive technologies, we have presented the emotional hearing aid [KR03, KR04a], an assistive tool designed to help children diagnosed with Asperger Syndrome read and respond to the facial expressions of people they interact with. There are equally interesting applications of automated mind-reading in mainstream (horizontal) computing domains such as computer-mediated communication, ubiquitous computing, and wearable devices. In computer-mediated communication, I have proposed the use of automated mind-reading to automatically recognize a user's mental state and broadcast it to people on his/her buddy list [KR04b].

9.2.4 Refinement of emotion taxonomies

This dissertation has presented several findings on the facial expressions of complex mental states, which have the potential to inform the development of psychological theories of how people read the minds of others. In Chapter 3, the ability of human participants to recognize various complex mental states from facial stimuli was tested. In Chapters 7 and 8, I took a different approach to explore the facial signals of complex mental states using statistical machine learning. The findings from both approaches constitute a strong starting point for further studies on facial signatures of complex mental states. The findings also show that some mental states are "closer" to each other and could co-occur such as *thinking* and *unsure*; other mental states such as *interested* and *concentrating* are grouped under the same mental state group (*interested*) even though their facial signatures are considerably different. One interesting extension of this work is to explore how complex mental states fit in a multi-dimensional state space. This space can then be used in psychometric tests to investigate whether closer mental states are in fact perceived as being close by humans, thereby refining emotion taxonomies such as that of Baron-Cohen *et al.* [BGW+04].

9.3 Summary

Existing human-computer interfaces are mind-blind—oblivious to the user's mental states and intentions. These user interfaces have zero persuasive power, cannot initiate interactions with the user, and are mostly limited to a command and control interaction paradigm. Even if they do take the initiative, like the now retired Microsoft Clip, they are often misguided and irrelevant, and end up frustrating the user. With the increasing complexity of HCI and the ubiquity of mobile and wearable devices, a new interaction paradigm is needed in which systems autonomously gather information about the user's mental state, intentions and surrounding context to adaptively respond to that.

In this dissertation I have described the design, implementation, and validation of a real time system for the inference of complex mental states from head and facial signals

in a video stream. The computational model of mind-reading presents a coherent framework for incorporating mind-reading functions in user interfaces. The implementation of the system has shown that it is possible to infer a wide range of complex mental states from the head and facial displays of people, and that it is possible to do so in real time and with minimal lag.

Moving forward, there are numerous research opportunities that warrant further research. The computational model of mind-reading can be extended to more modalities and context cues in order to recognize a wider range of mental states. A more rigourous learning mechanism needs to be implemented that fuses these different sensors in an efficient way. The model needs to generalize well to naturally evoked mental states, and applications of automated mind-reading in HCI need to be conceptualized, implemented and validated.

As the challenges presented in this dissertation are addressed over time, information about a user's mental state will become as readily available to computer applications as are keyboard, mouse, speech and video input today. Interaction designers will have at their disposal a powerful new tool that will open up intriguing possibilities not only in verticals such as assistive technologies and learning tools, but also in applications we use in our day-to-day lives to browse the web, read emails or write documents. The result will be next-generation applications that employ the user's emotional state to enrich and enhance the quality of interaction, a development that will undoubtedly raise the complexity of human-computer interactions to include concepts such as exaggeration, disguise and deception that were previously limited to human-to-human interaction.

The research presented here serves as an important step towards achieving this vision. By developing a computational model of mind-reading that infers complex mental states in real time, I have widened the scope of human-computer interaction scenarios in which automated facial analysis systems can be integrated. I have also motivated future research that takes full advantage of the rich modality of the human face and of nonverbal cues in general, to further the development of socially and emotionally intelligent interfaces.

Symbols

$\mathbf{X}$	Set of mental state events
$\mathbf{Y}$	Set of head and facial display events
$\mathbf{Z}$	Set of head and facial actions events
x	Number of supported mental states
y	Number of supported head and facial displays
z	Number of supported head and facial actions
$X_i[t]$	Mental state event at time t
$Y_j[t]$	Head or facial display event at time t
$Z_k[t]$	Head or facial action event at time t
$P(X_i[t])$	Probability of mental state i at time t
$P(Y_j[t])$	Probability of display j at time t
θ_i	Dynamic Bayesian Network of mental state i
λ_j	Hidden Markov Model of display j
B_ϕ	Observation function
A	Transition function
π	Prior
1:T	Time range from $t = 1$ up to $t = T$
S	Number of videos in training or test set
e_i	Classification error of mental state i
H	Discriminative power heuristic

Abbreviations

ASD	Autism Spectrum Disorders
AU	Action Unit
AVI	Audio Video Interleave
BNT	Bayes Net Toolbox
CVPR	Computer Vision and Pattern Recognition
DBN	Dynamic Bayesian Network
DVD	Digital Video Disc
FACS	Facial Action Coding System
FER	Facial Expression Recognition
FPS	Frames Per Second
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
MLE	Maximum Likelihood Estimation
PGM	Probabilistic Graphical Model
PNL	Intel's Probabilistic Networks Library
ROC	Receiver Operator Characteristic
SD	Standard Deviation
SVM	Support Vector Machines

Glossary

Action Unit — An independent motion of the face, the head or the eyes.

Actions — The basic spatial and motion characteristics of the head and the facial features in video input. The bottom level of the computational model of mind-reading.

Basic emotions — Emotions that have distinct, universal facial expressions.

Cognitive mental state — A feeling about one's state of knowledge, such as the *feeling of knowing* or the *feeling of not knowing*.

Complex Mental States — Mental states that are not part of the basic emotion set.

Conditional independence — Two events A and B are conditionally independent given a third event C if their occurrences are independent events in their conditional probability distribution given C.

Conditional probability — The probability $P(A|B)$ of some event A, assuming event B.

CVPR 2004 Corpus — Video corpus of facial expressions of complex mental states that I have recorded at the IEEE 2004 International Conference on Computer Vision and Pattern Recognition.

Displays — Logical unit with which people describe facial expressions that have meaning potential in the contexts of communication. Consists of a running sequence of head or facial actions. In the computational model of mind-reading, it is the intermediate level between tracked actions and inferred mental states.

Dynamic Bayesian Networks — A class of probabilistic graphical models in which nodes represent random variables or events, and the (lack of) arcs represent conditional independence assumptions. Additional arcs between consecutive time slices encode temporal dependencies between variables.

Dynamic variables — Variables that evolve in time.

FaceTracker Feature point tracking software that is part of Nevenvisions facial feature tracking SDK.

Facial Action Coding System Describes all possible movements of the face, head and eyes. Published by Ekman and Friesan in 1978.

Facial Expression Recognition The automated analysis of images or video clips of a person's face or facial movement with the objective of describing that person's facial expression and corresponding emotional state.

Forward-Backward Algorithm A special case of the Expectation-Maximization algorithm for parameter estimation. The algorithm runs forwards and backwards through each training example using the actual values of the observed and hidden states to successively refine an initial estimate of the parameters. Used in the training and inference of Hidden Markov Models, and for exact inference in Dynamic Bayesian Networks.

Generalization A system's performance when trained on one corpus and tested on a different one.

Graphical model A graph that represents dependencies among random variables or events. In a directed graphical model, also known as Bayesian Network, any two nodes that are not in a parent/child relationship are conditionally independent given the values of their parents.

Hidden Markov Model A model that represents the statistical behaviour of an observable symbol sequence in terms of a network of hidden states. The hidden states are temporally connected in a Markov chain.

Inter-expression dynamics The transition between consecutive facial expressions. Represents the information gained by processing a facial expression in the context of the one preceeding it.

Intra-expression dynamics The way facial movements unfold within a single facial expression.

Joint probability The probability $P(A, B)$ of two events A and B happening together.

Latency The time elapsed or lag between the onset of an event and when the system recognizes it. For the automated mind-reading system, the latency is the difference between the instant a frame is captured and the time when the system infers the mental state.

Marginal probability The probability $P(A)$ of an event A, ignoring any information about the other event B. It is obtained by summing or integrating the joint probability over B.

Markov chain A discrete-time stochastic process with the Markov property. In such a process future states depend on the present state but are independent of the past.

Maximum Likelihood Estimation A method of point estimation that estimates an unobservable population with parameters that maximize the likelihood function.

Mental states States of mind that people exhibit, express and attribute to each other. Include emotions, cognitive states, intentions, beliefs, desires and focus of attention. The top-most level of the computational model of mind-reading.

Mind-reading The ability to attribute mental states to others from their behaviour, and to use that knowledge to guide one's own actions and predict that of others.

Mind Reading DVD An interactive guide to emotions that includes a comprehensive video collection of mental state enactments.

Mind-reading Machines Human-computer interfaces that, through a computational model of mind-reading, have an awareness of the user's state of mind. This user-awareness is used to drive the functions of the interface accordingly.

Probabilistic graphical model A graph that represents the the prior knowledge of the causal probability and conditional independence relations among events. Examples include Naive Bayes Models and Dynamic Bayesian Networks.

Real time In human-computer interaction, a system is real time if it responds to the user without a noticeable delay.

Throughput The number of events that are processed per unit time. For the automated mind-reading system, the throughput is the number of mental state inferences made per second.

References

[ABD00] Sara B. Algoe, Brenda N. Buswell, and John D. DeLamater. Gender and Job Status as Contextual Cues for the Interpretation of Facial Expression of Emotion. *Sex Role*, 42(3):183–208, 2000.

[Ado02] Ralph Adolphs. Recognizing Emotion from Facial Expressions: Psychological and Neurological Mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1:21–61, 2002.

[AS02] Anne Aula and Veikko Surakka. Auditory Emotional Feedback Facilitates Human-Computer Interaction. In *Proceedings of British HCI Conference*, pages 337–349. Berlin: Springer-Verlag, 2002.

[Ass94] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition*. Washington, DC: American Psychiatric Association, DSM-IV, 1994.

[BA02] Cynthia Breazeal and Lijin Aryananda. Recognizing Affective Intent in Robot Directed Speech. *Autonomous Robots*, 12(1):83–104, 2002.

[BA03] Cynthia Breazeal and Lijin Aryananda. Emotion and Sociable Humanoid Robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, 2003.

[Bak91] Raimo Bakis. Coarticulation Modeling with Continuous-state HMMs. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition*, pages 20–21, 1991.

[Bar94] Simon Baron-Cohen. How to Build a Baby That Can Read Minds: Cognitive Mechanisms in Mindreading. *Current Psychology of Cognition*, 13(5):513–552, 1994.

[Bar95] Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press, 1995.

[Bar01] Simon Baron-Cohen. Theory of Mind and Autism: A Review. *Special Issue of the International Review of Mental Retardation*, 23(169):170–184, 2001.

[Bar03] Simon Baron-Cohen. *The Essential Difference: The Truth about the Male and Female Brain*. New York: Basic Books, 2003.

[BBLM86] Janet B. Bavelas, Alex Black, Charles R. Lemery, and Jennifer Mullett. “I Show How You Feel”: Motor Mimicry as a Communicative Act. *Journal of Personality and Social Psychology*, 50:322–329, 1986.

[BC92] Simon Baron-Cohen and Pippa Cross. Reading the Eyes: Evidence for the Role of Perception in the Development of a Theory of Mind. *Mind and Language*, 6:173–186, 1992.

[BC03] Andrea Bunt and Cristina Conati. Probabilistic Student Modelling to Improve Exploratory Behaviour. *Journal of User Modeling and User-Adapted Interaction*, 13:269–309, 2003.

[BCL01] Fabrice Bourel, Claude C. Chibelushi, and Adrian A. Low. Recognition of Facial Expressions in the Presence of Occlusion. In *Proceedings of the British Machine Vision Conference*, pages 213–222, 2001.

[BDD00] Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. Emotion, Decision-making and the Orbitofrontal Cortex. *Cereb Cortex*, 10(3):295–307, 2000.

[Bec04] Antoine Bechara. The Role of Emotion in Decision-making: Evidence from Neurological Patients with Orbitofrontal Damage. *Brain and Cognition*, 55:30–40, 2004.

[BEP96] Sumit Basu, Irfan Essa, and Alex Pentland. Motion Regularization for Model-Based Head Tracking. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 611–616. IEEE press, 1996.

[BGW+04] Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, , and Jacqueline Hill. A New Taxonomy of Human Emotions. *(under review)*, 2004.

[BGWH04] Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, and Jacqueline J. Hill. *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004.

[BHES99] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Measuring Facial Expressions by Computer Image Analysis. *Psychophysiology*, 36(2):253–263, 1999.

[BI98] Andrew Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*. New York: Springer-Verlag, 1998.

[Bil00] Jeffrey Bilmes. Dynamic Bayesian Multi-Networks. In *Proceedings of International Conference on Uncertainty in Aritificial Intelligence (UAI)*, pages 38–45. San Francisco, CA: Morgan Kaufmann Publishers, 2000.

[Bir70] Ray Birdwhistell. *Kinesics and Context*. Philadelphia: University of Pennsylvania Press, 1970.

[Bis95] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[BLB+03] Marian Stewart Bartlett, Gwen Littlewort, Bjorn Braathen, Terrence J. Sejnowski, and Javier R. Movellan. A Prototype for Automatic Recognition of Spontaneous Facial Actions. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 1271–1278. Cambridge, MA: MIT Press, 2003.

[BLF85] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the Autistic Child have a "Theory of Mind"? *European Journal of Neuroscience*, 21(1):37–46, 1985.

[BLFM03] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human-Computer Interaction. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

[Blu53] Daniel Blum. *A Pictorial History of the Silent Screen*. London: Spring Books, 1953.

[BML+04] Marian Stewart Bartlett, Javier R. Movellan, Gwen Littlewort, Bjorn Braathen, Mark G. Frank, and Terrence J. Sejnowski. *What the Face Reveals (2nd Edition): Basic and Applied Studies of Spontaneaous Expression using the Facial Action Coding System (FACS)*, chapter Towards Automatic Recognition of Spontaneous Facial Actions. New York: Oxford University Press Series in Affective Science, 2004.

[BRF+96] Simon Baron-Cohen, Angel Riviere, Masato Fukushima, Davina French, Julie Hadwin, Pippa Cross, Catherine Bryant, and Maria Sotillo. Reading the Mind in the Face: A Cross-cultural and Developmental Study. *Visual Cognition*, 3:39–59, 1996.

[Bru86] Vicki Bruce. *Recognizing Faces*. Hove, East Sussex: Lawrence Erlbaum, 1986.

[BRW+99] Simon Baron-Cohen, Howard A. Ring, Sally Wheelwright, Edward T. Bullmore, Mick J. Brammer, Andrew Simmons, and Steve C. R. Williams. Social Intelligence in the Normal and Autistic Brain: An fMRI Study. *European Journal of Neuroscience*, 11:1891–1898, 1999.

[BT03] Magali Batty and Margot J. Taylor. Early Processing of the Six Basic Facial Emotional Expressions. *Cognitive Brain Research*, 17:613–620, 2003.

[BV99] Volker Blanz and Thomas Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the ACM SIGGRAPH'99*, pages 187–194, 1999.

[BWH+01] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The Reading the Mind in the Eyes Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, 42(2):241–251, 2001.

[BWJ97] Simon Baron-Cohen, Sally Wheelwright, and Therese Jolliffe. Is There a Language of the Eyes? Evidence from Normal Adults, and Adults with Autism or Asperger Syndrome. *Visual Cognition*, 4(3):311–331, 1997.

[BWL+02] Simon Baron-Cohen, Sally Wheelwright, John Lawson, Rick Griffin, and Jacqueline Hill. *Handbook of Childhood Cognitive Development*, chapter The Exact Mind: Empathising and Systemising in Autism Spectrum Conditions, pages 491–508. Oxford: Blackwell Publishers, 2002.

[BY95] Michael Black and Yaser Yacoob. Tracking and Recognizing Rigid and Non-Rigid Facial Motions using Local Parametric Models of Image Motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 374–382, 1995.

[BY97] Michael Black and Yaser Yacoob. Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[BY98] Vicki Bruce and Andrew W. Young. *In the Eye of the Beholder: The Science of Face Perception*. Oxford: Oxford University Press, 1998.

[Cab02] Michael Cabanac. What is Emotion? *Behavioural Processes*, 60:69–83, 2002.

[CBM+01] Andrew J. Calder, A. Mike Burton, Paul Miller, Andrew W. Young, and Shigeru Akamatsu. A Principal Component Analysis of Facial Expressions. *Vision Research*, 41:1179–1208, 2001.

[CDLS99] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag, 1999.

[CEJ98] Timothy F. Cootes, Gareth J. Edwards, and Chris J.Taylor. Active Appearance Models. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 484–498, 1998.

[CFH99] Antonio Colmenarez, Brendan Frey, and Thomas S. Huang. Embedded Face and Facial Expression Recognition. In *Proceedings of IEEE International Conference on Image Processing*, volume 1, pages 633–637, 1999.

[CGV02] Cristina Conati, Abigail Gertner, and Kurt VanLehn. Using Bayesian Networks to Manage Uncertainty in Student Modeling. *Journal of User Modeling and User-Adapted Interaction*, 12:371–417, 2002.

[CJ92] Timothy F. Cootes and Chris J.Taylor. Active Shape Models. In *Proceedings of the British Machine Vision Conference*, pages 266–275, 1992.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm (Last accessed 25th March, 2005), 2001.

[CLK+02] Andrew J. Calder, Andrew D. Lawrence, Jill Keane, Sophie K. Scott, Adrian M. Owena, Ingrid Christoffels, and Andrew W. Young. Reading the Mind from Eye Gaze. *Neuropsychologia*, 40:1129–1138, 2002.

[CNH01] Naiwala P. Chandrasiri, Takeshi Naemura, and Hiroshi Harashima. Real Time Facial Expression Recognition with Applications to Facial Animation in MPEG-4. *IEICE Trans. INF. and SYST*, (8):1007–1017, 2001.

[CO00] Gerald L. Clore and Andrew Ortony. *Cognitive Neuroscience of Emotion*, chapter Cognition in Emotion: Always, Sometimes or Never?, pages 24–61. Oxford: Oxford University Press, 2000.

[Coh04] Jeffrey F. Cohn. *What the face reveals (2nd edition): Basic and Applied Studies of Spontaneaous Expression using the Facial Action Coding System (FACS)*, chapter Automated Analysis of the Configuration and Timing of Facial Expressions. New York: Oxford University Press Series in Affective Science, 2004.

[Con02] Cristina Conati. Probabilistic Assessment of User's Emotions in Educational Games. *Journal of Applied Artificial Intelligence Special Issue on Merging Cognition and Affect in HCI*, 16:555–575, 2002.

[CRA+04] Jeffrey F. Cohn, Lawrence Ian Reed, Zara Ambadar, Jing Xiao, and Tsuyoshi Moriyama. Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, 2004.

[CRPP02] Tanzeem Choudhury, James M. Rehg, Vladimir Pavlovic, and Alex Pentland. Boosting and Structure Learning in Dynamic Bayesian Networks for Audio-Visual Speaker Detection. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 789–794, 2002.

[Cru98] Josoeph L. Hernandez Cruz. Mindreading: Mental State Ascription and Cognitive Architecture. *Mind and Language*, 13(3):323–340, 1998.

[CSA00] Marco La Cascia, Stan Sclaroff, and Vassilis Athitsos. Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(4):322–336, 2000.

[CSC+03a] Ira Cohen, Nicu Sebe, Lawrence S. Chen, Ashutosh Garg, and Thomas S. Huang. Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding (CVIU) Special Issue on Face recognition*, 91(1-2):160–187, 2003.

[CSC+03b] Ira Cohen, Nicu Sebe, Fabio G. Cozman, Marcelo C. Cirelo, and Thomas S. Huang. Learning Bayesian Network Classifiers for Facial Expression Recognition with both Labeled and Unlabeled Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 595–604, 2003.

[CSGH02] Ira Cohen, Nicu Sebe, Ashutosh Garg, and Thomas S. Huang. Facial Expression Recognition from Video Sequences. In *Proceedings of International conference on Multimedia and Expo*, pages 121–124, 2002.

[CT99] Justine Cassell and Kristinn R. Thorisson. The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, 13(4):519–538, 1999.

[CYKD01] Andrew J. Calder, Andrew W. Young, Jill Keane, and Michael Dean. Configural Information in Facial Expression Perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):527–551, 2001.

[CZLK98] Jeffrey F. Cohn, Adena J. Zlochower, James J. Lien, and Tokeo Kanade. Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.

[CZLK99] Jeffrey F. Cohn, Adena J. Zlochower, James J. Lien, and Tokeo Kanade. Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual FACS Coding. *Psychophysiology*, 36:35–43, 1999.

[CZLK04] Jeffrey F. Cohn, Adena J. Zlochower, James J. Lien, and Takeo Kanade. Multimodal Coordination of Facial Action, Head Rotation, and Eye Motion during Spontaneous Smiles. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 129–138, 2004.

[Dam94] Antonio R. Damasio. *Descartes Error: Emotion, Reason and the Human Brain*. New York: Putnam Sons, 1994.

[Dar65] Charles Darwin. *The Expression of Emotions in Man and Animals*. Chicago, IL: University of Chicago Press, 1965.

[Dau01a] Kerstin Dautenhahn. Robots as Social Actors: AURORA and the Case of Autism. In *Proceedings of Third International Cognitive Technology Conference*, pages 359–374, 2001.

[Dau01b] Kerstin Dautenhahn. The Art of Designing Socially Intelligent Agents - Science, Fiction, and the Human in the Loop. *Applied Artificial Intelligence Journal Special Issue on Socially Intelligent Agents*, 12(7), 2001.

[DB02] Kerstin Dautenhahn and Aude Billard. *In Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT)*, chapter Games Children with Autism Can Play with Robota, a Humanoid Robotic Doll, pages 179–190. London: Springer-Verlag, 2002.

[DBH+99] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrance J. Sejnowski. Classifying Facial Actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 21(10):974–989, 1999.

[Den89] Daniel C. Dennett. *The Intentional Stance*. Cambridge, MA: MIT Press, 1989.

[DP97] Pedro Domingos and Michael Pazzani. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning*, 29:103–130, 1997.

[DV01] James Davis and Serge Vaks. A Perceptual User Interface for Recognizing Head Gesture Acknowledgements. In *Proceedings of the Workshop on Perceptive User Interfaces*, pages 1–7, 2001.

[Edw98] Kari Edwards. The Face of Time: Temporal Cues in Facial Expression of Emotion. *Psychological Science*, 9:270–276, 1998.

[EF69] Paul Ekman and Wallace V. Friesen. The Repertoire of Non-verbal Behaviour: Categories, Origins, Usage, and Coding. *Semiotica*, 1:49–98, 1969.

[EF71] Paul Ekman and Wallace V. Friesen. Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17:124–129, 1971.

[EF76] Paul Ekman and Wallace V. Friesen. *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press, 1976.

[EF78] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[EF86] Paul Ekman and Wallace V. Friesen. A New Pan-cultural Facial Expression of Emotion. *Motivation and Emotion*, 10:159–168, 1986.

[EFA80] Paul Ekman, Wallace V. Friesen, and Sonia Ancoli. Facial Signs of Emotional Experience. *Journal of Personality and Social Psychology*, 39:1125–1134, 1980.

[EHH+04] Nancy Edenborough, Riad Hammoud, Andrew P. Harbach, Alan Ingold, Branislav Kisacanin, Phillip Malawey, Timothy J. Newman, Gregory Scharenbroch, Steven Skiver, Matthew R. Smith, Andrew Wilhelm, Gerald J. Witt, Eric Yoder, and Harry Zhang. Driver Drowsiness Monitor from DELPHI. In *Demonstration at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[Ekm79] Paul Ekman. *Human Ethology*, chapter About Brows: Emotional and Conversational Signals, pages 169–200. London: Cambridge University Press, 1979.

[Ekm92a] Paul Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6:169–200, 1992.

[Ekm92b] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: Norton, 1992.

[Ekm94] Paul Ekman. *The Nature of Emotion: Fundamental questions*, chapter All Emotions are Basic, pages 15–19. New York: Oxford University Press, 1994.

[Ell03] Phoebe C. Ellsworth. Confusion, Concentration and Other Emotions of Interest. *Emotion*, 3(1):81–85, 2003.

[EP95] Irfan A. Essa and Alex Pentalnd. Facial Expression Recognition using a Dynamic Model and Motion Energy. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 360–367, 1995.

[ES02] Murat Erdem and Stan Sclaroff. Automatic Detection of Relevant Head Gestures in American Sign Language Communication. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 10460–10463, 2002.

[Fac02] FaceTracker. *Facial Feature Tracking SDK*. Neven Vision, 2002.

[Fas02a] Beat Fasel. Facial Expression Analysis using Shape and Motion Information Extracted by Convolutional Neural Networks. In *International IEEE Workshop on Neural Networks for Signal Processing (NNSP 02)*, pages 607–616, 2002.

[Fas02b] Beat Fasel. Head-pose Invariant Facial Expression Recognition using Convolutional Neural Networks. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 529–534, 2002.

[FDWS91] Jose-Miguel Fernandez-Dols, Harald G. Wallbott, and Flor Sanchez. Emotion Category Accessibility and the Decoding of Emotion from Facial Expression and Context. *Journal of Nonverbal Behavior*, 15:107–123, 1991.

[FGG97] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29:131–163, 1997.

[FHF+95] Paul C. Fletcher, Francesca Happè, Uta Frith, Susan C. Baker, Raymond J. Dolan, Richard S.J. Frackowiak, and Christopher D. Frith. Other Minds in the Brain: A Functional Imaging Study of "Theory of Mind" in Story Comprehension. *Cognition*, 57(2):109–128, 1995.

[FKtS89] Nico H. Frijda, Peter Kuipers, and Elisabeth ter Schure. Relations Among Emotion, Appraisal, and Emotional Action Readiness. *Journal of Personality and Social Psychology*, 57(2):212–228, 1989.

[FL00] Beat Fasel and Juergen Luettin. Recognition of Asymmetric Facial Action Unit Activities and Intensities. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 1100–1103, 2000.

[FL03] Beat Fasel and Juergen Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003.

[Fli00] Echo Fling. *Eating an Artichoke: A Mother's Perspective on Asperger Syndrome*. London: Jessica Kingsley Publishers, 2000.

[FND03] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A Survey of Socially Interactive Robots, Robotics and Autonomous Systems. *Journal of Nonverbal Behavior*, 42(3–4):143–166, 2003.

[FPHK94] Francesc J. Ferri, Pavel Pudil, Mohamad Hatef, and Josef Kittler. *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies and Hybrid Systems*, chapter Comparative Study of Techniques for Large Scale Feature Selection, pages 403–413. New York: Elsevier Science, 1994.

[Fri89a] Uta Frith. *Autism and Asperger Syndrome*. Cambridge, UK: Cambridge University Press, 1989.

[Fri89b] Uta Frith. *Autism: Explaining the Enigma*. Oxford: Blackwell Publishers, 1989.

[Fri92] Alan J. Fridlund. *Emotion*, chapter The Behavioral Ecology and Sociality of Human Faces. Newbury Park, CA: Sage, 1992.

[Fri97] Alan J. Fridlund. *The Psychology of Facial Expression*, chapter The New Ethology of Human Facial Expressions, pages 103–131. Newbury Park, CA: Sage, 1997.

[Fri01] Uta Frith. Mind Blindness and the Brain in Autism. *Neuron*, 32(6):969–979, 2001.

[Gam97] Dani Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman and Hall, 1997.

[GH02] Haipeng Guo and William H. Hsu. A Survey of Algorithms for Real-Time Bayesian Network Inference. In *AAAI/KDD/UAI Joint Workshop on Real-Time Decision Support and Diagnosis Systems*, 2002.

[GHB+00] Helen L. Gallagher, Francesca Happè, N. Brunswick, Paul C. Fletcher, Uta Frith, and Christopher D. Frith. Reading the Mind in Cartoons and Stories: an fMRI Study of 'Theory of Mind' in Verbal and Nonverbal Tasks. *Neuropsychologia*, 38(1):11–21, 2000.

[GJ04] Haisong Gu and Qiang Ji. Facial Event Classification with Task Oriented Dynamic Bayesian Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 870–875, 2004.

[GK01] Amr Goneid and Rana el Kaliouby. Enhanced Facial Feature Tracking for a Natural Model of Human Expression. In *Proceedings of the International Conference on Human-Computer Interaction (HCII)*, pages 323–326, 2001.

[GPH02] Ashutosh Garg, Vladimir Pavlovic, and Thomas S. Huang. Bayesian Networks as Ensemble of Classifiers. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 2, pages 779–784, 2002.

[GPR03] Ashutosh Garg, Vladimir Pavlovic, and James M. Rehg. Boosted Learning in Dynamic Bayesian Networks for Multimodal Speaker Detection. *Proceedings of IEEE*, 91(9):1355–1369, 2003.

[Gro04] Thomas Gross. The Perception of Four Basic Emotions in Human and Nonhuman Faces by Children with Autism and Other Developmental Disabilities. *Journal Of Abnormal Child Psychology*, 32(5):469–480, 2004.

[GS05] Alvin I. Goldman and Chandra Sekhar Sripada. Simulationist Models of Face-based Emotion Recognition. *Cognition*, 94(3):193–213, 2005.

[GSS+88] Karl Grammer, Wulf Schiefenhoevel, Margret Schleidt, Beatrice Lorenz, and Irenaus Eibl-Eibesfeldt. Patterns on the Face: The Eyebrow Flash in Crosscultural Comparison. *Ethology*, 77:279–299, 1988.

[GSV+03] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 529–536. ACM Press, 2003.

[HAMJ00] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K. Jain. Face Detection in Color Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(5):696–706, 2000.

[HBCH99] Patricia Howlin, Simon Baron-Cohen, and Julie Hadwin. *Teaching Children with Autism to Mind-read: A Practical Guide for Teachers and Parents*. New York: John Wiley and Sons, 1999.

[HCFT04] Changbo Hu, Ya Chang, Rogerio Feris, and Matthew Turk. Manifold Based Analysis of Facial Expression. In *In IEEE Workshop on Face Processing in Video at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[Hec95] David Heckerman. A Tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[Hes03] Ursula Hess. Now You See It, Now You Don't- The Confusing Case of Confusion as an Emotion: Commentary on Rozin and Cohen (2003). *Emotion*, 3(1):76–79, 2003.

[HK99] Jonathan Haidt and Dacher Keltner. Culture and Facial Expression: Open-ended Methods Find More Expressions and a Gradient of Recognition. *Cognition and Emotion*, 13(3):225–266, 1999.

[HL03] Jesse Hoey and James J. Little. Bayesian Clustering of Optical Flow Fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1086–1093, 2003.

[HL04] Jesse Hoey and James J. Little. Decision Theoretic Modeling of Human Facial Displays. In *Proceedings of 8th European Conference on Computer Vision (Part III)*, pages 26–38, 2004.

[HMC01] Francesca Happè, Gurjhinder S. Malhi, and Stuart Checkley. Acquired Mind-blindness following Frontal Lobe Surgery? A Single Case Study of Impaired "Theory of Mind" in a Patient Treated with Stereotactic Anterior Capsulotomy. *Neuropsychologia*, 39:83–90, 2001.

[Hoe04] Jesse Hoey. *Decision Theoretic Learning of Human Facial Displays and Gestures*. PhD Thesis, University of British Columbia, Computer Science Department, 2004.

[Hor87] Albert Sydney Hornby. *Oxford Advanced Learner's Dictionary of Current English (11th impression)*. New Delhi, India: Oxford University Press, 1987.

[HPB98] Ursula Hess, Pierre Philippot, and Sylvie Blairy. Facial Reactions to Emotional Facial Expressions: Affect or Cognition? *Cognition and Emotion*, 12(4):590–531, 1998.

[Ise00] Alice M. Isen. *Handbook of Emotions*, chapter Positive Affect and Decision Making, pages 417–435. New York: Guilford Press, 2000.

[JDM00] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao. Statistical Pattern Recognition: A Review. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):4–37, 2000.

[JLO89] Philip N. Johnson-Laird and Keith Oatley. The Language of Emotions: An Analysis of a Semantic Field. *Cognition and Emotion*, 3(2):81–123, 1989.

[JP00] Tony Jebara and Alex Pentland. Parameterized Structure from Motion for 3D Adaptive Feedback Tracking of Faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 144–150, 2000.

[JY02] Qiang Ji and Xiaojie Yang. Real-Time Eye Gaze and Face Pose Tracking for Monitoring Driver Vigilance. *Real-Time Imaging*, 8(5):357–377, 2002.

[JZ97] Anil K. Jain and Douglas Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 19(2):153–158, 1997.

[K00] Rana el Kaliouby. Enhanced Facial Feature Tracking of Spontaneous Facial Expression. M.Sc. Thesis, Computer Science Department, The American University in Cairo, 2000.

[KR03] Rana el Kaliouby and Peter Robinson. The Emotional Hearing Aid: An Assistive Tool for Autism. In *Proceedings of the International Conference on Human-Computer Interaction (HCII): Universal Access in HCI*, volume 4, pages 68–72. Lawrence Erlbaum Associates, 2003.

[KR04a] Rana el Kaliouby and Peter Robinson. *Designing a More Inclusive World*, chapter The Emotional Hearing Aid: An Assistive Tool for Children with Asperger Syndrome, pages 163–172. London: Springer-Verlag, 2004.

[KR04b] Rana el Kaliouby and Peter Robinson. FAIM: Integrating Automated Facial Affect Analysis in Instant Messaging. In *Proceedings of ACM International Conference on Intelligent User Interfaces (IUI)*, pages 244–246, 2004.

[KR04c] Rana el Kaliouby and Peter Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *In IEEE Workshop on Real-Time Vision for Human-Computer Interaction at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[KRK03] Rana el Kaliouby, Peter Robinson, and Simeon Keates. Temporal Context and the Recognition of Emotion from Facial Expression. In *Proceedings of the International Conference on Human-Computer Interaction (HCII): Human-Computer Interaction, Theory and Practice*, volume 2, pages 631–635. Lawrence Erlbaum Associates, 2003.

[KBM+01] Miyuki Kamachi, Vicki Bruce, Shigeru Mukaida, Jiro Gyoba, Sakiko Yoshikawa, and Shigeru Akamatsu. Dynamic Properties Influence the Perception of Facial Expressions. *Perception*, 30(7):875–887, 2001.

[KCT00] Tokeo Kanade, Jeffrey Cohn, and Ying-Li Tian. Comprehensive Database for Facial Expression Analysis. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[KEG+03] Clinton D. Kilts, Glenn Egan, Deborah A. Gideon, Timothy D. Ely, and John M. Hoffman. Dissociable Neural Pathways are Involved in the Recognition of Emotion in Static and Dynamic Facial Expressions. *NeuroImage*, 18(1):156–168, 2003.

[KK03] Eva Krumhuber and Arvid Kappas. Moving Smiles: The Influence of the Dynamic Components on the Perception of Smile-genuineness. In *10th European Conference on Facial Expressions: Measurement and Meaning*, 2003.

[KMP00] Jonathan Klein, Youngme Moon, and Rosalind W. Picard. This Computer Responds to User Frustration: Theory, Design, and Results. *Interacting with Computers*, 14(2):119–140, 2000.

[KO00] Shinjiro Kawato and Jun. Ohya. Real-Time Detection of Nodding and Head-shaking by Directly Detecting and Tracking the "Between-Eyes". In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 40–45, 2000.

[Koh95] Ron Kohavi. A Study of Cross-validation and Bootstrap for Accuracy estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1137–1143, 1995.

[KP01] Ashish Kapoor and Rosalind W. Picard. A Real-Time Head Nod and Shake Detector. In *Proceedings of the Workshop on Perceptive User Interfaces*, 2001.

[KP05] Branislav Kisacanin and Vladimir Pavlovic. *Real-Time Vision for Human-Computer Interaction*, chapter Real-Time Algorithms: From Signal Processing to Computer Vision. Springer-Verlag, 2005.

[KPI04] Ashish Kapoor, Rosalind W. Picard, and Yuri Ivanov. Probabilistic Combination of Multiple Modalities to Detect Interest. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 969–972, 2004.

[KQP03] Ashish Kapoor, Yuan Qi, and Rosalind W. Picard. Fully Automatic Upper Facial Action Recognition. In *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG) at the International Conference on Computer Vision (ICCV)*, 2003.

[KRP01] Barry Kort, Rob Reilly, and Rosalind Picard. An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy Building a Learning Companion. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 43–46, 2001.

[KSFVM01] Pierre Krolak-Salmon, Catherine Fischer, Alail Vighetto, and François Mauguière. Processing of Facial Emotional Expression: Spatio-temporal Data as Assessed by Scalp Event Related Potentials. *European Journal of Neuroscience*, 13(5):987–994, 2001.

[Kur03] Thomas Kurz. *Stretching Scientifically: A Guide to Flexibility Training*. Island Pond, VT: Stadion Publishing Co, 2003.

[LB03] Karen Lander and Vicki Bruce. *Advances in Audiovisual Speech Processing*, chapter Dynamic-varying information for person perception. Cambridge, MA: MIT Press, 2003.

[LBF+04a] Gwen Littlewort, Marian S. Bartlett, Ian Fasel, Joshua Susskind, and Javier R. Movellan. Dynamics of Facial Expression Extracted Automatically from Video. In *In IEEE Workshop on Face Processing in Video at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[LBF+04b] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joel Chenu, Takayuki Kanda, Hiroshi Ishiguro, and Javier R. Movellan. Towards Social Robots: Automatic Evaluation of Human-Robot Interaction by Face Detection and Expression Classification. In S. Thrun and B. Schoelkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, 2004.

[LeD96] Josesh E. LeDoux. *The Emotional Brain*. New York: Simon & Schuster, 1996.

[LNL+03] Christina Lisetti, Fatma Nasoza, Cynthia LeRouge, Onur Ozyer, and Kaye Alvarez. Developing Multimodal Intelligent Affective Interfaces for Tele-home Health Care. *International Journal of Human-Computer Studies*, 59, 2003.

[LWB00] Stephen R.H. Langton, Roger J. Watt, and Vicki Bruce. Do The Eyes Have It? Cues to the Direction of Social Attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[LZCK98] James J. Lien, Adena Zlochower, Jeffrey F. Cohn, and Takeo Kanade. Automated Facial Expression Recognition Based on FACS Action Units. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 390–395, 1998.

[Mac95] I. Scott MacKenzie. *Virtual Environments and Advanced Interface Design*, chapter Input Devices and Interaction Techniques for Advanced Computing, pages 437–470. Oxford: Oxford University Press, 1995.

[Mel04] Andrew N. Meltzoff. *Handbook of Childhood Cognitive Development*, chapter Imitation as a Mechanism of Social Cognition: Origins of Empathy, Theory of Mind, and the Representation of Action, pages 6–25. Oxford: Blackwell Publishers, 2004.

[MF91] Chris Moore and Douglas Frye. *Childrens Theories of Mind*, chapter The Acquisition and Utility of Theories of Mind, pages 1–14. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[MH05] Roberta Muramatsu and Yaniv Hanoch. Emotions as a Mechanism for Boundedly Rational Agents: The Fast and Frugal Way. *Journal of Economic Psychology*, 26(2):201–221, 2005.

[Mic03] Philipp Michel. Emotion Recognition using Support Vector Machines. Part II Dissertation, Computer Laboratory, University of Cambridge, 2003.

[Min01] Thomas Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of International Conference on Uncertainty in Aritificial Intelligence (UAI)*, pages 362–369, 2001.

[MK03a] Philipp Michel and Rana el Kaliouby. Facial Expression Recognition Using Support Vector Machines. In *Proceedings of the International Conference on Human-Computer Interaction (HCII): Human-Computer Interaction, Theory and Practice*, volume 2, pages 93–94. Lawrence Erlbaum Associates, 2003.

[MK03b] Philipp Michel and Rana el Kaliouby. Real Time Facial Expression Recognition in Video using Support Vector Machines. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 258–264, 2003.

[ML04] Sinjini Mitra and Yanxi Liu. Local Facial Asymmetry for Expression Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 889–894, 2004.

[MM63] Thomas Marill and David M.Green. On the Effectiveness of Receptors in Recognition Systems. *IEEE Transactions*, IT-9:11–27, 1963.

[Mur01] Kevin P. Murphy. The Bayes Net Toolbox for Matlab. Technical report, Computer Science Division, University of California, Berkeley, CA, 2001.

[Mur02] Kevin P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD Thesis, UC Berkeley, Computer Science Division, 2002.

[MW93] I. Scott MacKenzie and Colin Ware. Lag as a Determinant of Human Performance in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 488–493, 1993.

[MYD96] Carlos Morimoto, Yaser Yacoob, and Larry Davis. Recognition of Head Gestures using Hidden Markov Models. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 461–465, 1996.

[Mye94] Richard Myers. *Myer's Hidden Markov Model Software*. 1994.

[NM00] Clifford Nass and Youngme Moon. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[O'C98] Sanjida O'Connell. *Mindreading: An Investigation into How we Learn to Love and Lie*. New York: Doubleday Books, 1998.

[OHG02] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Hierarchical Representations for Learning and Inferring Office Activity from Multimodal Information. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 3–8, 2002.

[OPB97] Nuria Oliver, Alex Pentland, and François Bérard. LAFTER: Lips and Face Real Time Tracker. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1445–1449, 1997.

[PA03] Mette T. Posamentier and Hervé Abdi. Processing Faces and Facial Expressions. *Neuropsychology Review*, 13(3):113–144, 2003.

[PA04] Sangho Park and J.K. Aggarwal. Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy. In *IEEE Workshop on Articulated and Non Rigid Motion at the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[PAF+01] Dale Purves, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel LaMantia, James O. McNamara, and Mark S. Williams, editors. *Neuroscience*, chapter Emotions, pages 2030–2076. Sunderland: Sinauer Associates, 2001.

[PBL02] Montse Pardàs, Antonio Bonafonte, and José Luis Landabaso. Emotion Recognition based on MPEG4 Facial Animation Parameters. In *Proceedings of the IEEE Internatoinal Conference on Acoustics, Speech and Signal Procssing*, volume 4, pages 3624–3627, 2002.

[PC95] Curtis Padgett and Garrison W. Cottrell. Identifying Emotion in Static Images. In *Proceedings of the Second Joint Symposium of Neural Computation*, volume 5, pages 91–101, 1995.

[PCC+03] Ana Paiva, Marco Costaa, Ricardo Chavesa, Moiss Piedadea, Drio Mouroa, Daniel Sobrala, Kristina Höök, Gerd Andersson, and Adrian Bullock. SenToy: An Affective Sympathetic Interface. *International Journal of Human-Computer Studies*, 59:227–235, 2003.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, 1988.

[Pet05] Eric Petajan. *Real-Time Vision for Human-Computer Interaction*, chapter Vision-based HCI Applications. Springer-Verlag, 2005.

[Pic97] Rosalind W. Picard. *Affective Computing*. Cambridge, Massachusetts: MIT Press, 1997.

[PNL03] PNL. *Open-Source Probabilistic Networks Library*. Intel Corporation, 2003.

[PPB+05] Rosalind W. Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. Affective Learning – a Manifesto. *BT Technology Journal*, 22:253–269, 2005.

[PR00a] Maja Pantic and Leon J.M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22:1424–1445, 2000.

[PR00b] Maja Pantic and Leon J.M. Rothkrantz. Expert System for Automatic Analysis of Facial Expressions. *Image and Vision Computing*, 18:881–905, 2000.

[PR03] Maja Pantic and Leon J.M. Rothkrantz. Toward an Affect-sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[PS00] Rejean Plamondon and Sargur N. Srihari. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):63–84, 2000.

[PS01] Rosalind W. Picard and Jocelyn Scheirer. The Galvactivator: A Glove that Senses and Communicates Skin Conductivity. In *Proceedings of the International Conference on Human-Computer Interaction (HCII)*, pages 91–101, 2001.

[PS04] Timo Partala and Veikko Surakka. The Effects of Affective Interventions in Human-Computer Interaction. *Interacting with Computers*, 16:295–309, 2004.

[PW78] David Premack and Guy Woodruff. Does the Chimpanzee have a Theory of Mind? *Behavioral and Brain Sciences*, 4:515–526, 1978.

[PWS02] Ann T. Phillips, Henry M. Wellman, and Elizabeth S. Spelke. Infants' Ability to Connect Gaze and Emotional Expression to Intentional Actions. *Cognition*, 85:53–78, 2002.

[QP02] Yuan Qi and Rosalind W. Picard. Context-sensitive Bayesian Classifiers and Application to Mouse Pressure Pattern Classification. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume III, pages 448–451, 2002.

[Rab89] Lawrence Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE*, 77(2):257–286, 1989.

[RAN+03] Anu Realo, Jri Allik, Aire N¨lvak, Raivo Valk, Tuuli Ruus, Monika Schmidt, and Tiina Eilola. Mind-Reading Ability: Beliefs and Performance. *Journal of Research in Personality*, 37(5):420–445, 2003.

[Rat89] Carl Ratner. A Social Constructionist Critique of Naturalistic Theories of Emotion. *Journal of Mind and Behavior*, 10:211–230, 1989.

[RBCW02] Mel D. Rutherford, Simon Baron-Cohen, and Sally Wheelwright. Reading the Mind in the Voice: A Study with Normal Adults and Adults with Asperger Syndrome and High Functioning Autism. *Journal of Autism and Developmental Disorders*, 32(3):189–194, 2002.

[RC03] Paul Rozin and Adam B. Cohen. High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. *Emotion*, 3(1):68–75, 2003.

[RN96] Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, MA: Cambridge University Press, 1996.

[Rob94] Colin Robson. *Experiment, Design and Statistics in Psychology (3rd Edition)*. London: Penguin, 1994.

[Sab04] Mark Sabbagh. Understanding Orbitofrontal Contributions to Theory-of-mind Reasoning: Implications for Autism. *Brain and Cognition*, 55:209–219, 2004.

[SC01] Karen Schmidt and Jeffrey F. Cohn. Dynamics of Facial Expression: Normative Characteristics and Individual Differences. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 728–731, 2001.

[Sca01] Brian Scassellati. Foundations for a Theory of Mind for a Humanoid Robot. Technical report, Ph.D. Thesis, Department of Electronics Engineering and Computer Science, 2001.

[Sch00] Norbert Schwarz. Emotion, Cognition and Decision Making. *Cognition and Emotion*, 14(4):433–440, 2000.

[SCT03] Karen Schmidt, Jeffrey F. Cohn, and Ying-Li Tian. *What the Face Reveals (2nd edition): Basic and Applied Studies of Spontaneaous Expression using the Facial Action Coding System (FACS)*, chapter Signal Characteristics of Spontaneous Facial Expressions: Automatic Movement in Solitary and Social Smiles. New York: Oxford University Press Series in Affective Science, 2003.

[SFKP02] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W. Picard. Frustrating the User on Purpose: a Step Toward Building an Affective Computer. *Interacting with Computers*, 14(2):93–118, 2002.

[SFP99] Jocelyn Scheirer, Raul Fernandez, and Rosalind W. Picard. Expression Glasses: A Wearable Device for Facial Expression Recognition. In *Proceedings of the Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*, pages 262–263, 1999.

[SJ97] Mubarak Shah and Ramesh Jain. *Motion-Based Recognition*, chapter Visual Recognition of Activities, Gestures, Facial Expressions and Speech: An Introduction and a Perspective. Norwell, MA: Kluwer Academic Publihsers, 1997.

[SKR04] Tal Sobol Shikler, Rana el Kaliouby, and Peter Robinson. Design Challenges in Multi Modal Inference Systems for Human-Computer Interaction. In S. Keates, J. Clarkson, P. Langdon, and P. Robinson, editors, *Proceedings of the 2nd Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT)*, pages 55–58, 2004.

[SKY+04] Wataru Sato, Takanori Kochiyama, Sakiko Yoshikawa, Eiichi Naito, and Michikazu Matsumura. Enhanced Neural Activity in Response to Dynamic Facial Expressions of Emotion: an fMRI Study. *Cognitive Brain Research*, 20(1):81–91, 2004.

[SM90] Peter Salovey and John D. Mayer. Emotional Intelligence. *Imagination, Cognition and Personality*, 9(3):185–211, 1990.

[TBH+03] Ying-Li Tian, Lisa Brown, Arun Hampapur, Sharat Pankanti, Andrew W. Senior, and Ruud M. Bolle. Real World Real-Time Automatic Recognition of Facial Expressions. *IEEE workshop on Performance Evaluation of Tracking and Surveillance*, 2003.

[TC90] John Tooby and Leda Cosmides. The Past Explains the Present : Emotional Adaptations and the Structure of Ancestral Environments. *Ethology and Sociobiology*, 11(4–5):375–424, 1990.

[TFAS00] Jean-Christophe Terrillon, Hideo Fukamachi, Shigeru Akamatsu, and Mahdad N. Shirazi. Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 54–61, 2000.

[TFS00] Helen Tager-Flusberg and Kate Sullivan. A Componential View of Theory of Mind: Evidence from Williams Syndrome. *Cognition*, 76:59–89, 2000.

[TK04] Matthew Turk and Mathias Kolsch. *Emerging Topics in Computer Vision*, chapter Perceptual Interfaces. Prentice Hall PTR, 2004.

[TKC00a] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Eye-state Detection by Local Regional Information. In *Proceedings of the International Conference on Multimedia Interfaces*, pages 143–150, 2000.

[TKC00b] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Robust Lip Tracking by Combining Shape, Color and Motion. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2000.

[TKC01] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(2):97–115, 2001.

[TKC02] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 229–234, 2002.

[TN00] Jinshan Tang and Ryohei Nakatsu. A Head Gesture Recognition Algorithm. In *Proceedings of International Conference on Multimedia Interfaces*, pages 72–80, 2000.

[TP91] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[TR03] Wenzhao Tan and Gang Rong. A Real-Time Head Nod and Shake Detector using HMMs. *Expert Systems with Applications*, 25:461–466, 2003.

[Tro89] Edward Z. Tronick. Emotions and Emotional Communication in Infants. *American Psychologist*, 44:112–119, 1989.

[Tur05] Matthew Turk. *Real-Time Vision for Human-Computer Interaction*, chapter RTV4HCI: A Historical Overview. Springer-Verlag, 2005.

[TW01] Wataru Tsukahara and Nigel Ward. Responding to Subtle, Fleeting Changes in the Users Internal State. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 77–84, 2001.

[Vap95] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[Wal82] Andrew S. Walker. Intermodal Perception of Expressive Behaviours by Human Infants. *Journal of Experimental Child Psychology*, 33:514–535, 1982.

[Wal91] Harald G. Wallbott. In and out of context: Influences of Facial Expression and Context Information on Emotion Attributions. *British Journal of Social Psychology*, 27:357–369, 1991.

[WB05] John Winn and Christopher M. Bishop. Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

[WBC97] Andrew D. Wilson, Aaron F. Bobick, and Justine Cassell. Temporal Classification of Natural Gesture and Application to Video Coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 948–954, 1997.

[WCP00] Christopher R. Wren, Brian P. Clarkson, and Alex P. Pentland. Understanding Purposeful Human Motion. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 19–25, 2000.

[WCVM01] Galen S. Wachtman, Jeffrey F. Cohn, Jessie M. VanSwearingen, and Ernest K. Manders. Automated Tracking of Facial Features in Facial Neuromuscular Disorders. *Plastic and Reconstructive Surgery*, 107:1124–1133, 2001.

[Wel90] Henry Wellman. *The Childs Theory of Mind*. Cambridge, MA: Bradford Books/MIT Press, 1990.

[WFKdM97] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krger, and Christoph von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):775–779, 1997.

[Whi91] Andrew Whiten, editor. *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*. Oxford: Basil Blackwell, 1991.

[Win04] WinSTAT. *WinSTAT: Statistics Add-in for Microsoft Excel*. R. Fitch Software, 2004.

[WPG01] Michael Walter, Alexandra Psarrou, and Shaogang Gong. Data Driven Gesture Model Acquisition Using Minimum Description Length. In *Proceedings of the British Machine Vision Conference*, pages 673–683, 2001.

[WS04] Gillian M. Wilson and M. Angela Sasse. From Doing to Being: Getting Closer to the User Experience. *Interacting with Computers*, 16(4):697–705, 2004.

[XKC02] Jing Xiao, Toekeo Kanade, and Jeffrey F. Cohn. Robust Full Motion Recovery of Head by Dynamic Templates and Re-registration Techniques. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 163–169, 2002.

[YD96] Yasser Yacoob and Larry Davis. Recognizing Human Facial Expression from Long Image Sequences Using Optical Flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 18(6):636–642, 1996.

[YRC+97] Andrew W. Young, Duncan Rowland, Andrew J. Calder, Nancy L. Etcoff, Anil Seth, and David I. Perrett. Facial Expression Megamix: Tests of Dimensional and Category Accounts of Emotion Recognition. *Cognition*, 63(3):271–313, 1997.

[ZC03] Xiaoming Zhou and Cristina Conati. Inferring User Goals from Personality and Behavior in a Causal Model of User Affect. In *Proceedings of ACM International Conference on Intelligent User Interfaces (IUI)*, pages 211–281, 2003.

[ZGPB+04] Dong Zhang, Daniel Gatica-Perez, Samy Bengi, Iain McCowan, and Guillaume Lathoud. Modeling Individual and Group Actions in Meetings: a Two-Layer HMM Framework. In *The Detection and Recognition of Events in Video Workshop at CVPR*, 2004.

[ZJ03] Yongmian Zhang and Qiang Ji. Facial Expression Understanding in Image Sequences Using Dynamic and Active Visual Information Fusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1297–1304, 2003.

[ZLSA98] Zhengyou Zhang, Michael Lyons, Michael Schuster, and Shigeru Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 454–459, 1998.

[ZT01] Jeffrey M. Zacks and Barbara Tversky. Event Structure in Perception and Conception. *Psychological Bulletin*, 127(1):3–21, 2001.

[ZTI01] Jeffrey M. Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, Remembering, and Communicating Structure in Events. *Journal of Experimental Psychology*, 130(1):29–58, 2001.