Designing for Automatic Affect Inference in Learning Environments

Shazia Afzal and Peter Robinson
University of Cambridge, Computer Laboratory, CB3 0FD United Kingdom //Shazia.Afzal@cl.cam.ac.uk // Peter.Robinson@cl.cam.ac.uk

ABSTRACT

Emotions play a significant role in healthy cognitive functioning; they impact memory, attention, decision-making and attitude; and are therefore influential in learning and achievement. Consequently, affective diagnoses constitute an important aspect of human teacher-learner interactions motivating efforts to incorporate skills of affect perception within computer-based learning. This paper provides a discussion of the motivational and methodological issues involved in automatic affect inference in learning technologies. It draws on the recent surge of interest in studying emotions in learning, highlights available techniques for measuring emotions, and surveys recent efforts to automatically measure emotional experience in learning environments. Based on previous studies, six categories of pertinent affect states are identified; the visual modality for affect modelling is selected given the requirements of a viable measurement technique; and a bottom-up analysis approach based on context-relevant data is adopted. Finally, a dynamic emotion inference system that uses state of the art facial feature point tracking technology to encode the spatial and temporal signature of these affect states is described.

Keywords
Affective computing, emotions in learning, computer-based learning, facial affect analysis

Introduction

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

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

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

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

Building on a discussion of recent studies highlighting the relevance of emotions in learning, this paper describes different techniques for measuring emotions and efforts in automatic recognition and/or prediction of affect in learning contexts before proposing a parallel emotion inference system. This is not an exhaustive survey of the past work but a selected discussion of recent works highlighting the concern and those attempting to address it. Throughout this paper the terms ‘emotion’ and ‘affect’ will be used interchangeably.

Learning and Emotions

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

In a series of qualitative case-studies, Pekrun et al. (2002) explored the ‘occurrence and phenomenological structures of academic emotions’. They demonstrated that learners experience a rich diversity of positive and negative emotions; the most frequently reported being: anxiety, enjoyment, hope, pride, and relief, as well as anger, boredom and shame. Developing a multidimensional instrument, the Academic Emotions Questionnaire [AEQ], they conducted quantitative studies to test assumptions underlying Pekrun’s cognitive-motivational model (Pekrun, 1992). Using dimensions of valence (positive vs. negative) and activation (activating vs. deactivating) they distinguished four groups of emotions with reference to their performance effects – positive activating emotions (such as enjoyment of learning, hope, or pride); positive deactivating emotions (e.g., relief, relaxation after success, contentment); negative activating emotions (such as anger, anxiety, and shame); and negative deactivating emotions (e.g., boredom, hopelessness). Accordingly, they studied the effects of these emotions on learning and achievement with cognitive and motivational mechanisms like motivation to learn, strategies of learning, cognitive resources, and self-regulation. Instances of these mechanisms like interest and effort, learning strategies like elaboration or rehearsal, task irrelevant thinking diverting cognitive resources and self-regulated learning as compared to reliance on external guidance may all occur in the course of learning with a computer tutor and are thus directly relevant to this study.

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

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

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

In an attempt to understand the emotional dimension of online learning in qualitative terms, O’Regan (2003) explored the lived experience of students taking online learning courses. The study identifies both positive and negative emotions experienced by students, significantly - frustration, fear/anxiety, shame/embarrassment, enthusiasm/excitement and pride. These had a variable effect on the learning process depending on the strength and nature of the emotion, as well as the associated learning context. In another study, using a manual affect coding system, Craig, Graesser, Sullins, and Gholson (2004) observed the occurrence of six affect states during learning with an intelligent tutoring system. They analysed frustration, boredom, flow, confusion, eureka and neutral, and found significant relationships between learning and the affective states of boredom, flow and confusion.

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

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

**Measuring Emotions**

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

<table>
<thead>
<tr>
<th>Subjective (Qualitative)</th>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Interviews</td>
<td>Questionnaires</td>
<td>Emotional Diaries</td>
</tr>
<tr>
<td>Emotional Probes</td>
<td>Surveys</td>
<td>Think-aloud</td>
</tr>
<tr>
<td>Stimulated Recall</td>
<td></td>
<td>Experience / Time-Sampling</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Objective (Quantitative)</th>
<th>Qualitative</th>
<th>Continuous Type (During Learning)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured Interviews</td>
<td>Transcripts Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Video Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observational Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interactional Content</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Physiology / Nonverbal Behaviour Analysis</td>
<td></td>
</tr>
</tbody>
</table>

For intervention to be effective, remedial action has to be appropriately timed - particularly in the case of strong emotions. Given the complex and transient nature of emotions, any retrospective accounts are problematic because of issues related to the potential for multiple levels of awareness, reappraisals and reconstruction of meanings during recall (Schutz, Hong, Cross, & Obson, 2006). This necessitates dynamic evaluation of emotions but without disrupting the learning task itself. Ideally then, an unobtrusive, quantitative, and continuous account of emotional
experience is a suitable method of enquiry. Amongst the methods listed in Table 1, analysis of nonverbal behaviour in the lower right quadrant, offers a reasonable fit to this requirement (Pekrun, 2005; Picard, et al., 2004; Hudlicka, 2003). Analyses of tutoring sessions have indeed revealed that affective diagnoses, as an important aspect of expert human mentoring, depend heavily on inferences drawn from facial expressions, body language, intonation, and paralinguistic cues (Lepper, Woolverton, Mumme, & Gurtner, 1993). Advances in the field of affective computing have opened the possibility of emotion recognition from its nonverbal manifestations like facial expressions, head pose, body gestures, voice and physiology. The field is promising, yet in a formative stage as current technologies need to be validated for reliability outside controlled experimental conditions.

**Automatic Measurement of Affect**

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

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

**Prior Work**

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

Based on the modelling approach used, affect inference methods can be broadly categorised as (Liao et al., 2006; Alexander, Hill, & Sarrafzadeh, 2005):
- Predictive - those that predict emotions based on an understanding of their causes
- Diagnostic - those that detect emotions based upon their physical effects, and
- Hybrid - those that combine causal and diagnostic approaches

**Table 2. Affect modelling in learning environments**

<table>
<thead>
<tr>
<th>Citation</th>
<th>Affect Construct</th>
<th>Information Source</th>
<th>Learning Context</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Approach</td>
<td>Model</td>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------</td>
<td>------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Jaques &amp; Vicari (2007)</td>
<td>OCC Cognitive Theory of Emotions</td>
<td>User’s actions &amp; interaction patterns</td>
<td>Pedagogical agent based educational environment</td>
<td>Belief-Desire-Intention (BDI) reasoning; appraisal based inference</td>
</tr>
<tr>
<td>Kapoor, Burleson &amp; Picard (2007)</td>
<td>Pre-frustration &amp; Not pre-frustration</td>
<td>Facial expressions, posture, mouse pressure, skin conductance, task state</td>
<td>Automated Learning Companion</td>
<td>Gaussian process classification; Bayesian inference</td>
</tr>
<tr>
<td>Liao et al. (2006)</td>
<td>Stress &amp; fatigue</td>
<td>Physical appearance, physiological, behavioural and performance measures</td>
<td>Maths and audio based experimental tasks</td>
<td>Influence Diagram; Ensemble of classifiers</td>
</tr>
<tr>
<td>Amershi, Conati &amp; Maclaren (2006)</td>
<td>Affective reactions to game events</td>
<td>Skin conductance, heart rate, EMG</td>
<td>Educational game-Prime Climb</td>
<td>Unsupervised clustering</td>
</tr>
<tr>
<td>Kapoor &amp; Picard (2005)</td>
<td>Interest, Disinterest, break-taking behaviour</td>
<td>Facial expressions, posture patterns &amp; task state</td>
<td>Educational Puzzle</td>
<td>Ensemble of classifiers</td>
</tr>
<tr>
<td>Conati (2002); Conati &amp; Zhou (2004)</td>
<td>OCC Cognitive Theory of Emotions</td>
<td>Interaction patterns, personality, goals</td>
<td>Educational game-Prime Climb</td>
<td>Dynamic decision network; Appraisal based inference</td>
</tr>
</tbody>
</table>

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

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

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

D’Mello and Graesser (2007) use posture patterns along with dialogue, to discriminate between affect states during interaction with an intelligent tutoring system called Auto-Tutor. This is a dialogue based system achieving recognition of affect states like flow, confusion, boredom, eureka and neutral. Interestingly however, the ground truth used for validating their classification is mainly the facial action coding of recorded interaction by FACS experts. FACS or the Facial Action Coding System is the anatomic classification devised by Ekman and Friesen (1978) that defines 44 Action Units to describe any human facial expression.

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

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

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

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

Discussion and Scope of this Work

Ideally, automatic sensing should be able to function in real-time; measure multiple and co-occurring emotions unobtrusively and without causing disruption in the actual learning process. As reviewed in the previous section, numerous efforts are being made towards this goal to give computer-based tutoring some semblance of emotional intelligence. Table 2 lists the relevant works and categorises these according to their specific focus and approach. It highlights the variety in modelling techniques that range from rule-based systems to complex probabilistic models; the different ways in which affect is conceptualised in these systems based on whether a dimensional, discrete or appraisal-based stance is adopted; the array of interactional as well as behavioural measures used to infer affect; and importantly, the nature and focus of the learning setup used. Given this diversity in the measured affect constructs, the specific learning environments and the channels used as information sources; it is difficult to comment on the overall performance of a system and determine its efficiency in a broad sense. This inability to make generalisable claims is an acknowledged limitation of affect sensing technologies (Pantic & Rothkrantz, 2003) and makes it challenging to establish the merit and success of a particular system satisfactorily and with confidence. Nevertheless, what is apparent is a growing understanding of the importance of affect modelling in learning and this substantiates further research in the area. The following sections lay out some design choices that set the scope of this work and therefore the proposed system.

Conceptualisation of affect

The issue of representation is at the core of emotion research and therefore affective computing. This is because handling of emotion data by machines requires programmed representations of affect and a clear structure that will perform real-time interaction with a user. Selection of an appropriate descriptive framework embodies the way affect is conceptualised within a system, the way it is observed and assessed, and consequently, the way it is processed (Peter & Herbon, 2006). However, the question of representation is not a simple one as it requires an understanding of the typology and semantics of the whole range of emotion-related phenomena like short-lived, intense emotions; moods; long-lasting established emotions; stances; attitudes/preferences, traits/affect dispositions, etc (Cowie & Cornelius, 2003). All this complicates the task of describing emotional content and while no single best representation scheme exists, there are established psychological traditions that have been used effectively to formalise the behaviour of interest. One of most long-standing way by which affect has been described by psychologists is in terms of discrete categories – an approach rooted in everyday language and driven by historical tradition around the existence of universal emotions.

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

Following such an application-oriented approach, we considered emotion groups of annoyed, anxious, bored, confused, happy, interested, neutral and surprised using the taxonomy of complex mental states by Baron-Cohen (2004). This is a lexical taxonomy that groups together semantically similar emotion concepts so that each group
encompasses the finer shades of an emotion concept. Confusion for example includes states like unsure, puzzled, baffled and clueless while Happy includes pleased, cheerful, relaxed, calm, enjoying, etc. These encompass representative emotions from each of Kort, Reilly and Picard’s (2001) emotion axes as well as those of Pekrun et al.’s (2002) academic emotions with the exception of hope, pride and shame which have more complex social antecedents and meanings and are therefore excluded from this study. The selected emotion descriptors thus have a wider scope than considered by previous methods.

Choice of Modality

Emotion is expressed through visual, vocal and physiological channels. The visual channel includes facial expressions, body gestures, eye-gaze and head pose; the vocal channel focuses on measures of intonation and prosody; while the physiological channel includes measures of skin conductance, blood volume pressure, heart rate, temperature, etc. Lack of a consistent mapping between observable aspects of behaviour and actual affective states, technical feasibility, and practical issues complicate the choice of modality for sensing in a learning setting. Issues of ethics, privacy and comfort further constrain the design, use and deployment of appropriate sensing technologies. The use of physiological sensing in particular is challenging. Though relatively easy to detect and reasonably unobtrusive now, physiological sensing has some inherent shortcomings like requirement of specialised equipment, controlled conditions, baseline determination and normalising procedures, possible discomfort in usage, expertise in use of sensing apparatus and issues of privacy and comfort (Scherer, 2005; Hudlicka, 2003). Speech analysis may not always be suitable as not all learning environments are dialogue based. Table 3 below gives a brief comparative overview.

Table 3. Overview of the three dominant channels of nonverbal behavior

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

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

Given the pre-eminence of facial signs in human communication the face is a natural choice for inferring affective states. With the latest computer vision techniques facial information can be detected and analyzed unobtrusively and automatically in real-time requiring no specialized equipment except a simple video capture device. This makes facial affect analysis an attractive choice for evaluating learner states and together with head gestures is selected as the modality for affect inference in our system. Moreover, although recent studies have looked at the divergence in emotional information across modalities (Cowie, 2009; Cowie & McKeown, 2009), affect inference from facial expressions has been found to be consistent with other indicators of emotion (Cohn, 2006). However, facial expressions are not simple read-outs of mental states and their interpretation being context-driven is largely situational. Computer tutors can exploit this aspect to infer affective states from observed facial expressions using the knowledge state and navigation patterns from the learning situation as supporting evidence. Given the requirements of an affective computer tutor, the visual modality thus has a great potential for evaluating learner states thereby
facilitating an engaging and optimal learning experience. It is for these reasons that the visual modality was selected for affect analysis in this work.

Context and Corpora

For a meaningful interpretation and to ensure ecological validity, it is essential to study the occurrence and nature of affect displays *in situ*, as they occur. Although a number of face databases exist, these are mostly posed or recorded in scripted situations that may not be entirely relevant in a learning situation. We know that emotions are situated, have contextual antecedents and are influenced by social consequences. Knowledge of the learning setting is important then to ground a research work in a specific context and help assess its generalisation ability. The nature and dynamics of emotions in a solo learning setting e.g., Conati and Zhou (2002), Conati (2002), will no doubt differ from those generated within an agent-based learning environment like in Jaques and Vicari (2007), Heylen, Ghijsen, Nijholt, and Akker (2005), Kapoor, Burleson, and Picard (2007), or with those that involve dialogue, as in D'Mello, Picard, & Graesser (2007), Litman and Forbes (2003).

The nature of affect and its dependence on context thus makes the choice of a learning environment an important one. As such, we decided to use a solo, one to one learning setting for our study. By focusing on a self-regulated learning model our objective was to minimise the potential effects of design variables like instructional strategy, process of communication, collaboration, presence of an embodied agent, etc; in the assessment and interpretation of emotional experience. A data collection exercise was undertaken in which eight participants were video-recorded while doing two computer-based learning tasks. About four hours of data was collected which underwent three annotation levels to finally get samples of the six emotion groups. The pre-selected emotion categories were validated during the annotation process except for the addition of *surprise* which did not feature in the original list of relevant emotions. *Surprise* was added to the list of domain relevant affect states because of its frequent occurrence in the data as noted by the coders. The set of affect states thus represents the range of emotions observed in the collected video data. Furthermore, the proportion of labelled instances showed the predominance of confusion followed by surprised, interested, happy, bored and annoyed. A detailed description of the data collection and annotation process appears elsewhere in (Afzal & Robinson, 2009). Note that the emotions groups of annoyed and anxious had very few representative samples to merit proper statistical analyses and were therefore not included in the subsequent analysis. The compiled dataset was used as the ground-truth for the training of a fully automatic parallel inference system designed to continuously and unobtrusively model emotions in real-time, as described in the following sections.

Representation and Measurement of Facial Motion

Machine perception of affect can be posed as a pattern recognition problem, typically classification or categorisation, where the classes or categories correspond to the different emotion groups. Determining an optimal feature representation is then crucial to overall classifier design. Defining features implies developing a representation of the input pattern that can facilitate classification. Domain knowledge and human instinct play an important role in identifying such descriptors. Although a large body of work dealing with human perception of facial expressions exists, there have been very few attempts to develop objective methods for quantifying facial movements (Essa, 1997). One of the most significant works in this area is that of Ekman & Friesen (1978) who have devised a system for objectively coding all visually distinguishable facial movements called the *Facial Action Coding System* (FACS).

FACS associates facial expression changes with the actions of the muscles that produce them and by enumerating 44 action units (AUs) it encodes all anatomically possible facial expressions, singly or in combination. Since the AUs are purely descriptive measures of facial expression changes, they are independent of interpretation and provide a useful grammar for use as feature descriptors in expression studies as this. FACS remains a popular method for measuring facial behaviour and continues to have normative significance in automatic facial expression analysis as the only psychometrically rigorous and comprehensive grammar of facial actions available (Cohn, 2006).

The 2D face model (see Figure 1) of the Nevenvision FaceTracker is used to characterize the facial motion in terms of AUs. This FaceTracker is a state-of-art facial feature point tracking technology and requires no manual pre-processing or calibration. It is resilient to limited out-of-plane motion, can deal with a wide range of physiognomies
and can also track faces with glasses or facial hair. The FaceTracker uses a generic face template to capture the movement of 22 facial feature points over the video sequences. The displacement of these feature points over successive frames encodes the motion pattern of the face AUs in a feature vector. To remove the effects of variation in face scale and projection, the distance measurements are normalized with respect to a positional line connecting the inner eyes in the first frame. Statistically, the representative values of AUs in terms of local concentration (median) and dispersion (standard deviation) are selected as parameters, along with the first temporal derivative corresponding to speed as an additional attribute. The inclusion of speed helps qualify the dynamic information in expression changes and is found to increase the interpretive power and performance of classifiers (Tong, Liao, & Ji, 2007; Pantic & Patras, 2006; Ambadar, Schooler, & Cohn, 2005).

Preliminary statistical analysis using WEKA followed by a comparison of two popular class binarisation strategies namely, the one-versus-all approach (OvA), and the pairwise or round robin approach (AvA), indicated enhanced classification performance using OvA. Class binarisation reduces the complexity of multi-class discrimination by transforming the original multi-class learning problem \(y=\{1,2,...,k\}\) into a series of binary problems and evaluates the overall performance by combining the multiple outputs (Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006). OvA is the most common binary classification approach based on the assumption that there exists a single (simple) separator between a class and all others. Learning proceeds by learning \(k\) independent binary classifiers, one for each class, where the positive training examples are those belonging to the class while the negative examples are formed by the union of all other classes (Park & Furnkranz, 2007; Har-Peled, Roth, & Zimak, 2003). OvA classifiers operate by a winner-takes-all strategy so that a new example is assigned to the class corresponding to the maximum output value from the \(k\) binary classifiers. The OvA scheme is powerful because of its conceptual simplicity and comparative performance relative to other binarisation methods but at lower computational costs (Rifkin & Klautau, 2004). Applying OvA strategy therefore creates six binary classifiers, each differentiating a class from all others. Positive and negative samples of relevant emotion classes are randomly sub-sampled to learn each binary classifier. From an application perspective, a classifier should also be able to deal with real-time data input and be able to model the temporal evolution of facial expressions. To address this, we now describe the classification system that uses a class of dynamic probabilistic network to model the temporal signatures of the six emotion classes under study using an OvA design.

**Discriminative HMMs**

Hidden Markov Models (HMMs) are a popular statistical tool for modeling and recognition of sequential data and have been successfully used in applications like speech recognition, handwriting recognition, gesture recognition and even automatic facial expression classification (Rabiner, 1989). Based on whether the observations being modelled are discrete or continuous, HMMs can be constructed as having discrete or continuous output probability densities. Since it is intuitively more advantageous to use continuous HMMs (CHMM) to model continuous observations, we use CHMMs to model the temporal patterns of the emotion classes under study. Following OvA design, we use HMMs in a discriminatory manner which implies learning one HMM per class, running all HMMs in parallel and choosing the model with the highest likelihood as the most likely classification for a sequence. This way an HMM models the temporal signature of each emotion class so that the likelihood that an unseen sequence is emitted by each of the models can be estimated and be classified as belonging to the model most likely to have produced it (Oliver & Horvitz, 2005; Cohen, Sebe, Garg, Chen, & Huang, 2003).

Thus, a bank of HMMs is learned using the Baum-Welch algorithm (Rabiner, 1989) over the sample sequences. During training, the Gaussian mixtures with diagonal covariance are used and the initial estimates of state means and covariance matrices are found by k-means clustering. For classification, all HMMs are run in parallel and the forward-backward procedure (Rabiner, 1989) is used to select the model with the highest likelihood as the true class. See Figure 1 for illustration. The observation vector for the HMMs consists of the position and speed parameters sampled over a sliding-window of five frames. This results in a multi-dimensional feature vector characterizing a filtered pattern sequence of the temporally evolving facial and head motions. PCA is used to extract salient features and reduce dimensionality.

The overall classification accuracy is estimated by averaging the true positive rate using tenfold cross-validation. To determine the best performance empirically, recognition accuracies are computed by varying the free parameters - the number of states and the number of Gaussian mixtures. Table 4 shows the detailed confusion matrix for the best
classification achieved. Overall, for a mean false positive rate of just 1.01% the best average accuracy of 94.96% is obtained with eleven states and four Gaussian mixtures. Happy and surprised attain perfect true positive rates while others show satisfactory recognition. Individual classes attain optimal performance at varying number of states and mixtures suggesting that individual emotions have their own temporal signatures and can be modeled by aligning them along their optimal topologies. This, along with an assessment of the generalization ability, needs to be determined in future work as it requires evaluation of the system on a database that is comparable at least in terms of context and recording conditions.

Bank of HMM Models $\lambda_c, 1 \leq c \leq 6$

$$c^* = \text{argmax} \left[ \Pr(O_t|\lambda_c) \right]$$

$\lambda_1$
$$t-4 \quad t-3 \quad t-2 \quad t-1 \quad t \quad \Pr(E_1|\lambda_1)$$

$\lambda_2$
$$t-4 \quad t-3 \quad t-2 \quad t-1 \quad t \quad \Pr(E_2|\lambda_2)$$

$\lambda_6$
$$t-4 \quad t-3 \quad t-2 \quad t-1 \quad t \quad \Pr(E_6|\lambda_6)$$

Figure 1. Feature point measurements fed to the bank of discriminative HMMs

<table>
<thead>
<tr>
<th>Predicted</th>
<th>bored</th>
<th>confused</th>
<th>happy</th>
<th>interested</th>
<th>neutral</th>
<th>surprised</th>
<th>total</th>
<th>TP %</th>
</tr>
</thead>
<tbody>
<tr>
<td>bored</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>93.8</td>
</tr>
<tr>
<td>confused</td>
<td>0</td>
<td>57</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>59</td>
<td>96.6</td>
</tr>
<tr>
<td>happy</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>100.0</td>
</tr>
<tr>
<td>interested</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>87.1</td>
</tr>
<tr>
<td>neutral</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>24</td>
<td>1</td>
<td>26</td>
<td>92.3</td>
</tr>
<tr>
<td>surprised</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>32</td>
<td>100.0</td>
</tr>
<tr>
<td>total</td>
<td>15</td>
<td>62</td>
<td>34</td>
<td>27</td>
<td>24</td>
<td>34</td>
<td>196</td>
<td>95.0</td>
</tr>
<tr>
<td>FP %</td>
<td>0.0</td>
<td>3.6</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Best performance of discriminative HMMs.

Summary and Conclusions

A consistent theme that emerges from education literature is that teaching and learning are essentially emotional practices. Learners experience a wide range of both positive and negative emotions, and these influence their cognitive functioning and performance. Access to emotions is then important to ensure optimal learning, more so in the case of computer-based learning environments where the learner’s motivation is an important determinant of engagement and success. However, automatic measurement of affect is a challenging task. Emotions consist of multiple components that may include intentions, action tendencies, appraisal, other cognitions, central and peripheral changes in physiology, and subjective feelings. As a result they are not directly observable and can only be inferred from expressive behaviour, self-report, physiological indicators, and context (Cohn, 2006).

This paper has outlined the problem space with respect to the application of affect-sensitive technologies in computer-based learning. Building on a discussion of studies highlighting the relevance of emotions in learning, the different techniques for measuring emotions and recent advances in automatic recognition and/or prediction of affect in learning contexts were discussed. Six categories of pertinent affect states were identified; the visual modality for affect modelling was preferred given the requirements of a viable measurement technique; and a bottom-up analysis approach based on context-relevant data was adopted. Finally, a dynamic classification system using a bank of discriminative HMMs was described while the underlying differences in the temporal signatures of the individual affect states was also highlighted. Trained on the compiled corpus, it is designed to model multiple emotions simultaneously in real-time using automatic facial feature point tracking and will be optimized in future work on dataset(s) from potential learning contexts.
Acknowledgements

This research was supported by the Gates Cambridge Scholarships and the Overseas Research Studentship of the University of Cambridge.

References


