Automatic Detection of Learner's Affect from Gross Body Language

Sidney D'Mello and Art Graesser

The University of Memphis

Corresponding Author:

Sidney D'Mello

Department of Computer Science

209 Dunn Hall

The University of Memphis

Memphis, TN 38152, USA.

Email: sdmello@memphis.edu

Abstract

We explored the reliability of detecting learners' affect by monitoring their gross body language (body position and arousal) during interactions with an Intelligent Tutoring System called AutoTutor. Training and validation data on affective states were collected in a learning session with AutoTutor, after which the learners' affective states (i.e., emotions) were rated by the learner, a peer, and two trained judges. An automated body pressure measurement system was used to capture the pressure exerted by the learner on the seat and back of a chair during the tutoring session. We extracted two sets of features from the pressure maps. The first set focused on the average pressure exerted, along with the magnitude and direction of changes in the pressure during emotional experiences. The second set of features monitored the spatial and temporal properties of naturally occurring pockets of pressure. We constructed five data sets that temporally integrated the affective judgments with the two sets of pressure features. The first four datasets corresponded to judgments of the learner, a peer, and two trained judges, whereas the final data set integrated judgments of the two trained judges. Machine learning experiments yielded affect detection accuracies of 73%, 72%, 70%, 83%, and 74% respectively (chance=50%) in detecting boredom, confusion, delight, flow, and frustration, from neutral. Accuracies involving discriminations between two, three, four, and five affective states (excluding neutral) were 71%, 55%, 46%, and 40% with chance rates being 50%, 33%, 25%, and 20% respectively.

1. Introduction

Verbal and non-verbal channels show a remarkable degree of sophisticated coordination in human-human communication. While the linguistic channel mainly conveys the content of the message, non-verbal behaviors play a fundamental role in expressing the affective states, attitudes, and social dynamics of the communicators. Although ubiquitous to human-human interactions, the information expressed through non-verbal communicative channels is largely ignored in human-computer interactions. Simply put, there seems to be a great divide between the highly expressive human and the perceptually deficit computer.

In an attempt to alleviate this shortcoming in human-computer interactions, the last decade has witnessed a surge of research activities that are aimed at narrowing the communicative bandwidth between the human and the computer. Notable among these endeavors is the rapidly growing area of Affective Computing. Affective Computing is a subfield of Human Computer Interaction that focuses on the affective states (feelings, moods, emotions) of the user (Picard 1997). This emphasis of affect is quite critical because emotions are inextricably bound to cognition (Dweck 2002; Mandler 1984). Cognitive activities such as causal reasoning, deliberation, goal appraisal, and planning processes operate continually throughout the experience of emotion.

The primary practical goal of Affective Computing is to create technologies that can monitor and respond to the affective states of the user (Picard 1997).

This is achieved by integrating the affective states of a user into the decision cycle of the interface in order to provide more effective, user-friendly, and naturalistic applications (Bianchi-Berthouze and Lisetti 2002; Conati 2002; de Rosis 2002; Lisetti and Gmytrasiewicz 2002; Prendinger and Ishizuka 2005; Whang, Lim, and Boucsein 2003). Progress in achieving the primary goal requires an interdisciplinary integration of computer science, psychology, artificial intelligence, and artifact design.

Although, there are a number of obstacles that need to be overcome before functional affect-sensitive computer interfaces can be realized, the success of any affect-sensitive interface will ultimately depend upon the accuracy by which the user's affect can be detected. These interfaces ultimately are guided by the design goal of narrowing the communicative gap between the emotionally challenged computer and the emotionally rich human. Expectations are raised when humans recognize that a computer system is attempting to communicate at their level (i.e. with enhanced cognitive and emotional intelligence) far beyond traditional interaction paradigms (i.e. WIMP - window, icon, menu, pointing device). When these expectations are not met, users often get discouraged, disappointed, or even frustrated (Norman 1994; Shneiderman and Plaisant 2005). Therefore, robust recognition of the users' emotions is a crucial challenge that is hindering major progress towards the larger goal of developing affect-sensitive interfaces that work.

Consequently, the last decade has been ripe with technologies that attempt to automatically detect the affective states of a user. Many of these technologies analyze physiological signals for emotion detection (Rani, Sarkar, and Smith 2003; Picard, Vyzas, and Healey 2001; Whang, Lim, and Boucsein 2003). One potential pitfall to this approach is the reliance on obtrusive sensing technologies, such as skin conductance, heart rate monitoring, and measurement of brain activities. These obtrusive physiological sensors are acceptable in some applications and it is true that users habituate to the presence of these sensors, but they are not satisfactory in environments because the sensors distract users and interfere with the primary tasks. This has motivated designers of affect-sensitive technologies to focus on facial feature tracking and acoustic-prosodic vocal features, two technologies that are unobtrusive (see Pantic and Rothkrantz 2003 for a comprehensive review and the proceedings of ACII 2007 edited by Paiva, Prada, and Picard 2007 for recent updates).

State-of-the-art affect detection systems have overlooked posture as a serious contender when compared to facial expressions and acoustic-proposed features, so an analysis of posture merits more close examination. There apparently are some benefits to using posture as a means to diagnose the affective states of a user (Bull 1987; de Meijer 1989; Mehrabian 1972). Human bodies are relatively large and have multiple degrees of freedom, thereby providing them with the capability of assuming a myriad of unique configurations (Bernstein 1967). These static positions can be concurrently combined and temporarily aligned

with a multitude of movements, all of which makes posture a potentially ideal affective communicative channel (Coulson 2004; Montepare, Koff, Zaitchik, and Albert 1999). Posture can offer information that is sometimes unavailable from the conventional non-verbal measures such as the face and paralinguistic features of speech. For example, the affective state of a person can be decoded over long distances with posture, whereas recognition at the same distance from facial features is difficult or unreliable (Walk and Walters 1988). Perhaps the greatest advantage to posture based affect detection is that gross body motions are ordinarily unconscious, unintentional, and thereby not susceptible to social editing, at least compared with facial expressions, speech intonation, and some gestures. Ekman and Friesen (1969), in their studies of deception, have coined the term *nonverbal leakage* to refer to the increased difficulty faced by liars, who attempt to disguise deceit, through less controlled channels such as the body when compared to facial expressions. Furthermore, although facial expressions were once considered to be the objective gold standard for emotional expression in humans, there is converging evidence that disputes the adequacy of the face in expressing affect (see Barrett 2006 for a comprehensive review). At the very least, it is reasonable to operate on the assumption that some affective states are best conveyed through the face, while others are manifested through other non-verbal channels. This paper adopts this position and aspires to investigate the potential of gross body movements as a viable channel to detect affect.

One option towards automated posture analysis is to use cameras and associated computer vision techniques to monitor body position and movement of a user. However, this approach is riddled with the problems that accompany nearly all computer vision-based applications, such as lighting, background conditions, camera angles, and other factors (Mota and Picard 2003). Fortunately, there is a relatively new sensor that circumvents these challenges. In 1997 Tekscan™ released the Body Pressure Measurement System (BPMS) which consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The system provides pressure maps in real time that can be analyzed for a variety of applications. For example, Tan and colleagues demonstrated that the BPMS system could be used to detect several static postures (leaning right, right leg crossed, etc) quite reliably with an average recognition accuracy of 85% (Tan, Lu, and Pentland 1997).

Mota and Picard (2003) reported the first substantial body of work that used the automated posture analysis via the BPMS system to infer the affective states of a user in a learning environment. They analyzed temporal transitions of posture patterns to classify the interest level of children while they performed a learning task on a computer. A neural network provided real time classification of nine static postures (leaning back, sitting upright, etc) with an overall accuracy of 87.6%. Their system then recognized interest (high interest, low interest, and taking a break) by analyzing posture sequences over a 3 second interval, yielding an overall accuracy of 82.3%.

In this paper we explore the possibility of using posture to automatically detect the affective states of college students during a tutoring session with the AutoTutor learning environment (Graesser et al. 2001; VanLehn, Graesser, et al. 2007. We focus on Intelligent Tutoring Systems (ITSs) because they represent a domain that is on the forefront of affect-sensitive interface research (Conati 2002; D'Mello et al. 2005; Kort, Reilly, and Picard 2001; Litman and Forbes-Riley 2004). Affect-sensitive ITSs operate on the fundamental assumption that affect is inextricable linked to cognition. There is also some evidence that tracking and responding to human emotions on a computer increases students' persistence and learning (Aist et al. 2002; Kim 2005; Linnenbrink and Pintrich 2002). Hence, affect-sensitive ITSs attempt to incorporate the affective and cognitive states of a learner into their pedagogical strategies to increase engagement, reduce attrition, boost self-efficacy, and ultimately promote active learning.

The larger goal of the project is to reengineer AutoTutor to enable it to adapt to the learner's affective states in addition to cognitive states. This adaptation would increase the bandwidth of communication and allow AutoTutor to respond at a more sophisticated metacognitive level. Quite clearly, robust affect recognition is an important requirement for the affect-sensitive AutoTutor because the system will never be able to respond to a learners' emotion if it cannot sense the emotion. Therefore, we explore the potential of affect-detection from body language as an incremental step towards this goal.

Our research differs from the previous research in automated affect detection from posture in four significant ways. First, much of the research in affect-detection has focused on the "basic" emotions (i.e. anger, fear, sadness, enjoyment, disgust, and surprise, Ekman and Friesen 1978). While these basic emotions are ubiquitous in everyday experience, there is a growing body of evidence that suggests that they rarely play a significant role in deep learning of conceptual information (D'Mello, et al. 2006; Graesser et al. 2006; Kort, Reilly, and Picard 2001). While it is conceivable that the more basic extreme emotions identified by Ekman are relevant to learning in some circumstances, as discussed later, the present study concentrated on emotions that we already know are relevant to learning with AutoTutor. Second, we monitored a larger set of affective states than Mota and Picard (2003), specifically the affective states identified by Craig at al. (2004): boredom, flow, confusion, frustration, delight, and neutral. It is important to consider a larger set of affective states that encompass the entire gamut of learning (Conati 2002) because a person's reaction to the presented material can change as a function of their goals, preferences, expectations and knowledge states. Third, some additional considerations arise because we monitored college students rather than children as in the Mota & Picard (2003) work. Children are much more active than the college students, so the algorithms used to detect affective states may differ. The movements of college students are more subtle so it is important to pick up fleeting transitions in body pressure. We ultimately developed two different

methods to infer affect from body movement. Both of these methods monitor gross body movements, rather than explicit postures, and hence did not require an additional training phase for static posture detection, as in the Mota and Picard (2003) work. The fourth difference between this research and other efforts is the method of establishing ground-truth categories of affect, which is a requirement for most supervised learning methods. A number of researchers have relied on a single operational measure when inferring a learner's emotion, such as self reports (De Vicente and Pain 2002; Klein, Moon, and Picard 2002; Matsubara and Nagamachi 1996) or ratings by independent judges (Liscombe, Riccardi, and Hakkani-Tür 2005; Litman and Forbes-Riley 2004; Mota and Picard 2003). In contrast, we propose the combination of several different measures of a learner's affect. Our measures of emotion incorporate judgments made by the learner, a peer, and two trained judges, as will be elaborated later.

The paper is organized in four sections. First we describe a study that collected data from the BPMS system and affect labels from multiple judges in order to train and validate the posture-based affect classifier. Second, we describe the BPMS system in some detail as well as two sets of posture features used to develop the affect-detector. Third, the Results section begins with a description of a series of experimental simulations that attempt to measure affect recognition accuracy. We compare performance of affect classification from posture with classification accuracies obtained via a conversational sensor and facial feature tracking. And fourth, we present a summary of our major

findings, limitations of our methodology, potential resolutions, and future work. Our ultimate goal is to explore how the learner's affective states may be integrated into AutoTutor's pedagogical strategies and thereby improve learning.

2. Empirical Data Collection: The Multiple Judge Study

Modeling affect involves determining what emotion a learner is experiencing at particular points in time. Emotion is a construct (i.e., an inferred conceptual entity), so one can only approximate its true value. Therefore, in contrast to a single operational measure to inferring a learner's emotion, we propose the combination of several different measures of a learner's affect. Our measures consist of emotion judgments made by the learner, a peer, and two trained judges. Employing multiple measures of affect is compatible with the standard criterion for establishing convergent validity (Campbell and Fiske 1959).

2.1. Participants

The participants were 28 college students from a Southern university in the United States of America.

2.2. Materials

2.2.1. Sensors. Three streams of information were recorded during the participant's interaction with AutoTutor. A video of the participants face was captured using the IBM® blue-eyes camera (Morimoto, Koons, Amir, and Flickner 1998). Posture patterns were captured by the Tekscan® Body Pressure Measurement System (Tekscan 1997) which is described in some detail below.

A screen-capturing software program called Camtasia Studio (developed by TechSmith) was used to capture the audio and video of the participant's entire tutoring session with AutoTutor. The captured audio included the speech generated by the AutoTutor animated conversational agent.

2.2.2. AutoTutor. AutoTutor is a fully automated computer tutor that simulates human tutors and holds conversations with students in natural language (Graesser et al. 2001; VanLehn, Graesser, et al. 2007). AutoTutor attempts to comprehend the students' natural language contributions and then responds to the students' typed input with adaptive dialogue moves similar to human tutors. AutoTutor helps students learn by presenting challenging problems (or questions) from a curriculum script (a set of questions, ideal answers, and expected misconceptions) and engaging in a mixed-initiative dialogue while the learner constructs an answer.

AutoTutor has different classes of dialogue moves that manage the interaction systematically. AutoTutor provides *feedback* on what the student types in (positive, neutral, or negative feedback), *pumps* the student for more information ("What else?"), *prompts* the student to fill in missing words, gives *hints*, fills in missing information with *assertions***,** identifies and corrects *misconceptions* and erroneous ideas, *answers* the student's questions, and *summarizes* topics. A full answer to a question is eventually constructed during this dialogue, which normally takes between 30 and 100 turns between the student and tutor for one particular problem or main question.

The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in over a dozen experiments on college students as learners for topics in introductory computer literacy (Graesser, Lu et al. 2004) and conceptual physics (VanLehn, Graesser, et al. 2007). Tests of AutoTutor have produced gains of .4 to 1.5 sigma (a mean of .8), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor. From the standpoint of the present study, we will take it as given that AutoTutor helps learning whereas our direct focus is on the emotions that occur in the learning process.

2.3. Procedure

2.3.1. Interacting with AutoTutor. The participants interacted with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, internet, or operating systems. During the interaction process we recorded data from the three sensors listed above. Participant completed a multiple choice pre-test before interacting with AutoTutor and a multiple choice post-test after the tutoring session.

2.3.2. Judging Affective States. The affect judging process was conducted by synchronizing and displaying to the judges the video streams from the screen and the face. Judges were instructed to make judgments on what affective states were present at 20-second intervals; at each of these points, the video automatically paused (freeze-framed). Additionally, if participants were experiencing more than one affective state in a 20-second block, judges were

instructed to mark each state and indicate which was most pronounced. However, in these situations only the more prominent affective state was considered in the current analyses. At the end of the study, participants were asked to identify any affective states they may have experienced that were not included in the specified list of 7 emotions. However, a cursory look at the data did not reveal any new affective states.

Four sets of emotion judgments were made for the observed affective states of each participant's AutoTutor session. First, for the *self* judgments, the participant watched his or her own session with AutoTutor immediately after having interacted with the tutor. Second, for the *peer* judgments, participants returned approximately a week later to watch and judge another participant's session on the same topic in computer literacy. Finally, two additional judges (called *trained judges*) judged all of the sessions individually; these trained judges had been trained on how to detect facial action units according to Paul Ekman's Facial Action Coding System (FACS) (Ekman and Friesen 1978). The trained judges also had considerable experience interacting with AutoTutor. Hence, their emotion judgments were based on contextual dialogue information as well as the FACS system.

A list of the affective states and definitions was provided for all judges. The states were frustration, confusion, flow, delight, surprise, boredom, and neutral[1].

[1] An interesting alternative to the categorical affect coding scheme used here is a dimensional based scheme. In such a scheme affect is coded along dimensions of valence and intensity rather than specific emotion categories (e.g. Bianchi-Berthouze et al. 2006). However, we used the categorical coding scheme in this research

Frustration was defined as dissatisfaction or annoyance. Confusion was defined as a noticeable lack of understanding, whereas flow was a state of interest that results from involvement in an activity. Delight was a high degree of satisfaction. Surprise was defined as wonder or amazement, especially from the unexpected. Boredom was defined as being weary or restless through lack of interest. Neutral was defined as no apparent emotion or feeling.

2.4. Proportions of Emotions Experienced

We examined the proportion of judgments that were made for each of the affect categories, averaging over the four judges. The most common affective state was neutral (.32), followed by confusion (.24), flow (.17), and boredom (.16). The frequency of occurrence of the remaining states of delight, frustration and surprise were significantly lower, comprising .04, .06, and .02 of the observations respectively. This distribution of affective states implies that most of the time learners are either in a neutral state or in a subtle affective state (boredom or flow). There is also a reasonable amount of confusion since the participants in this study were typically low domain knowledge students as indicated by their low pretest scores.

2.5. Agreement between Affect Judges

Interjudge reliability was computed using Cohen's kappa for all possible pairs of judges: self, peer, trained judge1, and trained judge2. Cohen's kappa

---

because specific emotional categories are needed to fortify AutoTutor with the ability to respond to the learners' affective states.

.

measures the proportion of agreements between two judges with correction for baserate levels and random guessing (Cohen 1960). There were 6 possible pairs altogether. The kappas were reported in Graesser et al. (2006): self-peer (.08), self-judge1 (.14), self-judge2 (.16), peer-judge1 (.14), peer-judge2 (.18), and judge1-judge2 (.36). While these kappas appear to be low, the kappas for the 2 trained judges are on par with data reported by other researchers who have assessed identification of emotions by humans (Ang et al. 2002; Grimm et. al. 2006; Litman and Forbes-Riley 2004; Shafran, Riley, and Mohri 2003). For example, Litman and Forbes-Riley (2004) reported kappa scores of .40 in distinguishing between positive, negative, negative and neutral affect. Ang et al. (2002) reported that human judges making a binary frustration-annoyance discrimination obtained a kappa score of .47. Shafran, Riley, and Mohri achieved kappa scores ranging from .32 - .42 in distinguishing among 6 emotions. In general, these results highlight the difficulty that humans experience in detecting affect.

3. Architecture of the Posture Based Affect Detector

3.1. The Body Pressure Measurement System (BPMS)

The BPMS system, developed by Tekscan™ (1997), consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The pad is paper thin with a rectangular grid of sensing elements that is enclosed in a protective pouch. Each sensing element provides a 8-bit pressure output in

mmHg. Our setup had one sensing pad placed on the seat of a Steelcase™ Leap Chair and another placed on the back of the chair (see Figure 1A).

INSERT FIGURE 1 ABOUT HERE

The output of the BPMS system consists of a 38 × 41 pressure matrix (rows × columns) for each pad. Each cell in the matrix monitors the amount of pressure exerted on a single sensing element (see Figure 1B). During the tutoring intervention, at each sampling instance (1/4 second for our study), matrices corresponding to the pressure exerted on the back and the seat of the chair were recorded for offline analyses.

3.2. High Level Pressure Features

This feature set monitored the average pressure exerted on the back and seat of the chair along with the magnitude and direction of changes in pressure over a brief time window. Several features were computed by analyzing the pressure maps of the 28 participants recorded in the study. We individually computed 6 pressure-related features and 2 features related to the pressure coverage for the back and the seat, yielding 16 features in all. Each of the features was computed by examining the pressure map at the time of an emotional episode (called the *current frame* or the frame at time *t*).

Perhaps the most significant pressure related feature was the *average pressure,* which measured the mean pressure exerted in the current frame. This was computed by summing the pressure exerted on each sensing element and dividing the sum by the total number of elements. The average pressure is

expressed in Equation (1) where R is the number of rows in the pressure matrix, C the number of columns, and p_{ij} is the pressure of a sensing element in row i and column j. For the current study, $R = 38$ and $C = 41$.

$$\mu = \frac{1}{R \times C} \sum_{i=1}^{R} \sum_{j=1}^{C} p_{ij} \qquad \text{(Equation 1)}$$

We introduced another feature to detect the incidence of sharp forward versus backward leans, which ostensibly occurs when a learner is modulating his or her engagement levels. This feature measured the pressure exerted on the top of the back and seat pads. This was obtained by first dividing the pressure matrix into 4 triangular regions of equal area (see Figure 1C) and then computing the average pressure for the *top* triangular region. For the seat, this feature measured the force exerted on the frontal portion of the chair (sharp forward lean), while for the back it indicated whether the shoulder blades of the learner were on the back rest of the chair (heightened backward lean).

The next two features measured the direction of pressure change. These include the *prior change* and *post* change, which measured the difference between the average pressure in the current frame (t) and the frame J seconds earlier ($t - J$) and K seconds later, ($t + K$) respectively (See Equations 2 and 3). For the current analyses, $J = K = 3$ seconds. A positive prior change value is indicative of an increase in the pressure exerted, while a positive post change value reflects a reduction in pressure.

$$\Delta_{prior} = \mu_t - \mu_{t-J} \qquad \text{(Equation 2)}$$

$$\Delta_{post} = \mu_t - \mu_{t+K} \quad \text{(Equation 3)}$$

The *reference change* (See Equation (See Equation) measured the difference between the average pressure in the current frame (*t*) and the frame for the last known affective rating (*r*). The motivation behind this measure was to calibrate the impact of the last emotional experience on the current affective state. It should be noted that unlike *J* and *K ,* which were used to compute the prior and post changes respectively, *r* is not a constant time difference. Instead *r* varies in time across different affective experiences. For the current analyses, *r* was 20 seconds for a majority of the instances since affect judgments were elucidated every 20 seconds. However, since the affect judges voluntarily offered judgments between the 20 second time slots, in several cases, $r < 20$ seconds.

$$\Delta_{ref} = \mu_t - \mu_{t-r} \quad \text{(Equation 4)}$$

Finally, the *average pressure change* ($a_{pressure}$) measured the mean change in the average pressure across a predefined window of length N (see Equation 5). The window was typically 4 seconds, which spanned two seconds before and two seconds after an emotion judgment.

$$a_{pressure} = \frac{1}{N}\sum_{t=1}^{N}\left|\mu_t - \mu_{t+1}\right| \quad \text{(Equation 5)}$$

The two coverage features examined the proportion of non-zero sensing units (*average coverage*) on each pad along with the mean change of this feature across a 4-second window (*average coverage change*). The computations for average coverage can be depicted as follows. Consider x_{ij} to

be an indicator variable that determines whether the pressure (p_{ij}) on sensing element *ij* is non-zero. Then:

$x_{ij} = 1, \quad \text{if } p_{ij} > 0$

$x_{ij} = 0, \quad \text{if } p_{ij} = 0$

The average coverage was the proportion of x_{ij} values that were non-zero as indicated by Equation 6. Analogous to Equation 4, the average coverage change is expressed in Equation 7 as:

$$c = \frac{1}{R \times C} \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} \qquad \text{(Equation 6)}$$

$$a_{coverage} = \frac{1}{N} \sum_{t=1}^{N} \left| c_t - c_{t+1} \right| \qquad \text{(Equation 7)}$$

3.3. Spatial-Temporal Features

The second set of features used for the posture affect detector involved monitoring the spatial distribution of pressure contours and the magnitude by which they changed over time. Pressure contours were obtained by clustering the pressure maps for the back and seat of the chair using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). The input to the EM algorithm were pressure values for each of the 1558 (38 × 41) pressure sensing elements from each sensor pad (i.e. back and seat). Each sensing element, along with the corresponding pressure exerted on it, was represented as a 3 dimensional point. The first two dimensions represented the relative position of the sensing element (i.e. its X and Y coordinate) on the sensor map while the third dimension was the pressure exerted on it. The EM algorithm was

configured to output 4 clusters based on earlier findings by Mota and Picard (2003) and preliminary experimental simulations in which the number of clusters was varied (k = 2, 3, 4, 5, 6). Figure 3 shows an example of the clustering data from the pressure maps with the EM algorithm.

Since each data point was 3-dimensional (i.e., X and Y position co-ordinates and pressure), each cluster was represented by a 3D mean, a 3D standard deviation, and a prior probability (i.e. proportion of the 1558 data points that are included in the cluster). Consequently, we extracted 7 features from each cluster: 3 for the means, 3 for the standard deviations, and 1 for the prior probability. By tracking 4 clusters on each pad we obtained 28 (4 × 7) features in all. Additionally, since we were tracking both the back and the seat we obtained 28 × 2 = 56 features.

The aforementioned 56 features provide a snapshot of the spatial distribution of pressure exerted on the back and the seat of the chair while the learner interacts with AutoTutor. In order to obtain a measurement of arousal we tracked the change in each pressure contour (cluster) over a short time. In particular, the pressure contours across a 4 second window were monitored and the means and standard deviations of each of the 56 features were computed. Therefore, effective dimensionality was 112. In this manner the features selected were spatial (distribution of pressure on pad) and temporal (changes in distribution over time).

INSERT FIGURE 2 ABOUT HERE

3.4. Hierarchical Classification via an Associated Pandemonium

Perhaps the simplest method to develop an affect classifier is to experiment with several standard classifiers (neural networks, Naïve Bayes classifiers, etc) and select the one that yields the highest performance in collectively discriminating between the affective states of boredom, confusion, delight, flow, frustration, and neutral (surprise was not included in the affect detector since its occurrence was quite rare). However, since affect detection is a very challenging pattern recognition problem, on par with automatic speech recognition, developing a 6-way affect classifier that is sufficiently robust is quite challenging. An alternative approach is to divide the 6-way classification problem into several smaller 2 or 3 way classification problems.

One way to decompose the 6-way classification problem into several smaller problems is to include a collection of affect-neutral classifiers that would first determine whether the incoming posture pattern resonated with any one or more of the emotions (versus a neutral state). If there is resonance with only one emotion, then that emotion would be declared as being experienced by the learner. If there is resonance with 2 or more emotions, then a conflict resolution module would be launched to decide between the alternatives. This would essentially be a second-level affect classifier. If 3 or more emotions are detected, then the second level classifier would perform a 3-way classification task. In situations where the emotion expression is very murky, 4- or 5-way distinctions might be required as well.

INSERT FIGURE 3 ABOUT HERE

Figure 3 depicts the manner in which the various classifiers are organized and interact. Central to the model lie 5 affect-neutral classifiers, each performing an emotion vs. neutral discrimination. To the left we find 10 classifiers that specialize in making binary discriminations among the affective states. On the top there are 10 three-way emotion classifiers that are recruited when three or more affective states are detected by the affect-neutral classification layer. On the right the various possibilities for 4-way emotion classifiers are listed. Finally, the single 5-way classifier lies at the bottom.

As an example consider a situation where the affect-neutral classifiers output [Boredom, Neutral, Neutral, Neutral, Neutral]. In this case the Boredom-Neutral classifier has detect boredom while the other four affect-neutral classifiers have detected neutral. In this situation, we would declare that the learner is experiencing the boredom emotion. If instead, the output of the affect-neutral level is [Boredom, Neutral, Neutral, Flow, Neutral], where the Boredom-Neutral classifier detects boredom, the Flow-Neutral classifier detects flow, and the other affect-neutral classifiers declare neutral, then the Boredom-Flow binary discriminator would be recruited to resolve the conflict (see Figure 3).

Such a classification scheme is strongly motivated by the Pandemonium model (Selfridge 1959). It is expected that in most cases the level 1 classifier (affect-neutral) or a 2-way affect classifier would suffice. When more subtle distinctions are required from ambiguous input, a 3-way or higher order

classifier may also be necessary. However, 4-way or 5-way discriminations are expected to be much more rare, as discussed later.

4. Measuring the Accuracy of the Posture Based Affect Detector

In order to address the larger goal of extending AutoTutor into an affect-sensitive intelligent tutoring system, the need for real time automatic affect detection becomes paramount. An emotion classifier need not be perfect but should possess some modicum of accuracy. The subsequent analyses include a series of classification experiments to evaluate the reliability of affect detection from gross body language.

4.1. Experimental Setup

4.1.1. Data Set Creation. The data used for the analyses was from the multiple judge study that was described earlier in which 28 participants interacted with AutoTutor on topics in computer literacy. Posture feature vectors for each method (high-level pressure features and spatial-temporal pressure contours) were extracted from the BPMS data stored offline. The feature vector was then associated with an emotion category on the basis of each of the four human judges' affect ratings. More specifically, each emotion judgment was temporally bound to each posture based feature vector. This data collection procedure yielded four ground truth models of the learner's affect (self, peer, 2 trained judges), so we were able to construct four labeled data sets. When aggregated across each 32-minute session for each of the 28 participants, we obtained 2967,

3012, 3816, and 3723 labeled data points for the self, peer, trained judge 1, and trained judge 2, respectively.

Affect judgment reliabilities between the human judges presented above revealed that the highest agreement was obtained between the trained judges (kappa = .36). However, it is still not firmly established whether the trained judges or the self judgments are closer to ground truth. We addressed this issue by combining affect judgments from the trained judges in order to obtain a better approximation of the learner's emotion. In particular, an additional data set was constructed on the basis of judgments in which both trained judges agreed; this sample therefore focused on observations in which we had some confidence about the emotion. The frequencies of the emotions in each data set are listed in Table 1.

INSERT TABLE 1 ABOUT HERE

4.1.2. Classification Analyses. The Waikato Environment for Knowledge Analysis (WEKA) (Witten and Frank 2005) was used to comparatively evaluate the performance of various standard classification techniques (N = 17) in detecting affect from posture. The classification algorithms tested were selected from a list of categories including Bayesian classifiers (Naive Bayes and Naive Bayes Updatable), functions (Logistic Regression and Support Vector Machines), instance based techniques (K-Nearest Neighbor with k = 1 and k = 3, K*, Locally Weighted Learning), meta classification schemes (AdaBoost, Bagging Predictors, Additive Logistic Regression), trees (C4.5 Decision Trees,

Logistic Model Trees, REP Tree), and rules (Decision Tables, Nearest Neighbor Generalization, PART).

The classification analyses proceeded in two phases. In phase 1 the higher level pressure features (N = 16) were inputs to the classifiers. For phase 2, the spatial-temporal features (N = 112) were used to detect the affective states. Each phase was independent of the other since the primary goal here was to evaluate the accuracy of each method. Therefore, for each phase we evaluated the accuracy of each of the 17 classifiers in discriminating the affective states grouped in the 5 categories of the hierarchy (see Figure 3). There were 31 different classification experiments conducted for each feature set. These included 5 affect-neutral discriminations, 10 two-way discriminations, 10 three-way discriminations, 5 four-way discriminations, and a single 5-way discrimination.

We established a uniform baseline for the different emotions by randomly sampling an equal number of observations from each affective state category. This sampling process was repeated for 10 iterations and all reported reliability statistics were averaged across these 10 iterations. For example, consider the task of detecting confusion from neutral with affect labels provided by the self. In this case we would randomly select an equal number of confusion and neutral samples, thus creating a data set with equal prior probabilities of both these emotions. Each randomly sampled data set was evaluated on the 17

classification algorithms and reliability statistics were obtained using k-fold cross-validation (k = 10).

4.2. Trends in Classification Accuracy

A 3 factor repeated measures analysis of variance (ANOVA) was performed in order to comparatively evaluate the performance of the classifiers in detecting affect from the posture features. The first factor (*feature*) was the feature set used as input into the classifier and had two levels: *pressure* and *contours* for the high level pressure features and the spatial-temporal contours respectively. The second factor involved the *emotions* classified and was composed of 5 levels: affect-neutral discriminations (chance = 50%), 2-way affect discriminations (chance = 50%), 3-way affect discriminations (chance = 33%), 4-way affect discriminations (chance = 25%), and 5-way affect discriminations (chance = 20%). The third factor was the *judge* that provided the affect judgments. This factor also had 5 levels: self, peer, trained judges 1, trained judge 2, and observations in which trained judges agree. The unit of analysis for the 2×5×5 ANOVA was the accuracy obtained by each of the 17 classifiers. The kappa score was utilized as the metric to evaluate performance of each classifier because this metric partials out random guessing.

The ANOVA indicated that there were significant differences in kappa scores across all three factors, as well as for various interactions between the factors. On the basis of the ANOVA we report comparisons between the various levels of our three factors (feature, emotion, and judge). Figure 4 graphically

depicts the mean kappa score obtained from the emotion classification for each level of each factor of the ANOVA.

INSERT FIGURE 4 ABOUT HERE

4.2.1. Comparison across Feature Sets. The results of the ANOVA indicated that there was a statistically significant difference in classification accuracy obtained from each feature set, $F(1, 16) = 55$, $MSe = .003$, $p < .001$ (partial $\eta^2 = .775$). In particular, the classifiers based on the spatial-temporal contours ($M_{\text{CONTOUR}} = .20$) outperformed those trained on the higher level pressure features ($M_{\text{PRESSURE}} = .17$, see Figure 4A). However, the performance increments attributed to the spatial-temporal contours were marginal (an 18% increase in kappa over high level pressure features). This marginal improvement may be indicative of problems commonly associated with high dimensional feature spaces (N = 112 for spatial-temporal contours) due to cross-correlations among features.

4.2.2. Comparison across Emotions. The ANOVA revealed statistically significant differences in kappa scores for the emotions classified, $F(4, 64) = 269.14$, $MSe = .002$, $p < .001$ (partial $\eta^2 = .944$). Bonferroni post hoc tests revealed that classification accuracy associated with discriminating each emotion from neutral ($M_{\text{AF-NU}} = .243$) and 2-way classifications ($M_{\text{2-WAY}} = .238$) were on par and quantitatively higher than classification accuracy associated with 3-way ($M_{\text{3-WAY}} = .177$), 4-way ($M_{\text{4-WAY}} = .143$), and 5-way ($M_{\text{5-WAY}} = .123$) discriminations (see Figure 4B).

Discriminating a larger number of affective states is challenging, particularly when the states are collected in an ecologically valid setting (i.e. no actors were used to express emotions and no emotions were intentionally induced). As expected, there appears to be a linear relationship between the number of emotions simultaneously being discriminated and the associated classification accuracy score (R^2= .91). It appears that each additional affective state included in the classification model is accompanied by a .04 (kappa) reduction in classification accuracy.

4.2.3. Comparison Across Affect Judges. The ANOVA revealed that there were statistically significant differences in kappa scores based on which judge provided the affect ratings used to train and validate the classifiers, $F(4, 64) = 26.42$, $MSe = .001$, $p < .001$ (partial $\eta^2 = .623$). Bonferroni post hoc tests revealed that classifiers based on affect ratings where the trained judges agreed ($M_{J1J2} = .203$, $p < .01$) yielded the best performance as depicted in Figure 4C. We recommend that this finding be interpreted with some caution since this data set probably consists of some of the more obvious cases, namely since the trained judges were able to agree on an affective state.

Figure 4C indicates that overall classification performance between the self, peer, and 2 trained judges were on par with each other ($M_{SELF} = .188$, $M_{PEER} = .183$, $M_{JDG1} = .178$, and $M_{JDG2} = .172$). However, interesting patterns appear when one considers interactions between the affect judge and the emotions classified (see Figure 4D), which was statistically significant, $F(16, 256) =$

167.76, MSe =0, $p < .001$ (partial $\eta 2 = .913$). When one considers simple affect-neutral distinctions, it appears that classifiers trained on data sets in which affect judgments were provided by the novice judges (self and peer, $M_{NOVICES} = .309$) were much higher than classifiers based on affect judgments from the trained judges ($M_{TRAINED} = .199$). However, a reverse pattern was discovered for more complex discriminations between the various emotions (obtained by averaging accuracy scored for 2-way, 3-way, 4-way, and 5-way classifications). These classifiers were best for the trained judges ($M_{TRAINED} = .179$) compared with the novices ($M_{NOVICES} = .153$). This suggests that the novices were more adept at making affect-neutral distinctions whereas the trained judges are more capable at performing complex emotion discriminations. Perhaps this phenomenon may be explained by the fact that the trained judges had considerable experience interacting with AutoTutor, and they make use of this contextual knowledge coupled with their facial expression training to discriminate between the affective states.

4.2.4. Comparisons across Classifier Schemes. We performed an additional repeated measures ANOVA to determine whether there was any significant differences among the various classifier schemes described above. This analysis had two factors: *feature* and *classifier*. Similar to the 3-way ANOVA described above, the first factor (*feature*) was the feature set used as input to the classifier and had 2 levels: *pressure* and *contours*. The second factor of the ANOVA was the classification scheme (called *classifier*) divided across 6 levels for Bayesian

classifiers, functions, instance based learners, meta classifiers, rules, and trees. The unit of analysis for this 2× 6 ANOVA was kappa scores associated with each of the affective models (affect-neutral, 2, 3, 4, and 5 way classifications).

As expected from the previous analyses, there was a statistically significant difference among the two feature sets used with the spatial-temporal contours feature set outperforming the high level pressure feature set. There were also a significant differences in the kappa scores across the various classifier schemes $F(5, 20) = 34.81$, $MSe = .006$, $p < .001$ (partial $\eta^2 = .807$, see Figure 4E). Bonferroni post hoc tests revealed that the kappa scores of the instance based classifiers ($M_{INST} = .22$) were significantly higher than the others. Performance of the function-based classifiers, meta classifiers, and trees ($M_{FNCN} = .193$, $M_{META} = .188$, $M_{TREE} = .155$) were similar quantitatively and higher than Bayesian classifiers and rule-based learning schemes ($M_{BAYS} = .158$, $M_{RULE} = .177$).

It is informative to note that the results showed no statistically significant interactions between the feature set and the classification schemes ($p = .113$). This result indicates that the relative performance of the 6 classification schemes derived above is independent of the feature set (pressure or contours).

INSERT TABLE 2 ABOUT HERE

4.3. Maximum Classification Accuracy

The use of multiple assessments of the learner's affect (N=5) and a large number of classifiers (N=17) was useful to investigate the effect of different

factors (feature set, affect judge, etc) on affect detection accuracy. However, in order to achieve our goal of developing a real time emotion classification system, we will shift our focus to the classifier that yielded the best performance. Table 2 presents the maximum classification accuracies obtained across all 17 classifiers across the 5 data sets in discriminating the various combinations of affective states specified by the hierarchy (See Figure 3).

The results revealed that the accuracy for affect-neutral discrimination and 2-way emotion resolutions are reasonable (74% and 71% respectively), but the accuracy drops when we attempt to resolve conflicts between 3 or more emotions (3-way = 55%, 4-way = 46%, 5-way = 39%). Therefore, it appears that the efficacy of the hierarchical classification scheme is inversely proportional to the probability of requiring the higher order emotion classification models to resolve discrepancies that arise during the affect-neutral discrimination phase. Simply put, the hierarchical method for affect detection would be feasible if we are able to get by with affect-neutral, 2-way, and the occasional 3-way classifications.

There is some evidence that suggests that human disagreements among the affective states usually occur at the affect-neutral stage or the 2-way classification stage. An analysis on the source of classification errors made by the human judges (self, peer, judge1, and judge 2) revealed that 63.5% of the time each emotion was confused with neutral (affect-neutral detection). The 2-way conflicts occurred 30.4% of the time, while 3-way conflicts were much

rarer (5.5%). The 4-way and 5-way discriminations almost never occurred. Taken together, the ideal model for emotion classification would involve (a) the detection of single emotions compared to neutral states, a resonance-based process that fits the Pandemonium model very well and (b) the resolution of pairs of emotions that have some modicum of activation.

4.4. Comparison to Other Sensors

In addition to using posture as a diagnostic channel in inferring the affective states of the learner, the larger project of extending AutoTutor into an affect-sensitive Intelligent Tutoring System also relies on facial expressions and conversational cues (or dialogue features) that are obtained from AutoTutor's log files. We compared the affect detection accuracy scores associated with the two posture feature sets (pressure features and spatial-temporal contours) with previously established reliabilities from dialogue and facial features (D'Mello, Picard, and Graesser 2007). The comparisons were restricted to affect-neutral discriminations since the more complex affective models (2-way, 3-way, etc) have not yet been developed for the other sensors. It should also be noted that the data set used in the current study was also used to train and validate the dialogue and facial feature based classifiers.

An initial analysis was performed by averaging across the accuracies associated with the 5 affect-neutral discriminations. It appears that the classification accuracy attributed to the posture sensor with the high level pressure features was on par with the dialogue features ($M_{\text{PRESSURE}} = M_{\text{DIALOGUE}}$

= 72%). Classification accuracies for the posture sensor with spatial-temporal contour features (M_{CONTOURS} = 74%) rivaled accuracies obtained by monitoring facial features (M_{FACE} = 75%). It should be noted, however, that the facial features used as predictors were manually annotated, as opposed to being automatically computed as in the case of the posture and dialogue features. This is a technological advantage of posture detection over the face as a channel for affect detection.

A finer grained comparison of the accuracies of detecting each affective state from neutral revealed that the posture sensor (with spatial-temporal features) was the most effective for affective states that do not generate overly expressive facial expressions, such as boredom (74%) and flow (83%). On the other hand, the affective states of confusion (76%) and delight (90%), which are accompanied by significant arousal, were best detected by monitoring facial features. The negative affective state of frustration is typically disguised and therefore difficult to detect with the bodily measures of face and posture. Frustration was best detected by examining the dialogue features in the tutoring context (78%). Taken together, detection accuracies were 80% when particular emotions were aligned with the optimal sensor channels.

5. General Discussion

This research was motivated by the belief that the affective states of learners are manifested through their gross body language via configurations of body position and modulations of arousal. We achieved several milestones that

suggest that significant information can be obtained from body position and movement. Although, the challenges of measuring emotions is beset with murky, noisy, and incomplete data, and is compounded with individual differences in experiencing and expressing emotions., we have found that the characteristics of the body posture are quite diagnostic of the affect states of learners. On the basis of two sets of body pressure features alone, our results showed that conventional classifiers are moderately successful in discriminating the affective states of boredom, confusion, delight, flow, and frustration from each other, as well as from the baseline state of neutral.

One may object to our use of the term *moderate* to characterize our classification results. However, it is imperative to note that an upper bound on automated classification accuracy of affect has yet to be established. While human classifications may be considered to be the ultimate upper bound on system performance, human performance is variable and not necessarily the best gold standard. As discussed above, our own results suggest that humans do not achieve a very high degree of concordance in judging emotions. Our findings with respect to low interrater reliability scores associated with emotion recognition independently replicate findings by a number of researchers (Ang et al. 2002; Grimm et. al. 2006; Litman and Forbes-Riley 2004; Shafran, Riley, and Mohri 2003).

Statisticians have sometimes claimed, with hedges and qualifications, that kappa scores ranging from 0.4 – 0.6 are typically considered to be fair, 0.6 –

0.75 are good, and scores greater than 0.75 are excellent (Robson 1993). On the basis of this categorization, the kappa scores obtained by our best classifiers would range from poor to fair. However, such claims of statisticians address the reliability of multiple judges or sensors when the researcher is asserting that the decisions are clear-cut and decidable. The present goal is very different. Instead, our goal is to use the kappa score as an unbiased metric of the reliability of making affect decisions, knowing full well that such judgments are fuzzy, ill-defined, and possibly indeterminate. A kappa score greater than 0.6 is expected when judges code some simple human behaviors, such as facial action units, basic gestures, and other visible behavior. However, in our case the human judges and computer algorithms are inferring a complex mental state. Moreover, it is the relative magnitude of these measures among judges, sensors, and conditions that matter, not the absolute magnitude of the scores. We argue that the lower kappa scores are meaningful and interpretable as dependent measures (as opposed to checks for reliability of coding), especially since it is unlikely that perfect agreement will ever be achieved and there is no objective gold standard.

In this paper we introduced a 2-step affect detection model where affect-neutral classifiers first determined whether the whether the incoming pressure maps resonated with any one or more of the emotions (versus a neutral state). If there is resonance with only one emotion, then that emotion would be declared as being experienced by the learner. In situations where there is resonance with

2 or more emotions additional 2, 3, 4 or 5-way conflict resolution modules are recruited. Comprehensive evaluations of this model were not presented in this paper because the focus was primarily on a exploring the potential of a posture based affect-detection. However, initial analyses with this model revealed that classification accuracy scores were notably lower for 4 and 5 way emotion classifications than for affect-neutral, 2-way and 3-way emotion discriminations.

Although the lower accuracy for the higher level classification models might seem problematic, it is important to note that it is not imperative for AutoTutor to detect and respond to *all* affective experiences of the learner. What is important, however, is that for a given turn if AutoTutor decides to adapt its pedagogical strategy to incorporate the emotions of the learner, then it should be sufficiently confident that the correct emotion has been detected. Taking inappropriate action, like incorrectly acknowledging frustration, can have very negative effects on the learner's perception of AutoTutor's capabilities which are presumably linked to learning gains.

Therefore, one strategy for AutoTutor in situations that require 4 or 5-way classifications might be to simply ignore the affective element and choose its next action on the basis of the learner's cognitive state alone. Perhaps more attractive alternatives exist as well. For example, AutoTutor could bias the confidence of the tutor's actions as a function of the confidence of the emotion estimate. For example if the system lacks confidence in its assessment of

frustration, then an empathetic response may be preferred over AutoTutor's directly acknowledging the frustration and drastically altering its dialogue strategy. We are in the process of experimenting with these strategies to compensate for non-perfect affect recognition rates.

6. Concluding Remarks

This research provides an alternative to the long standing notion that extoll the virtues of the face as the primary modality through which emotion is communicated. We hope to have established the foundation for the use of gross body language as a serious contender to traditional measures for emotion detection such as facial feature tracking and monitoring the paralinguistic features of speech. When directly compared to other sensors our results, suggest that boredom and flow might best be detected from body language while the face plays a significant role in conveying confusion and delight. It is tempting to speculate, from an evolutionary perspective, that learners use their face as a social cue to indicate that they are confused, to potentially recruits resources to alleviate their perplexity. However, it appears that learners do not readily display frustration on the face, perhaps due to the negative connotations associated with this emotion. This finding is consistent with Ekman's (1969) theory of social display rules, in which social pressures may result in the disguising of negative emotions such as frustration. It is the contextual information obtained by mining AutoTutor's log files that best detects frustration.

Although the face might reign supreme in the communication of the basic emotions (i.e. anger, fear, sadness, enjoyment, disgust, and surprise, Ekman and Friesen 1978), our results clearly indicate that the face is not the most significant communicative channel for some of the learning centered affective states such as boredom and flow. Instead, it is the body that best conveys these emotions. Furthermore, the face can be quite deceptive when learners' attempt to disguise negative emotions such as frustration. But bodily motions are ordinarily unconscious, unintentional, and thereby not susceptible to social editing. These factors make the body an ideal channel for non-intrusive affect monitoring.

References

Aist, G., B. Kort, R. Reilly, J. Mostow, and R. Picard. 2002. Adding Human-Provided Emotional Awareness To An Automated Reading Tutor That Listens. *Intelligent Tutoring Systems***:** Pages 992-93.

Ang, J., R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. 2002. Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. *In Proceedings of the International Conference on Spoken Language Processing.* Pages 2037-2039.

Barrett, L. F. 2006. Are Emotions Natural Kinds? *Perspectives on Psychological Science* 1:28-58.

Bernstein, N. 1967. *The co-ordination and regulation of movement.* London: Pergamon Press.

Bianchi-Berthouze, N. and C. L. Lisetti. 2002. Modeling Multimodal Expression of Users Affective Subjective Experience. *User Modeling and User-Adapted Interaction* 12(1): 49-84.

Bianchi-Berthouze, N., Cairns, P., Cox, A., Jennett, C., and Kim, W. 2006. On posture as a modality for expressing and recognizing emotions. In: Emotion and HCI Workshop, BCS HCI, London

Bull, E. P. 1987. *Posture and Gesture*. Pergamon Press.

Campbell, D. T. and D. W. Fiske. 1959. Convergent And Discriminant Validation By The Multitrait-Multimethod Matrix. *Psychological Bulletin* 56: 81-105.

Cohen, J.1920. A Coefficient Of Agreement For Nominal Scales. *Educational And Psychological Measurement* 20: 37-46.

Conati C. 2002. Probabilistic Assessment Of User's Emotions In Educational Games. *Journal Of Applied Artificial Intelligence* 16:555-575.

Coulson, M. 2004. Attributing Emotion To Static Body Postures: Recognition Accuracy, Confusions, And Viewpoint Dependence. *Journal of Nonverbal Behavior* 28:117-139.

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

de Meijer, M. 1989. The Contribution Of General Features Of Body Movement To The Attribution Of Emotions. *Journal Of Nonverbal Behavior* 13:247–268.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society* 39:1-38.

de Rosis, Fiorella. 2002. Toward Merging Cognition and affect in HCI. *Applied Artificial Intelligence* 16:487-494.

De Vicente, A. and H. Pain. 2002. Informing The Detection Of Students' Motivational State: An Empirical Study. *Intelligent Tutoring Systems*, pages 933-943.

D'Mello, S. K., S. D. Craig, B. Gholson, S. Franklin, R. Picard, and A. C. Graesser. 2005. Integrating Affect Sensors In An Intelligent Tutoring System. Affective Interactions: The Computer. *In The Affective Loop Workshop At 2005 International Conference On Intelligent User Interfaces*, New York: AMC Press.

D'Mello, S. K., S. D. Craig, J. Sullins, and A. C. Graesser. 2006. Predicting Affective States Through An Emote-Aloud Procedure From AutoTutor's Mixed-Initiative Dialogue. *International Journal Of Artificial Intelligence In Education* 16:3-28.

D'Mello, S. K., R. Picard, and A. C. Graesser. 2007. Towards an Affect Sensitive AutoTutor. *IEEE Intelligent Systems* 22:53-61.

D'Mello, S. K., S. D. Craig, A. W. Witherspoon, B. T. McDaniel, and A. C. Graesser. 2008. Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User-Adapted Interaction* 18:45-80.

Dweck, C. S. 2002. *Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways).* In J. Aronson (Ed.), Improving academic achievement: Impact of psychological factors on education, pages 61-87. Orlando, FL: Academic Press.

Ekman, P. and W. V. Friesen. 1969. Nonverbal Leakage And Clues To Deception. *Psychiatry* 32:88-105.

Ekman, P. and W. V. Friesen. 1978. *The Facial Action Coding System: A Technique For The Measurement Of Facial Movement*. Palo Alto: Consulting Psychologists Press.

Forbes-Riley, K., and D. Litman. 2004. Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, MA. Pages 201-208.

Graesser, A.C., S. Lu, G. T. Jackson, H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers* 36:180-193.

Graesser, A.C., B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson. 2006. Detection Of Emotions During Learning With AutoTutor. Proceedings of the 28th Annual Conference of the Cognitive Science Society, Mahwah, NJ: Erlbaum.

Graesser, A. C., N. Person, D. Harter, and Tutoring Research Group. 2001. Teaching Tactics And Dialogue In AutoTutor. *International Journal of Artificial Intelligence in Education* 12:257-279.

Grimm, M., E. Mower, K. Kroschel, and S. Narayan. 2006. Combining Categorical and Primitives-Based Emotion Recognition. *In 14th European Signal Processing Conference (EUSIPCO),* Florence, Italy.

Kim, Y. 2005. Empathetic Virtual Peers Enhanced Learner Interest and Self-Efficacy. *In Workshop on Motivation and Affect in Educational Software at the 12th International Conference on Artificial Intelligence in Education. Amsterdam, Netherlands.*

Klein, J., Y. Moon, and R. Picard. 2002. This Computer Responds To User Frustration – Theory, Design, And Results. *Interacting with Computers* 14: 19-140.

Kort, B., R. Reilly, and R. Picard. 2001. An Affective Model Of Interplay Between Emotions And Learning: Reengineering Educational Pedagogy—Building A Learning Companion. *In Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges,* Pages 43-48. Madison, Wisconsin.

Linnenbrink, E. A. and P. Pintrich, 2002. The Role Of Motivational Beliefs In Conceptual Change. In M. Limon and L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice.* Dordretch, Netherlands: Kluwer Academic Publishers.

Liscombe, J., G. Riccardi, and D. Hakkani-Tür. 2005. Using Context to Improve Emotion Detection in Spoken Dialog Systems. *In EUROSPEECH'05, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Lisetti, C. L., and P. Gmytrasiewicz. 2002. Can a Rational Agent Afford to be Affectless? A Formal Approach. *Applied Artificial Intelligence, An International Journal* 16:577-609.

Litman, D. J., and K. Forbes-Riley. 2004. Predicting Student Emotions In Computer-Human Tutoring Dialogues. *In Proceedings Of The 42nd Annual Meeting Of The Association For Computational Linguistics*, East Stroudsburg, PA: Association for Computational Linguistics.

Mandler, G. 1976. *Mind and emotion*. New York: Wiley.

Mandler, G. 1984. *Mind and body: Psychology of emotion and stress*. New York: Norton.

Matsubara, Y. and M. Nagamachi. 1996. Motivation Systems and Motivation Models for Intelligent Tutoring. *In Proceedings of the Third International Conference in Intelligent Tutoring Systems.*

Mehrabian, A.1972. *Nonverbal communication*. Aldine-Atherton, Chicago, Illinois.

Montepare, J., E. Koff, D. Zaitchik, and M. Albert. 1999. The Use Of Body Movements And Gestures As Cues To Emotions In Younger And Older Adults. *Journal of Nonverbal Behavior* 23:133–152.

Morimoto, C., D. Koons, A. Amir,. and M. Flickner, 1998. Pupil Detection and Tracking using Multiple Light Sources. Technical Report, IBM: Almaden Research Center.

Mota, S. and R. W. Picard, 2003. Automated Posture Analysis for Detecting Learner's Interest Level. *In Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI*.

Norman, D. A. 1994. How might people interact with agents? *Communication of the ACM* 37: 68-71.

Paiva, A., R. Prada, and R. W. Picard. 2007. (Eds.). *Affective Computing and Intelligent Interaction*. Springer.

Pantic, M. and L. J. M. Rothkrantz. 2003. Towards an Affect-sensitive Multimodal Human-Computer Interaction. In *Proceedings of the IEEE, Special Issue on Multimodal Human- Computer Interaction* (HCI). 91 (9): 1370-1390.

Picard, R.W.1997. *Affective Computing*. Boston, MA: MIT Press.

Picard, R.W., E. Vyzas, and J. Healey.2001. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions Pattern Analysis and Machine Intelligence* 23: 1175-1191.

Prendinger, H. and M. Ishizuka. 2005. The Empathic Companion: A Character-Based Interface That Addresses Users' Affective States. *International Journal of Applied Artificial Intelligence* 19:267-285.

Rani, P., N. Sarkar, and C. A. Smith, C. A. 2003. An Affect-Sensitive Human-Robot Cooperation – Theory and Experiments. *In Proceedings of the IEEE Conference on Robotics and Automation*, Pages 2382 – 2387. Taipei, Taiwan: IEEE.

Robson C.1993. *Real word research: A resource for social scientist and practitioner researchers*. Oxford: Blackwell.

Selfridge, O. G. 1959. Pandemonium: A Paradigm For Learning. In *Symposium on the Mechanization of Thought Processes,* Pages 511-531. London: Her Majesty's Stationary Office.

Shafran, I., M. Riley, and M. Mohri. 2003. Voice signatures. *In Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, Pages 31-36. Piscataway, NJ: IEEE.

Shneiderman, B., and C. Plaisant. 2005. *Designing the user interface: Strategies for effective human-computer interaction (Ed. 4)*. Reading, MA: Addison-Wesley.

Tan, H. Z. Lu, I. and A. Pentland. 1997. The Chair As A Novel Haptic User Interface. *In Proc. Workshop of Perceptual User Interface*, Pages 56-57.

Tekscan. 1997. *Tekscan Body Pressure Measurement System User's Manual*. South Boston, MA: Tekscan Inc.

VanLehn, K., A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rose. 2007. When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science.*

Walk, R. D., and K. L. Walters. 1988. Perception Of The Smile And Other Emotions Of The Body And Face At Different Distances. *Bulletin of the Psychonomic Society* 26:510–510.

Whang, M. C., J. S. Lim, and W. Boucsein. 2003. Preparing Computers for Affective Communication: A Psychophysiological Concept and Preliminary Results. *Human Factors* 45:623-634.

Witten, I. H. and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques (2nd Ed.).* San Francisco, CA: Morgan Kaufmann.

Acknowledgments

We thank our research colleagues in the Emotive Computing Group and the Tutoring Research Group (TRG) at the University of Memphis (http://emotion.autotutor.org). Special thanks to Patrick Chipman, Scotty Craig, Barry Gholson, Bethany McDaniel, Jeremiah Sullins, Amy Witherspoon for their valuable contributions to this study. We gratefully acknowledge our partners at the Affective Computing Research Group at MIT including Rosalind Picard, Rana el Kaliouby, and Barry Kort.

This research was supported by the National Science Foundation (REC 0106965 and ITR 0325428). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

Table 1. Frequency of affective states in each data set

Affect	**Frequency of Affective States**							**Sum**
Judge	*Boredom*	*Confusion*	*Delight*	*Flow*	*Frustration*	*Neutral*	*Surprise*	
Self	483	533	94	593	335	849	80	2967
Peer	582	555	50	605	207	942	71	3012
Trained Judge 1	379	1151	184	568	291	1192	46	3811
Trained Judge 2	770	1101	120	357	131	1214	30	3723
Trained Judges Agree	268	701	102	224	97	663	17	2072

Table 2. Maximum Classification Accuracy in Detecting Affect

Model	Affective States	Classification Accuracy (%)				Baseline (%)
		Max		Mean		
		P	C	P	C	
Affect-Neutral Detection	BO-NU	70.70	73.05	72.0	74.20	50.00
	CF-NU	68.50	71.65			
	DL-NU	70.55	70.30			
	FL-NU	78.70	82.45			
	FR-NU	71.65	73.70			
2-Way Affect Classification	BO-CF	67.50	70.15	69.40	71.20	50.00
	BO-DL	72.70	74.80			
	BO-FL	76.20	78.25			
	BO-FR	68.70	68.00			
	CF-DL	67.80	64.00			
	CF-FL	68.20	71.65			
	CF-FR	64.05	66.00			
	DL-FL	72.40	78.25			
	DL-FR	66.75	69.85			
	FL-FR	70.10	70.50			
3-Way Affect Classification	BO-CF-DL	51.63	52.37	52.50	55.20	33.33
	BO-CF-FR	50.69	51.49			
	BO-CF-FL	54.84	58.80			
	BO-DL-FR	53.03	55.18			
	BO-DL-FL	56.58	64.09			
	CF-DL-FR	48.14	49.15			
	CF-DL-FL	53.77	54.57			
	CF-FL-FR	50.35	52.50			
	BO-FL-FR	53.44	55.31			
	DL-FL-FR	52.03	58.19			
4-Way Affect Classification	BO-CF-DL-FR	41.35	42.33	42.00	45.90	25.00
	BO-CF-DL-FL	42.10	47.88			
	BO-CF-FL-FR	41.80	44.88			
	BO-DL-FL-FR	45.03	49.45			
	CF-DL-FL-FR	39.85	45.18			
5-Way Affect Classification	BO-CF-DL-FLFR	36.00	39.20	36.00	39.20	20.00

Notes. P – High level pressure features, *C* – Spatial-Temporal Pressure Contours. *BO* – Boredom, *CF*- Confusion, *DL* – Delight, *FL*- Flow, *FR*- Frustration, *NU* - Neutral

Figure 1: Body Pressure Measurement System. (A) The two pressure pads placed on the chair. (B) Pressure maps obtained from the pressure pads on the back and seat. (C) Pressure maps divided into 4 triangular areas for spatial analyses.

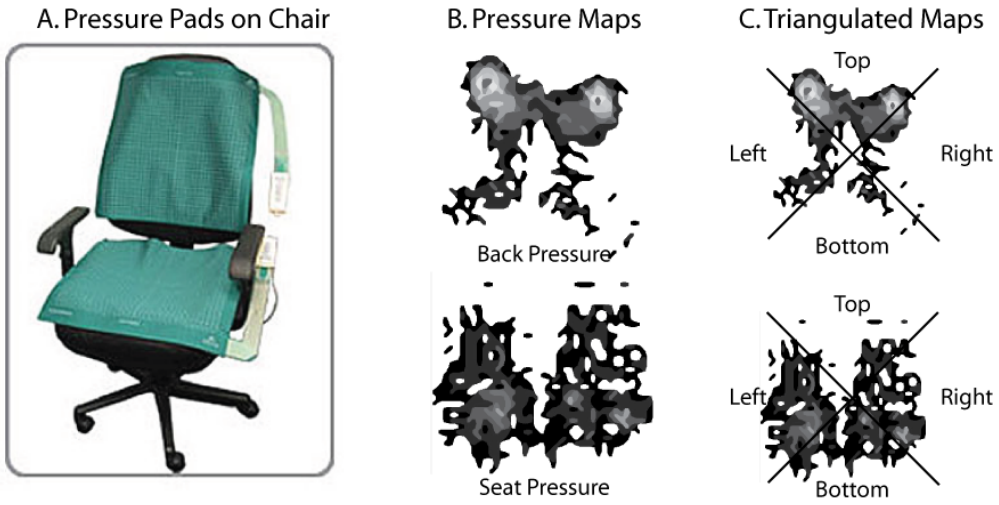

Figure 2. Clustering Pressure Maps on the Back and the Seat with the EM algorithm. The left half of the image is the output from the back while the right half is the output from the seat of the chair. The top quadrants (left and right) show the pressure maps on the back and seat. The 8 plots on the bottom depict each individual clusters - 4 for the back (left) and 4 for the seat (right). Note that the clusters are based on position (X and Y co-ordinates of each sensing element) as well as intensity (pressure exerted on the sensing element).

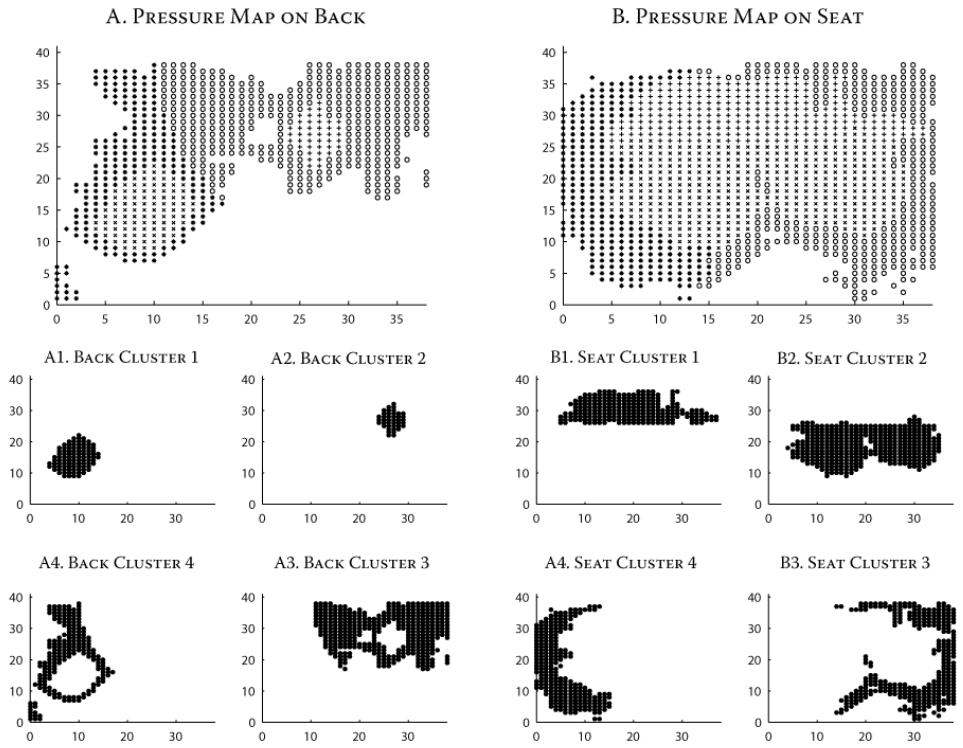

Figure 3. Hierarchical Classification via an Associated Pandemonium

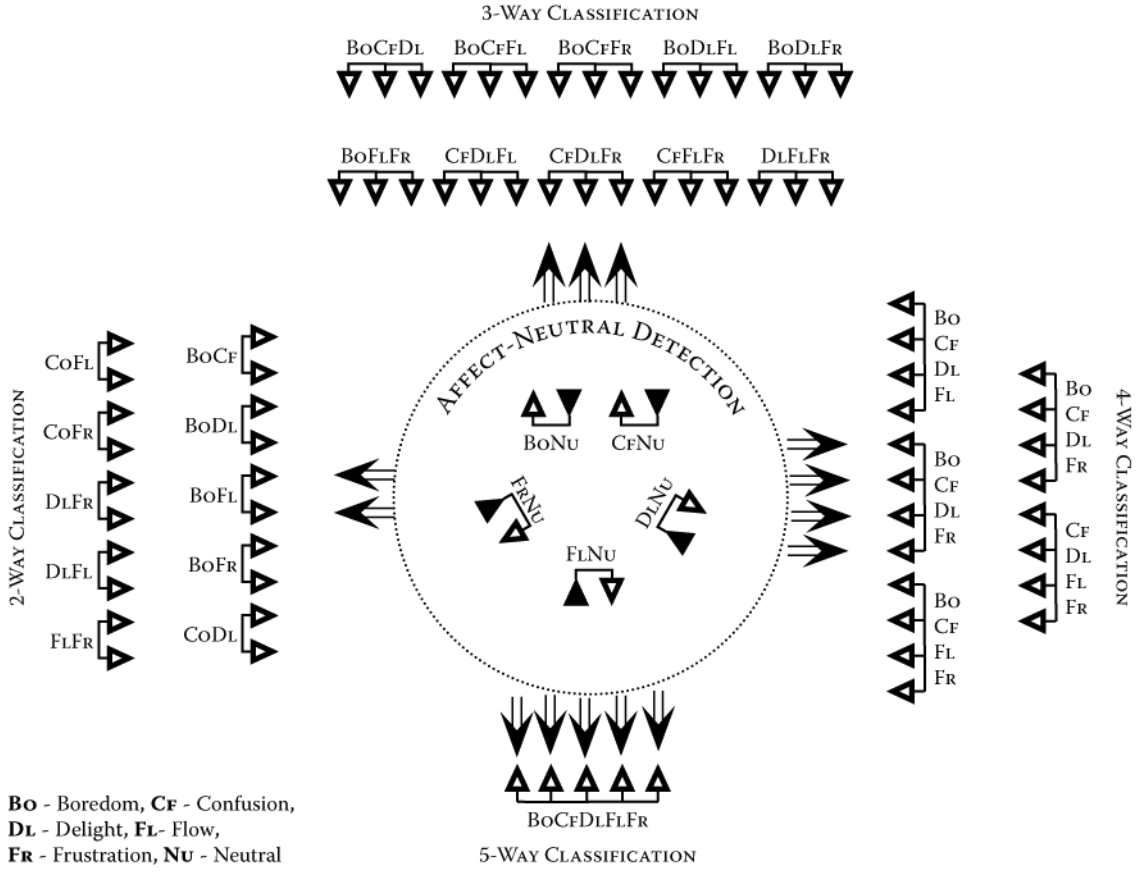

Figure 4. Mean kappa across: (A) Feature Type ; (B) Emotions Classified; (C) Affect Judge (D) Interaction between Emotions Classified and Affect Judge; (E) Classification Scheme