

Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments

Sidney D'MELLO and Arthur GRAESSER

*Institute for Intelligent Systems, The University of Memphis
365 Innovation Drive, University of Memphis, Memphis, TN, 38152, USA*

Abstract. We investigated the potential of automatic detection of a learner's affective states from posture patterns and dialogue features obtained from an interaction with AutoTutor, an intelligent tutoring system with conversational dialogue. Training and validation data were collected from the sensors in a learning session with AutoTutor, after which the affective states of the learner were rated by the learner, a peer, and two trained judges. Machine learning experiments with several standard classifiers indicated that the dialogue and posture features could individually discriminate between the affective states of boredom, confusion, flow (engagement), and frustration. Our results also indicate that a combination of the dialogue and posture features does improve classification accuracy. However, the incremental gains associated with the combination of the two sensors were not sufficient to exhibit superadditivity (i.e., performance superior to an additive combination of individual channels). Instead, the combination of posture and dialogue reflected a modest amount of redundancy among these channels.

Keywords. Affect, emotions, tutorial dialogue, posture patterns, AutoTutor

1. Introduction

Learning inevitably involves emotions. Negative emotions (such as confusion, irritation, frustration, anger, and sometimes rage) are ordinarily associated with failure, making mistakes, diagnosing what went wrong, struggling with troublesome impasses, and starting a complex plan over again. Positive emotions (such as engagement, flow, delight, excitement and eureka) are experienced when tasks are completed, challenges are conquered, insights are unveiled, and major discoveries are made. We are convinced that a broad array of different emotions (or more generally, affect states) accompany the mastery of deep level concepts and complex learning. Consequently, we claim that Intelligent Tutoring Systems (ITS) should include a mechanism for motivating the learner, detecting a learner's emotional/motivational state, and appropriately responding to that state [1, 2, 3].

One of the challenges of such an endeavor involves developing computational systems to reliably detect the learner's emotions. An emotion classifier need not be perfect but should have some modicum of accuracy. Although emotion recognition is an extremely difficult problem, on par with automatic speech recognition, a number of sustained efforts have yielded encouraging results. The majority of these affect detection systems are based on monitoring physiological signals, such as galvanic skin response, heart rate, or other bodily measures that include facial expressions, voice stress analysis, and acoustic-prosodic features [see 4 for a comprehensive review]. As an interesting

alternative, this paper explores the possibility of affect detection from two relatively unexplored sensors: dialogue features and posture patterns. Features from these sensors were extracted from an interaction session with AutoTutor, an ITS that helps learners construct explanations by interacting with them in natural language [5]. The overall goal of the project is to transform AutoTutor into an affect-sensitive intelligent tutoring system.

There are good reasons for expecting that dialogue features are diagnostic of affect in learning environments. The dialogue in one-on-one tutoring sessions yields a rich trace of contextual information, characteristics of the learner, episodes during the coverage of the topic, and social dynamics between the tutor and learner. Within the context of tutoring systems, we would predict that dialogue features provide a very robust diagnostic channel to infer a learner's affect because it has both a broad and deep feature set that covers deep meaning, world knowledge, and pragmatic aspects of communication.

Our use of posture for affect detection is motivated by a number of embodied theories of cognition [e.g. 6, 7]. Theories of embodied cognition postulate that cognitive processes are constrained substantially by the environment and by the coupling of perception and action. If embodied theories are correct, the cognitive and emotional states of a learner are implicitly or intentionally manifested through their gross body language. Therefore, the monitoring of posture patterns may lead to insights about the corresponding cognitive states and affective arousal. An added advantage of monitoring posture patterns is that these motions are ordinarily unconscious and thereby not susceptible to social editing, at least compared with speech acts in conversation.

The fact that two sensors will be monitored raises the issue of how the sensory fusion will impact emotion classification accuracy. One intriguing hypothesis is that classification performance from multiple channels will exhibit super-additivity, that is, classification performance from multiple channels will be superior to an additive combination of individual channels. Simply put, the whole will be greater than the sum of the parts. An alternative hypothesis would be that there is redundancy between the channels. When there is redundancy, the addition of one channel to another channel yields negligible incremental gains; the features of the two channels are manifestations of very similar mechanisms.

Much of the known research in affect detection has involved the use of a single modality to infer affect. Some notable exceptions involve the use of four physiological signals to detect 8 basic emotions [8] and various combinations of audio-visual features [9, 10, 11]. More recently, and of relevance to the context of learning, Kapoor and Picard [12] developed a probabilistic system to infer a child's interest level on the basis of upper and lower facial feature tracking, posture patterns (current posture and level of activity), and some contextual information (difficulty level and state of the game). The combination of these modalities yielded a recognition accuracy of 86%, which was quantitatively greater than that achieved from the facial features (67% upper face, 53% lower face) and contextual information (57%). However, the posture features alone yielded an accuracy of 82% which would indicate that posture was redundant with the other channels.

Before proceeding with a description of the experiments on automated affect detection, it is important to identify some differences between our approach and previously established research in this area. Compared to previous research on affect detection, we monitored a larger number of learning-relevant affective states ($N = 7$) and attempted to automatically distinguish between a subset ($N=4$) of these. We collected affect judgments

from multiple human judges in order to establish ground truth (or a gold standard) regarding the learners' affect. This can be contrasted with a number of researchers who have relied on a single operational measure when inferring a learner's emotion, such as self reports or ratings by independent judges. The present data collection procedure also used ecologically valid methods while monitoring the affect of a learner, namely learners interacting with AutoTutor in a bona fide tutoring session. No actors were used and no attempts were made to intentionally invoke affect. Finally, this research is novel by virtue of tracking relatively unexplored sensors to detect affect, namely dialogue and posture.

2. Empirical Data Collection

Our study collected tutorial dialogues and posture patterns from 28 college students who 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, a video of the participant's face and a video of the screen were recorded. The gross body language of the learner was also recorded using the Body Pressure Measurement System (BPMS, described below). The judging process was initiated by synchronizing the video streams from the screen and the face and displaying them to the judge. Judges were instructed to make judgments on what affective states were present in 20-second intervals, at which time the video automatically paused (freeze-framed). They were also instructed to indicate any affective states that were present in between the 20-second stops.

Four sets of emotion judgments were made for the observed affective states of each participant's AutoTutor session. For the self judgments, the participant watched his or her own session with AutoTutor immediately after having interacted with the tutor. For the peer judgments, the 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), who had been trained on how to detect facial action units according to Paul Ekman's Facial Action Coding System (FACS) [13], judged all of the sessions separately. The trained judges also had considerable interaction experience 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 boredom, confusion, flow, frustration, delight, neutral and surprise. The selection of emotions was based on previous studies of AutoTutor [14, 15] that collected observational data (i.e., trained judges observing learners) and emote aloud protocols while college students learned with AutoTutor.

Interjudge reliability was computed using Cohen's kappa for all possible pairs of judges: self, peer, trained judge1, and trained judge2. Cohen's kappa measures the proportion of agreements between two judges, with correction for baserate levels and random guessing. There were 6 possible pairs altogether. The kappa's were reported in Graesser et al. [16]: self-peer (.08), self-judge1 (.14), self-judge2 (.16), peer-judge1 (.14), peer-judge2 (.18), and judge1-judge2 (.36). These kappa scores revealed that the trained judges had the highest agreement, the self-peer pair had lowest agreement, and the other pairs of judges were in between. It should be noted, however, that the kappa scores increase substantially (.12, .31, .24, .36, .37, and .71) when we focus on observations

in which the learner declares they have an emotion, as opposed to points when they are essentially neutral. The kappa scores are on par with data reported by other researchers who have assessed identification of emotions by humans [e.g. 17].

3. Automated Affect Detection from Dialogue and Posture

One important question that arises during the design of a multisensory emotion classification system involves determining the appropriate level of abstraction at which to fuse the output from the sensors. Fusion at the feature level involves grouping features from the various sensors before attempting to classify emotions. Alternatively, in decision level fusion, the affective states would first be classified from each sensor and then integrated to obtain a global view across the various sensors. While decision level fusion is more common in HCI applications [18], Pantic and Rothkrantz [4] have questioned the validity of using decision level fusion in the affective domain because audio, visual, and tactile signals of a user are typically displayed in conjunction and with some redundancy. Consequently, in this paper we explore the use of feature level fusion alone. We also restrict our scope to a naïve fusion algorithm in which the features from the individual sensors are additively combined before attempting classification.

3.1. Dialogue Features

We mined several features from AutoTutor's log files in order to explore the links between the dialogue features and the affective states of the learners (see [15], for additional details about mining the dialogues). These features included temporal assessments for each student-tutor turn, such as the subtopic number, the turn number, and the student's reaction time (interval between presentation of the question and the submission of the student's answer). Assessments of response verbosity included the number of characters (letters, numbers) and speech act (that is, whether the student's speech act was a contribution towards an answer which was coded as a 1 versus a frozen expression, e.g., "I don't know," "Uh huh," coded as -1). The conceptual quality of the student's response was represented by 4 features obtained with Latent Semantic Analysis (LSA, [19]). LSA is a statistical technique that measures the conceptual similarity of two text sources. AutoTutor's major dialogue moves were ordered onto a scale of conversational directness, ranging from -1 to 1, in terms of the amount of information the tutor explicitly provides the student (e.g. pumps < hints < prompts < assertions < summaries). AutoTutor's short feedback (negative, neutral, and positive) was displayed in its verbal content, intonation, and a host of other non-verbal cues. The feedback was transformed to a 5-point scale ranging from -1 (negative feedback) to 1 (positive feedback).

3.2. Posture Features

The Body Posture Measurement System (BPMS), developed by Tekscan™, was used to monitor the gross body language of a student during a session with AutoTutor. The BPMS consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The output of the BPMS system consisted of two 38x41 matrices (for the back and seat) with each cell in the matrix corresponding to the amount of pressure exerted on

the corresponding element in the sensor grid.

Several features were computed by analyzing the pressure maps of the 28 participants recorded in the study. We computed 5 pressure related features and 2 features related to the pressure coverage for both the back and the seat, yielding 14 features in all. Each of the features was computed by examining the pressure map during an emotional episode (called the current frame). The pressure-related features include the net pressure, which measures the average pressure exerted. The prior change and post change measure the difference between the net pressure in the current frame and the frame three seconds earlier or later, respectively. The reference change measures the difference between the net pressure in the current frame and the frame for the last known affective rating. Finally, the net pressure change measures the mean change in the net pressure across a predefined window, typically 4 seconds, that spans two seconds before and two seconds after an emotion judgment. The two coverage features examined the proportion of non-negative sensing units (net coverage) on each pad along with the mean change of this feature across a 4-second window (net coverage change).

3.3. Classification Experiments

Four data sets, corresponding to each of the four judge's emotion judgments, were obtained by extracting the set of dialogue features for each turn and temporally integrating them with the emotion judgments. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15 second interval) was bound to that dialogue move. Similarly, posture features were extracted from the pressure maps obtained at the time of the emotional experience. This allowed us to obtain four sets of labeled dialogue and posture data aggregated across the 28 participants. The sizes of these data sets were 1022, 1038, 1113, and 1107 for affect labels provided by the self, the peer, trained judge1, and trained judge2, respectively.

The Waikato Environment for Knowledge Analysis (WEKA) was used to evaluate the performance of various standard classification techniques in an attempt to detect affect from dialogue and posture. The classifiers used were a Naïve Bayes classifier, a simple logistic regression, support vector machines, k-nearest neighbor with $k = 1$, C4.5 decision trees, and an additive logistic regression meta classifier. The classification algorithms were compared in their ability to discriminate between boredom, confusion, flow, and frustration. Delight and surprise was excluded due to a comparatively low number of observations. To establish a uniform baseline, we randomly sampled an equal number of observations from each affective state category. This process was repeated for 10 iterations and all reported reliability statistics were averaged across these 10 iterations. Classification reliability was evaluated on the 6 classification algorithms using k-fold cross-validation ($k = 10$).

Figure 1 presents kappa scores (classification accuracy after adjusting for the 25% base rate), averaged across the 6 classifiers. The discrimination was among the affective states of boredom, confusion, flow, and frustration. These results support a number of conclusions regarding automated affect detection. First, both dialogue and posture were moderately successful in discriminating between the affective states. 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. In particular, the interrater reliability (kappa) scores between the various human judges for a subset of the affective states (boredom, confusion, flow, and frustration) was: $\kappa_{\text{self-peer}} = .13$, $\kappa_{\text{self-judge1}} = .30$, $\kappa_{\text{self-judge2}} = .29$, $\kappa_{\text{peer-judge1}} = .33$, $\kappa_{\text{peer-judge2}} = .34$, and $\kappa_{\text{judge1-judge2}} = .58$. The highest kappa score for the machine generated emotion categories was $\kappa_{\text{machine}} = .32$ (or 50% accuracy for a four-way classification). This places the affect detection capabilities of the computer on par with novice judges (self and peer), but significantly lower than trained judges.

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 [20].

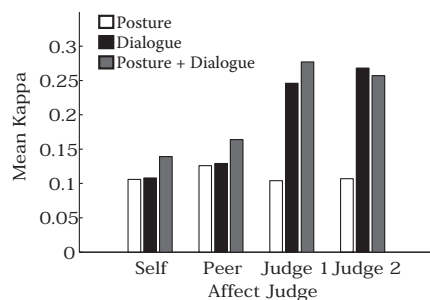


Figure 1. Classification Accuracy

On the basis of this categorization, the kappa scores obtained by our best classifier (.32) 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..

The results unveil interesting patterns between the various sensors and the type of judge that provided the emotion ratings. We find that the use of posture features yielded an average kappa of .112 for the novice datasets (self and peer), which is on par with the kappa obtained on the data sets in which affect judgments were provided by the trained judges (.106). For the dialogue features, we note that kappa scores associated with classifiers trained on the novice data sets (.119) were quantitatively lower than the kappa obtained from the trained judges' data sets (.257). One possible explanation for this difference is that the trained judges had considerable experience interacting with AutoTutor and conceivably tapped into this knowledge while they rated the learner's emotions. This use of contextual knowledge coupled with a video of the participant's face is reflected in the kappa scores obtained by the classifiers.

The results also indicate that a simple combination of dialogue and posture features does yield performance gains. That is, the combination of two sensors is higher than either one alone. However, it is unclear as to whether these performance gains reflect a degree of superadditivity or redundancy. One possible metric to assess superadditivity is on the basis of assessing improvements afforded by multisensory fusion over and above the maximum unisensory response. Specifically, if k_p and k_d are the kappa scores

associated with detecting affect when posture and dialogue are considered individually, and k_{p+d} is the kappa score associated with a combination of these two channels, then the degree of superadditivity obtained would be $[(k_{p+d} - \max(k_p, k_d))/\max(k_p, k_d)]$. This metric is widely used by neuroscientists studying multisensory integration with respect to visual, audio, and tactile senses in humans and animals [21]. The use of this metric to assess superadditivity yielded incremental gains of .3, .3, .1, and 0 from the data sets of the self, peer, and the 2 trained judges.

As an alternative, one could consider incremental gains obtained above and beyond an additive combination of the two sensors. This can be expressed as $k_{add} = k_p + k_d - k_p * k_d$. Superadditivity would be achieved if $k_{p+d} > k_{add}$, additivity if $k_{p+d} = k_{add}$. Redundancy occurs in situations where the combination of two or more sensory channels does not produce incremental gains (i.e. $k_{p+d} < k_{add}$). On the basis of this more conservative metric we conclude that a combination of posture and dialogue features resulted in redundancy.

4. Discussion

Multimodal systems for affect detection in general user modeling has been widely advocated but rarely implemented [22]. This is mainly due to the inherent challenges with unisensory affect detection, which no doubt increase in multisensor environments. The present research is a modest but important step in affect detection research.

This paper investigated an affect classifier to discriminate among the affective states of boredom, confusion, flow, and frustration. Although, this classifier did not consider neutral affect, neutral observations will be important to consider in the next step of the project. The next step will adaptively tailor AutoTutor's dialogue moves to the learner's affective states in addition to their cognitive states. In our view, the comparative classifier developed here would serve as a tie-breaker for several separate individual affect-neutral classifiers. Specifically, we envision a set of affect-neutral classifiers that determine whether the learner is experiencing an emotion (boredom, confusion, flow, or frustration) or is in the neutral state. If there is resonance with one and only one emotion, then that emotion is declared as being experienced by the learner. If there is resonance with 2 or more emotions, then the comparative classifier developed in this study would be used to discriminate among candidate emotions. We are currently in the process of calibrating the reliability of such affect-neutral classifiers. Our most recent results reveal that kappas associated with detecting each emotion from neutral are fair, at least with respect to the criterion established above: boredom = .40, confusion = .52, flow = .48, and frustration = .54. More details of this analysis are reported in D'Mello, Picard, and Graesser [23].

The broader scope of this research ventures into the areas of embodied cognition. The discovery of redundancy between the dialogue and posture features indicate that there are significant relationships between cognition and bodily movements. However, many research questions remain unanswered. To what extent can affect and cognition individually predict bodily activity? Does a combination of these channels increase their predictive power? Answers to these questions will help us explore theories of embodied cognition in addition to the synchronization of emotions with complex learning.

Acknowledgements

This research was supported by the National Science Foundation (REC 0106965 , ITR

0325428, and REC 0633918). 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.

References

- [1] Picard, R. (1997). *Affective Computing*. The MIT Press, Cambridge, MA.
- [2] Issroff, K. & del Soldato, T. (1996). Incorporating motivation into computer-supported collaborative learning. *Proceedings of the European Conference on Artificial Intelligence in Education*.
- [3] Lepper, M. R., & Chabay, R. W. (1988). Socializing the intelligent tutor: Bringing empathy to computer tutors. In H. Mandl & A. Lesgold (Eds.), *Learning Issues for Intelligent Tutoring Systems* (pp. 242-257). Hillsdale, NJ: Erlbaum.
- [4] Pantic, M., & Rothkrantz, L. J. M. (2003). Towards an affect-sensitive multimodal human-computer interaction. *IEEE Special Issue on Multimodal Human-Computer Interaction*, 91(9), 370-1390.
- [5] Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group (2000). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1, 35-51.
- [6] Clark, A. (1997). *Being there: Putting brain body and world together again*. Cambridge, MIT Press.
- [7] Glenberg, A. M., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558-565.
- [8] Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 23(10), 1175-1191
- [9] Chen, L. S., Huang, T. S., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proc. FG* (pp. 396-401).
- [10] Dasarathy, B. V. (1997). Sensor fusion potential exploitation: Innovative architectures and illustrative approaches. *Proc. IEEE*, 85, 24-38.
- [11] Yoshitomi, Y., Kim, S., Kawano, T., & Kitazoe, T. (2000). Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proc. ROMAN* (pp. 178-183).
- [12] Kapoor, A. and Picard, R. E. (2005), Multimodal Affect Recognition in Learning Environments, ACM MM'05, November 6-11, 2005, Singapore.
- [13] Ekman, P. & Friesen, W. V. (1978). *The facial action coding system: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.
- [14] Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241-250.
- [15] D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (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.
- [16] Graesser, A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In R. Son (Ed.), *Proceedings of the 28th Annual Meetings of the Cognitive Science Society* (pp. 285-290). Mahwah, NJ: Erlbaum.
- [17] Marsic, I., Medl, A., and Flanagan, J. L. (2000) Natural Communication with Information Systems, *Proceedings of the IEEE*, 88(8), 1354-1366.
- [18] Pentland, A. (2000). Looking at people: sensing for ubiquitous and sensible computing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22, 107-119.
- [19] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- [20] Robson C. (1993). *Real word research: A resource for social scientist and practitioner researchers*. Oxford: Blackwell.
- [21] Holmes, N., & Spence, C. (2005). Multisensory Integration: Space, Time and Superadditivity. *Current Biology*, 15(18), 762-764.
- [22] Jaimes, A., & Sebe, N. (2005) Multimodal human computer interaction: A survey. *IEEE International Workshop on Human-Computer Interaction (HCI 2005)*, Beijing, China.
- [23] D'Mello, S. K., Picard, R., & Graesser, A. C. (in review). Towards an affect-sensitive AutoTutor.