Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor

M. S. Hussain[1,2], Omar AlZoubi[2], Rafael A. Calvo[2] and Sidney D'Mello[3]

[1] National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
[2] School of Electrical and Information Engineering, University of Sydney, Australia
[3] Institute for Intelligent Systems, University of Memphis, Memphis, USA
Sazzad.Hussain@nicta.com.au
{omar.alzoubi, Rafael.Calvo}@sydney.edu.au
sdmello@memphis.edu

Abstract. It is widely acknowledged that learners experience a variety of emotions while interacting with Intelligent Tutoring Systems (ITS), hence, detecting and responding to emotions might improve learning outcomes. This study uses machine learning techniques to detect learners' affective states from multichannel physiological signals (heart activity, respiration, facial muscle activity, and skin conductivity) during tutorial interactions with AutoTutor, an ITS with conversational dialogues. Learners were asked to self-report (both discrete emotions and degrees of valence/arousal) the affective states they experienced during their sessions with AutoTutor via a retrospective judgment protocol immediately after the tutorial sessions. In addition to mapping the discrete learning-centered emotions (e.g., confusion, frustration, etc) on a dimensional valence/arousal space, we developed and validated an automatic affect classifier using physiological signals. Results indicate that the classifier was moderately successful at detecting naturally occurring emotions during the AutoTutor sessions.

Keywords: Affective computing, emotion, AutoTutor, multichannel physiology, learning interaction, self reports.

1 Introduction

It has been widely acknowledged that cognition, motivation, and emotion are the key components of learning. During tutorial sessions with Intelligent Tutoring Systems (ITS) or human tutors, learners experience a host of learning-centered emotions such as confusion, boredom, engagement/flow, curiosity, interest, surprise, delight, anxiety, and frustration. These affective states are highly relevant and influential to both the processes and products of learning [1]. Therefore, researchers in the interdisciplinary arena encompassing psychology, education, neuroscience, and computer science have recently been focused on understanding the relationship between affect and learning [1-4].

Affect-sensitive ITSs aspire to detect and respond to learner emotions in order to improve learning gains along with increasing motivation and task interest [3]. These systems aim to reduce the gap between human tutors and computer tutors by endowing ITSs with a degree of emotional intelligence. Whether it is human or computer, a learning environment requires some degree of accuracy in classifying the learner's affective states. Detecting affective states with reasonable accuracy is an essential challenge for achieving functional affect-sensitive ITS [5].

There has been some research on learners' affect recognition from facial expression, speech, posture and dialog [4, 6]. A study by Arroyo et al. [7] explored how students' experience with tutoring systems shape their feelings and proposed a data-driven model for emotion using four sensors (camera, mouse, chair, and wrist). Physiological signal analysis is another possible approach to affect detection, and the focus of this paper. Here, heart rate, respiration, muscle activity, galvanic skin response, skin temperature, blood pressure etc might be suitable channels for recognizing affective states provided appropriate pattern recognition techniques are utilized. There is some evidence that some of these physiological signals correlate with the "basic emotions" such as anger, sadness, and disgust [5]. Unfortunately, these basic emotions are not very prominent in learning situations, at least for the short learning sessions with ITSs [8], where the learning-centered emotions listed above play a more prominent role. Challenges emerge during the process of collecting physiological data in learning interactions. Sensors for measuring physiological signals are often unsuitable for learning environments as they tend to interfere with learning activities. Due to these challenges, affect recognition with physiological signals is quite rare in educational settings (exception includes [9]). It is important to note that recent advances in wearable physiological sensors circumvents some of these practical challenges and create new opportunities to infer learner affect from physiology. In this paper we revisit the physiological-based learning-centered affect detection problem by using machine learning techniques to classify affective states from learners' physiological patterns (heart activity, skin response, respiration, facial muscle activity) during learning sessions with AutoTutor, an ITS with conversational dialogues [10].

It is important to emphasize two points before proceeding with a description of our Methods and Results. First, although several theories of emotion focus on *categorical* models, which consider discrete emotions such as fear, anger, etc, the concept, the value, and even the existence of such 'labeled' states is still a matter of considerable debate. Others have proposed *dimensional* models, where a person's affective states are represented as a point in a multi-dimensional space such as a valence-arousal space (see [11] for a discussion). Russell and Barrett [11] proposed a theory that somewhat unites these two views. According to this theory, physiological features are not necessarily correlated with specific emotional states (discrete or categorical emotions), but instead to the underlying dimensions of these states. For example, there is some evidence that valence correlates positively with heart rate while arousal correlates positively with skin conductance level [12]. Perhaps the most defensible position is to adopt a model that incorporates both perspectives by mapping discrete emotions on a valence/arousal space. However, while such a mapping has been proposed for the basic emotions [11], no such empirically grounded mapping exists for the learning-centered emotions. One model has been proposed by Kort, Reilly, and Picard [13], however, this model has yet to be supported with empirical data. Consequently, one of the aims of this study is to provide an empirically grounded mapping of a set of discrete learning-centered affective states into a valence/arousal space. This was achieved by asking learners to provide self-reports of affect based on both categorical and dimensional (valence/arousal) models.

Second, the present focus is on detecting naturally occurring affective states. This is an important point because many physiological-based affect detection systems have relied on artificially-induced emotions using different affect elicitation methods (e.g. photos, films, music, self imagining) [14, 15]. People express their emotions in variable ways, and the same emotion can be expressed differently in different situations. This raises the question of whether physiological-based affect detection will be equally effective in naturalistic contexts. We addressed this question by providing a comparison of the classification performance of affect detection from physiological data for two scenarios: (a) induced emotions via IAPS (International Affective Picture System) [16] and (b) emotions that naturally arise during interactions with AutoTutor.

2 Method

2.1 Participants, Materials and Procedures

Participants were 20 healthy volunteers from the University of Sydney. Participants' age ranged from 18 to 30 years and there were 8 males and 12 females. Participants were instructed not to take any drugs and to avoid caffeine consumption prior to the experiment. Participants signed an informed consent prior to the experiment. The experiment took approximately two hours and participants were rewarded with $20 book vouchers for their participation.

Participants were equipped with physiological sensors that monitored electrocardiogram (ECG), facial electromyogram (EMG), respiration, and galvanic skin response (GSR). The physiological signals were acquired using a BIOPAC MP150 system with AcqKnowledge software at 1000 samples per second for all channels. ECG was collected with two electrodes placed on the wrists. Two channels of EMG were recorded from the zygomatic and corrugator muscles respectively. A respiration band was strapped around the chest and GSR was recorded from the index and middle finger of the left hand.

The experiment consisted of two parts. The first part involved a 40 min recording of physiological signals while participants viewed emotionally charged photos from the IAPS collection [16]. A total number of 90 images (three blocks of 30 images each) for 10 seconds each were presented, followed by 6 seconds pauses between the images. The images were selected so that the IAPS valence and arousal scores for the stimuli spanned a 3×3 valence/arousal space (IAPS normed ratings). Participants also self-reported their emotions by clicking radio buttons on the appropriate location of 3×3 valence/arousal grid after viewing each image [17].

In the second part of the experiment, subjects completed a 20-minute tutorial session with AutoTutor on topics in computer literacy. AutoTutor is a dialogue based ITS for Newtonian physics, computer literacy, and critical thinking. AutoTutor's dialogues are organized around difficult questions and problems (called main questions) that require reasoning and explanations in the answers [10]. During this interaction, a video of the participant's face and a video of the computer screen were recorded. Participants made affect judgments (video annotation) immediately after the learning session at 10 seconds fixed intervals over the course of viewing their face and screen videos [6]. They were asked to provide two types of judgments: (a) categorical judgments which included eight learning-centered affective states (frustration, confusion, flow/engagement, delight, surprise, boredom, curiosity, and neutral) [6, 9] and (b) dimensional judgments consisting of valence/arousal (low, medium, high) ratings using the 3×3 grid described earlier.

2.2 Computational Models for Affect Detection

The Augsburg Matlab toolbox [18] for physiological signal processing was used for extracting statistical features. Video annotations were synchronized with the physiological signals and features were extracted using a 10 seconds window. The feature vectors were also labeled with the corresponding video annotations (1-3 degrees of valence/arousal). A total of 214 features were extracted from the five physiological signals and were merged to achieve feature-level fusion. Some features were common for all signals (e.g. mean, median, and standard deviation, range, ratio, minimum, and maximum) and others were related to their characteristics (e.g. heart rate variability, respiration pulse, frequency). The detailed description of the features can be found in [18]. To reduce the dimensionality of the large number of features, chi-square (X^2) feature selection was used for ranking the ten best features. The X^2 feature selection technique evaluates features by computing the value of the chi-squared statistic with respect to the class, in this case affective states.

The Waikato Environment for Knowledge Analysis (Weka), a data mining package [19], was used for classification. We selected three machine learning algorithms; k-nearest neighbor (KNN), linear support vector machine (SVM), and decision trees for classification Finally, a Vote classifier for combining classifiers was applied with the *average probability* rule [20]. The training and testing for both IAPS dataset and AutoTutor dataset was performed separately with a 10-fold cross validation. The kappa statistic was used as the overall classification performance metric and the F-measure (from precision and recall) was calculated as an indication of how well each affective state was classified. For the classification scores of precision (P) and recall (R), the F-measure (F1) is calculated by; $F1=2((P*R)/(P+R))$.

3 Results and Discussion

3.1 Discrete Emotions Mapping onto the Dimensional Valence/Arousal Plane

The key self-reported states were *neutral* (20%), *boredom* (21%), *confusion* (15%), *flow/engagement* (14%), *curiosity* (10%), and *frustration* (14%), whereas *surprise* (2%), *delight* (4%) were comparatively rare. Mapping of the discrete affective states onto the dimensional (valence/arousal) plane was performed by computing the mean valence and arousal (across 20 participants) associated with each emotion and projecting these on the valence/arousal space. The mapping is presented in Figure 1. It should be noted that a small translation procedure was adopted so that *neutral* was mapped onto the origin.

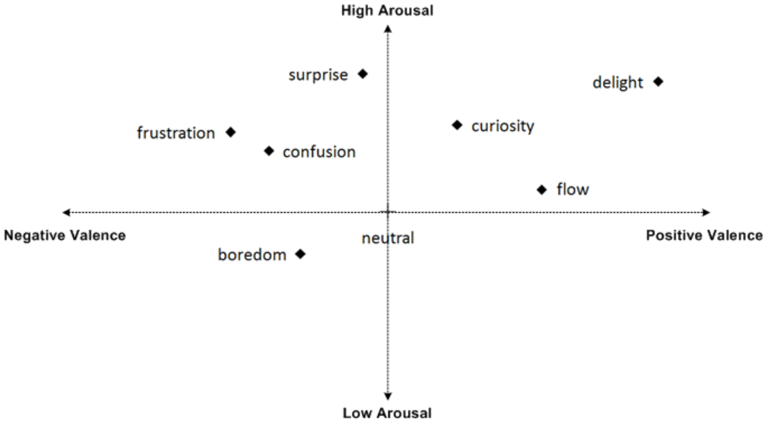

Fig. 1. Mapping of the discrete emotion labels on the valence/arousal plane (horizontal & vertical axes representing dimensions for valence and arousal respectively).

As Figure 1 indicates, *surprise* has no notable valence but has the highest arousal. In contrast, *flow/engagement* has arousal levels similar to neutral but is positively valenced. *Delight* and *curiosity* are characterized by high arousal and valence (especially delight). Both *confusion* and *frustration* have high arousal and negative valence. As could be expected, *boredom* is also negative valence with lower arousal. Most previous studies [e.g. 1, 6, 10] only used discrete affective states to annotate ITS interactions. Our mapping of discrete affective states onto a dimensional model (based on the empirical data) is a novel approach to combining results for the two models.

3.2 Classification Results from Physiological Signals

In this section we present the classification results for detecting 1-3 degrees (low, medium, high) of valence and arousal from physiological features, and leave classification of discrete emotions as part of future work. Self reports normally produce highly skewed class distribution, therefore up sampling and down sampling techniques are commonly used. For the initial analysis presented in this paper, we selected datasets/subjects with approximately balanced distribution of classes without using any up/down sampling techniques. Finally, the classes with extremely low or high number of instances were removed at the subject level. Separate classification analyses were performed for the valence and arousal dimensions. Table 1 presents the mean and standard deviation of kappa scores across learners for detecting 1-3 degrees of valance and arousal from physiological features (for both IAPS and AutoTutor sessions).

Table 1. Mean (M) and standard deviation (SD) of kappa scores for detecting 1-3 degrees of valance and arousal from physioligical signals across leaners

Affect	**IAPS**		**AutoTutor**	
	M	*SD*	*M*	*SD*
Valence	0.49	0.27	0.35	0.22
Arousal	0.31	0.16	0.23	0.03

We note that the overall performance (kappa scores) of affect detection using IAPS is higher than performance during the AutoTutor interaction. This is expected because the IAPS is designed to elicit basic emotions of higher intensity than the learning emotions obtained over the course of the AutoTutor sessions. Despite the lower overall performance, however, kappa scores are clearly greater than chance (kappa = 0) for the naturalistic emotions. In a previous study by D'Mello & Graesser [6] kappa score of 0.29 was achieved from face, dialog and posture using a similar AutoTutor setup. This indicates that learners' valence and arousal can be detected from physiological signals and the performance is quite satisfactory even when compared to controlled emotion elicitation.

While the kappa score provides a measurement for the overall performance, the F-measure indicates how well the individual affective categories were classified. Figure 2 presents the mean and standard deviation of the F-measure for detecting 1-3 degrees of valance and arousal from physiological signals across learners for both IAPS and AutoTutor sessions.

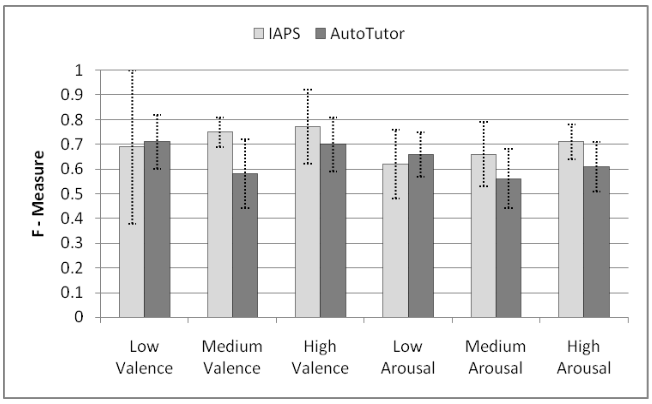

Fig 2. Mean and standard deviation of the F-measure for detecting 1-3 degrees of valance and arousal from physiological signals across learners for both IAPS and AutoTutor sessions

Observing the results from Figure 2 separately for IAPS and AutoTutor; the performance (F-measure) of detecting the degrees of valance and arousal for IAPS increases from low to high for both valance and arousal. On the contrary, during AutoTutor sessions, a curvilinear relationship was observed. Highest performances occur for low valence and low arousal and also for high valence and high arousal. Performance for medium valence and medium arousal is in between these two extremes. While comparing results for IAPS and AutoTutor, we note that the performance of detecting low valence and low arousal from physiology during naturalistic interactions is comparable to controlled emotion elicitation. The performance of detecting medium and high valence/arousal is also quite satisfactory. A paired t-test for comparing the F-measure means for the six categories of IAPS (*M*= .70) and AutoTutor (*M= .64*) revealed no significant difference (*p >0.05*), which indicates that the accuracy of detecting affective states were not very different for the two models. As part of future work, this could be very suitable for creating a model where the classifier can be trained using the IAPS dataset and tested for the AutoTutor interactions.

4 Conclusion

The implementation of an adaptive, multimodal, robust affective sensitive ITS with sufficient reliability is still far from reality. Despite the challenges of affect recognition from physiological signals, this research presents an automatic affect classifier to detect learners' affective states from multichannel physiological signals with the support of a systematic experimental setup, feature selection techniques, and machine learning approaches. Results show that for the AutoTutor interaction, valence and arousal can be classified with moderate accuracy from multichannel physiology. Other modalities such as facial expressions, dialog and posture features [6] can be included along with physiological channels which may improve the performance of affect detection during ITS interactions. Classification of descrete affective states and finding their relationships with the dimensional model using multichannel physiology will be explored in the future.

Acknowledgments. M. S. Hussain was supported by Endeavour Awards and National ICT Australia (NICTA). Sidney D'Mello was supported by the National Science Foundation (ITR 0325428, HCC 0834847). 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. Craig, S., Graesser, A., Sullins, J., Gholson, B.: Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Learning, Media and Technology 29, 241-250 (2004)
2. Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction 19, 267-303 (2009)
3. Calvo, R.A., D'Mello, S.: (in preparation) New perspectives on affect and learning technologies. New York: Springer
4. Kapoor, A., Picard, R. W.: Multimodal affect recognition in learning environments. Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore 677–682 (2005)

5. Calvo, R.A., D'Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. IEEE Transactions on Affective Computing 1, 18-37 (2010)
6. D'Mello, S., Graesser, A.: Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction 20, 147-187 (2010)
7. Arroyo, I., Cooper, D., Burleson, W., Woolf, B.P., Muldner, K., Christopherson, R.: Emotion Sensors Go To School. Proceeding of the 2009 conference on Artificial Intelligence in Education, Vol. 200, Amsterdam 17-24 (2009)
8. Lehman, B., Matthews, M., D'Mello, S., Person, N.: What are you feeling? Investigating student affective states during expert human tutoring sessions. Intelligent Tutoring Systems, Berlin 50-59 (2008)
9. Aghaei Pour, P., Hussain, M., AlZoubi, O., D'Mello, S., Calvo, R.: The Impact of System Feedback on Learners' Affective and Physiological States. Intelligent Tutoring Systems Springer LNCS 6094, Pittsburgh, USA. 264-273 (2010)
10. Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A.: AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education 48, 612-618 (2005)
11. Russell, J.A., Barrett, L.F.: Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76, 805-819 (1999)
12. Lichtenstein, A., Oehme, A., Kupschick, S., Jürgensohn, T.: Comparing Two Emotion Models for Deriving Affective States from Physiological Data. Affect and Emotion in Human-Computer Interaction. LNCS 4868, (2008)
13. Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. IEEE International Conference on Advanced Learning Technologies, Madison, Wisconsin 43-46 (2001)
14. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE transactions on pattern analysis and machine intelligence 23, 1175-1191 (2001)
15. Wagner, J., Kim, J., Andre, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, Amsterdam, The Netherlands 940-943 (2005)
16. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Technical manual and affective ratings. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida (1995)
17. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
18. Wagner, J., Kim, J., Andre, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. IEEE International Conference on Multimedia and Expo, 2005, Amsterdam, The Netherlands 940-943 (2005)
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann (2005)
20. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. Wiley-Interscience (2004)