Predicting Affect from Gaze Data During Interaction with an Intelligent Tutoring System

Natasha Jaques[1], Cristina Conati[1], Jason Harley[2], Roger Azevedo[3]

[1] University of British Columbia, 2366 Main Mall,
Vancouver, BC, Canada V6T1Z4
{jaquesn, conati}@cs.ubc.ca
[2] McGill University, 845 Sherbrook Street West,
Montreal, QC, Canada H3A0G4
Jason.Harley@mail.mcgill.ca
[2] North Carolina State University, 2310 Stinson Drive,
Raleigh, NC, USA 27695
razeved@ncsu.edu

Abstract. In this paper we investigate the usefulness of eye tracking data for predicting emotions relevant to learning, specifically boredom and curiosity. The data was collected during a study with MetaTutor, an intelligent tutoring system (ITS) designed to promote the use of self-regulated learning strategies. We used a variety of machine learning and feature selection techniques to predict students' self-reported emotions from gaze data features. We examined the optimal amount of interaction time needed to make predictions, as well as which features are most predictive of each emotion. The findings provide insight into how to detect when students disengage from MetaTutor.

1 Introduction

Emotions play a critical role in human behavior, thought, motivation, and social interaction [21]. An affect-adaptive interface can react and adapt to clues about the user's emotional state; such systems can increase task success [30], motivation [19], and user satisfaction [18]. Affect sensitivity can be especially beneficial in educational contexts, where maintaining positive emotions can lead to increased learning [21].

Our study focuses on predicting feelings of boredom and curiosity experienced during learner interactions with MetaTutor, an ITS designed to support effective self-regulated learning (SRL) [4]. The main contribution of our work is that we explore the usefulness of eye tracking data alone in predicting learner affect in MetaTutor via machine learning. The only other research that has used eye-tracking data to predict emotions has been limited to using hand engineered heuristics to generate gaze-based interventions [29], or has focused on non-gaze features such as pupil dilation [20] [29]. Unlike pupil dilation, gaze features provide insight into the user's attention to various interface elements, and are not sensitive to changes in luminosity. A second contribution is that we investigate curiosity, an emotion not frequently studied in the affective computing literature. Curiosity is considered an emotion related to interest

[27], and was included based on Pekrun's research into academic emotions [22].We are aware of few other studies that include curiosity [9][11][24]. Finally, by uncovering which features are most predictive of each emotion, we gain insights into effective methods for constructing an affect-adaptive MetaTutor.

2 Related Work

Emotions experienced in an academic setting are related to students' motivation and academic achievement [21]. For example, boredom is linked to decreased task success, while engagement is associated with user satisfaction [13]. Further, the presence of an empathetic and supportive tutor or pedagogical agent has been shown to enhance learning [32], and reduce stress [23]. For these reasons, researchers have begun investigating how to detect and respond to learners' emotional states. Conati and Maclaren [7] used information about learners' personalities and interaction logs to model emotions using a Dynamic Bayesian Network (DBN). Forbes-Riley et al. predicted disengagement from acoustic and dialog features [13].

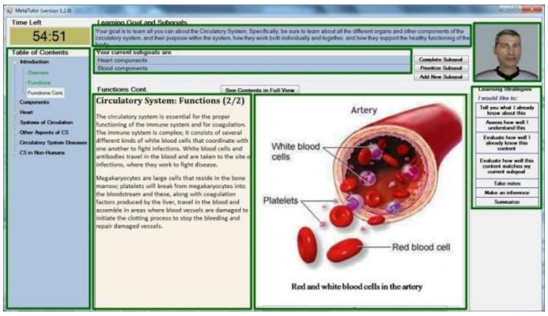
Physiological sensors, including wireless skin conductance bracelets, pressure sensitive seat cushions, and accelerometers, have been used to predict affect in an ITS context [3]. By combining several data sources, including heart rate, skin conductance, posture, questionnaires and interaction logs, Sabourin and colleagues achieved prediction accuracies of 75% for boredom and 85% for curiosity [24]. Affect can also be detected with a single sensor; D'Mello and colleagues. obtained 60%, 64%, and 70% accuracy in predicting boredom using facial expressions, dialog, and posture, respectively [9].

Eye gaze has been used to detect affect. Findings from psychological research have suggested that blinking often or a lack of fixations on interface text may help predict boredom [28], and that increased pupil diameter may be indicative of stronger emotion [20], [29]. This finding was incorporated in an affect-sensitive ITS that responded in real time to heuristic signs of boredom, such as decreased pupil size or wandering gaze [29]. Gaze Tutor [10] also uses heuristics to respond to gaze, by sending an intervention message if a student does not look at the tutor or the pedagogical content for ten seconds. In the broader domain of education, eye gaze has been used to predict learning gains [6] [17], problem solving [2], and reading performance [26].

Most closely related to our study is the work by Harley, Bouchet and Azevedo [15] on correlating the emotions experienced during interactions with MetaTutor with output from FaceReader 5.0 software. Because the FaceReader emotions do not map directly to the academic emotions of the study, the authors had to develop their own mapping scheme, but still achieved 75.6% agreement. This suggests that the emotion self-reports collected during the MetaTutor study closely matched participants' actual behavior [15]. Unfortunately, positive emotions (including curiosity) declined over the course of the interaction, demonstrating a need for affective interventions [15].

3 MetaTutor User Study

MetaTutor is an adaptive ITS designed to encourage students to employ metacognitive SRL strategies, while teaching concepts about the human circulatory system [5]. SRL is the ability to manage learning through monitoring and strategy use, and can be a powerful predictor of students' learning gains and academic success [25]. For this reason, the MetaTutor learning environment (Fig. 1) contains an overall learning goal (OLG) and subgoal completion bar (at the top of the screen), for setting and viewing progress toward learning objectives. There are four pedagogical agents (PAs) which appear in turn in the top right corner of the screen. The learning strategies palette (LSP) is located beneath the PAs, and allows the user to initiate interactions such as requesting an evaluation of her current understanding of content [19]. Finally, MetaTutor's text and image contents are displayed in the center of the screen, and are organized via the table of contents (TOC) on the far left.



Your goal is to learn all you can about the Circulatory System. Specifically, be sure to learn about all the different organs and other components of the circulatory system, and their purpose within the system, how they work both individually and together, and how they support the healthy functioning of the

Circulatory System: Functions (2/2)

The circulatory system is essential for the proper functioning of the immune system and for coagulation. The immune system is complex; it consists of several different kinds of white blood cells that coordinate with one another to fight infections. White blood cells and antibodies travel in the blood and are taken to the site of infections, where they work to fight disease.

Megakaryocytes are large cells that reside in the bone marrow; platelets will break from megakaryocytes into the bloodstream and these, along with coagulation factors produced by the liver, travel in the blood and assemble in areas where blood vessels are damaged to initiate the clotting process to stop the bleeding and repair damaged vessels.

Red and white blood cells in the artery

Fig. 1. The MetaTutor Interface

The data used in this analysis was collected from a study of 67 undergraduate students with a variety of academic program concentrations which were not necessarily related to MetaTutor's content. Participants used Meta Tutor for approximately 90 minutes while they were recorded using a number of sensors, including a Tobii T60 eye tracker [15]. Participants also self-reported their concurrent emotions using an Emotions-Value questionnaire (EVQ) developed by researchers at McGill University. The EVQ consists of 19 basic and learning-centered emotion items, and is based on a modified subscale of Pekrun's Academic Emotions Questionnaire [21]. Each item consists of a statement about an emotion (e.g., "Right now I feel bored"), and was rated on a 5-point Likert scale where 1 indicated "strongly disagree" and 5 indicated "strongly agree". The EVQ was filled out at the beginning, and every 14 minutes thereafter during the one hour learning session with Meta Tutor, for a total of 5 self-reports per student. For the purposes of this study, we will focus on two of the most strongly reported emotions (those most frequently rated as 4 or 5 on the Likert scale): boredom (*M* = 2.60, *SD* = 0.69) and curiosity (*M* = 2.93, *SD* = 0.71) [15].

4 Eye Tracking Data Analysis

The gaze data in this study was collected using a Tobii T60 eye tracker, and takes the form of *fixations* on a single point, and *saccades*, which are the paths between two consecutive fixations. Following the data validation process described in [6], we discarded participants with too few valid gaze samples overall, and were left with a total of 51 participants for analysis. The data was then processed into aggregate features using EMDAT, an open source package for gaze data analysis[1]. The extracted features include application-independent gaze features related to the number of fixations, fixation duration, and saccade length, as well as the angle between two consecutive saccades (the relative path angle) and the angle between a saccade and the horizontal plane (absolute path angle) [6]. We did not include features related to pupil dilation because the data was collected in a room with a window.[2]

In addition to application-independent features, we include features related to specific Areas of Interest (AOIs) within the MetaTutor interface. Following [6], we defined seven AOIs (which are outlined in green boxes in Fig. 1): Text Content, Image Content, Overall Learning Goal (OLG), Subgoals, Learning Strategies Palette (LSP), Agent, and Table of Contents (TOC). We include features such as the duration of the longest fixation on a given AOI, the proportion of fixations and time spent on an AOI, and the number and proportion of gaze transitions between each pair of AOIs. We also include time to first fixation on the AOI, time to last fixation, and the total fixation time. In total, we have 166 features.

5 Machine Learning Experiments

We treat predicting boredom and curiosity as two separate binary classification problems. Although boredom and curiosity could be considered mutually exclusive states, the data does not support this approach. While there was a significant negative correlation between the ratings of boredom and curiosity ($r = -.333$, $p < .001$), in 18% of the self-reports both curiosity and boredom were rated as present simultaneously, and in 13% they were both absent.

Classification labels were based on the EV self-reports. We did not include the first round of reports, because they were collected before participants began using the learning environment. Ratings of 3 or higher were labeled as Emotion Present (EP), and ratings of less than 3 were labeled as Emotion Absent (EA), as in [15]. For classification, we used 10-fold cross validation (CV), and four algorithms available in the Weka data mining toolkit: Random Forests (RF), Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM), chosen because they showed the most promising performance in initial tests. We also use 10-fold CV to tune the parameters of the algorithms. Results are reported in terms of both accuracy (percentage of correctly classified data points), as well as Cohen's kappa, a measure of classification perfor-

[1] http://www.cs.ubc.ca/~skardan/EMDAT/index.html

[2] Pupil dilation is more sensitive to luminance than to affect [31]

mance that accounts for correct predictions occurring by chance [8]. Kappa scores are 1 when classification labels exactly match the ground truth values, and 0 if the predictions were no more accurate than chance. A good kappa score for trained human judges rating emotion might be .5 [13] or .6 [8], while a typical score for a machine predicting emotion might be .3 [8] [16] [1].

Due to the small size of our dataset and the large number of features available, our classifiers will tend to over-fit the training data without an effective feature reduction method. We tested two techniques, Principal Component Analysis (PCA) and Wrapper Feature Selection (WFS), using 10-fold CV, and performing feature selection using only the training data. PCA reduces the dimension of a feature set by creating components based on highly correlated subsets of features [12]. WFS finds useful subsets of features by testing them with a specific classifier [16]. In order to obtain more robust feature sets with WFS, we performed nested cross validation, by further subdividing each training fold into another 10 train/test sets, performing wrapper selection on each, and using those features that were selected in more than 10% of the sub-folds. We found that WFS achieved better results overall, and that the features selected are more interpretable than PCA components. For these reasons, we focus on WFS when reporting results in the rest of the paper.

6 Results

In this section we present the results of several classification experiments. We begin by training classifiers using all available self-reports and gaze features computed using various time intervals preceding each report. We discuss the features chosen as most predictive by WFS, and the effectiveness of predicting reports independently.

6.1 Predicting Self-Reports Across the Interaction

Our first experiment involved training classifiers to predict the affective labels derived from any self-report, regardless of when it was generated. We wished to determine the amount of gaze data preceding the self-report that should be used for prediction. Many studies make use of a window of 20 seconds for affect labeling [14]. In a study of the same dataset, Harley et al., [15] used a 10 second window. We tested window lengths ranging from 100% of the available data (14 minutes) to 1% (8 seconds), and the results are shown in Fig. 2.

We used a 4 (classifier) x 6 (window length) General Linear Model (GLM) to analyze the results, treating the score obtained for one train/test split as a single data point. We ran four of these models, one with each of boredom accuracy, boredom kappa, curiosity accuracy, and curiosity kappa as the dependent variables, and applied Bonferroni corrections to adjust for family-wise error. In cases where the accuracy and kappa results are analogous, we present only the accuracy results. We compare the results to a majority-class baseline using t-tests with a Bonferroni adjustment.

The GLM results for both emotions were similar; there were no significant effects of classifier or interaction effects, likely because we have already restricted our focus to the best classifiers. There was, however, a significant main effect of window length for both boredom, *F*(5,216) = 8.390, η^2 = .163, *p* < .001, and curiosity, *F*(5,216) = 7.382, η^2 = .146, *p* < .001. The curiosity results significantly exceeded the baseline at a window of 100% or 14 minutes (*M* = 63.45%, *SD* = 1.03), *t*(3) = 7.12, *p* < .05. For boredom it was a window of 100%, (*M* = 55.79%, SD = 1.95), *t*(3) = 3.92, p < .05, and a window of 75% or 10.5 minutes (*M* = 57.83%, *SD* = 1.35), *t*(3) = 8.29, *p* < .01. Although certain classifiers (like RF) can still achieve good performance with a small interval of data, in general it seems that more gaze data generates better results. A large body of previous affect prediction research has focused on using a 20 second interval for affect labeling [14]. This study provides empirical evidence that this type of interval may not always be appropriate, depending on the data used for prediction.

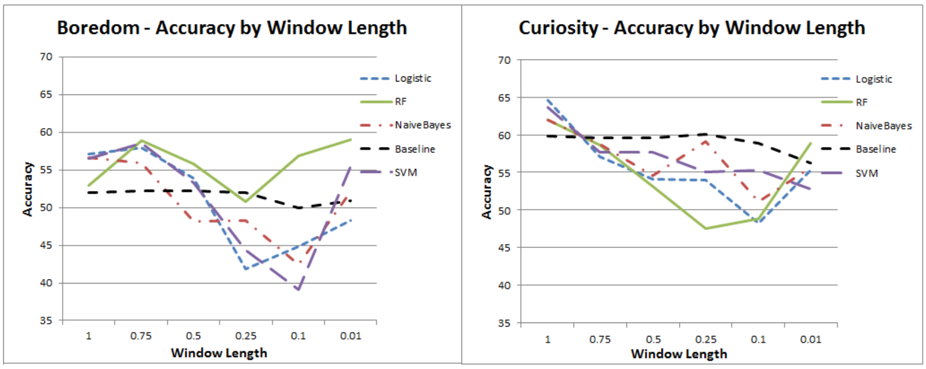


Fig. 2. Accuracy as a function of the fraction of interaction time used for training classifiers

6.2 Important Features

In this section we examine the features that were selected by the WFS process. We focus on the features selected most frequently for the windows that achieved the best performance, reasoning that these features must have been the most informative.

The general trends that seem to have emerged are depicted in Fig. 3, where arrows indicate gaze transitions and circles indicate features related to the AOI itself (circle size increases with the number of these features found). One trend is that students who are engaged (curious and/or not bored) make frequent use of the table of contents (TOC). This is evidenced by the fact that increased fixation length in the TOC and more TOC-to-TOC transfers are predictive of curiosity, while lower TOC fixation rate and fewer OLG-to-TOC and TOC-to-LSP transfers indicate boredom.

Engagement also appears to be linked to use of the image and Overall Learning Goal (OLG) AOIs. For example, bored students spend a smaller proportion of time and number of fixations on the image, and have fewer image-to-image and text-to-image transfers. They also have a shorter maximum fixation on the OLG, and fewer image-to-OLG, text-to-OLG, and OLG-to-subgoals transfers.

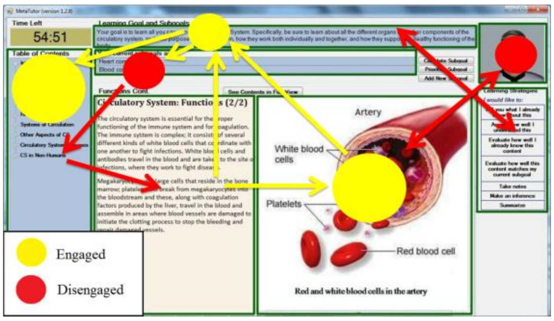

54:51
Time Left
Table of Contents
Learning Goal and Subgoals
Circulatory System: Functions (2/2)
Artery
White blood cells
Platelets
Red blood cell
Red and white blood cells in the artery
Engaged
Disengaged

Fig. 3. Depiction of gaze trends for engaged and disengaged students.

While the features listed so far provide evidence that engaged students look between the text, image, OLG, and TOC in a way that may suggest strategic learning, disengaged students seem to have frequent, scattered gaze transitions, without remaining focused on a single AOI. This is evidenced by subgoals-to-TOC, TOC-to-text, OLG-to-LSP and LSP-to-OLG transitions all being predictive of boredom or a lack of curiosity, while with the exception of frequent fixations on the subgoals, no features related to prolonged attention to the remaining AOIs were found.

Finally, attention to the agent may be associated with disengagement. Curious students fixate for a shorter time on the Agent, and have fewer Agent-to-image and image-to-Agent transfers. This is especially interesting given findings from [6] on the same dataset, which showed that the Agent was the only AOI not predictive of learning gains.

6.3 Time-Dependent Effects on Prediction

In our initial tests, the data from all self-reports was collected together and used in the training set with no indication of the point during the interaction when the report occurred. In the following tests we treat each self-report time as its own classification problem, to see if this timing information can improve performance. We use a full 14-minute window for prediction, based on findings from the previous section.

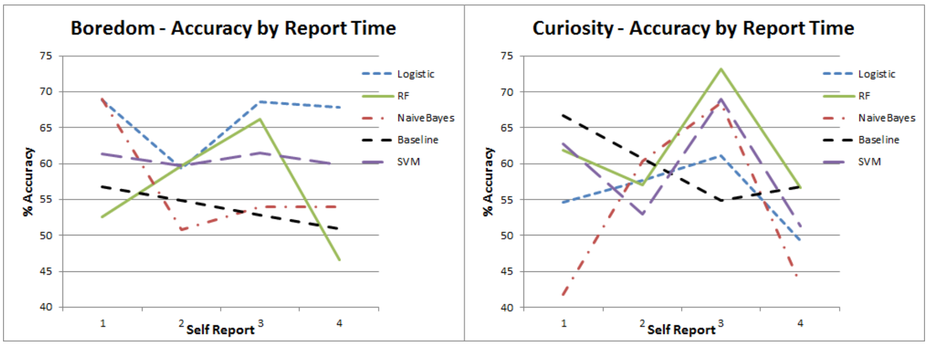


Fig. 4. Accuracy as a function of the self-report time

We conducted a similar 4 (classifier) x 4 (report time) General Linear Model on the results, which are shown in Fig. 4. For boredom, there were no significant effects, although on average the classifiers significantly exceeded the baseline, $t(163) = 2.68$, $p < .01$, with Logistic Regression achieving the highest average of 66.17% (kappa = .306), and peaking at report 1 (68.83%, kappa = .330). For curiosity, we found a main effect of report time for both kappa and accuracy, $F(3,144) = 5.953$, $\eta^2 = .110$, $p < .005$. This suggests that the time of the self-report, which corresponds to the amount of time a student has been interacting with MetaTutor, strongly affects the relationship between gaze and affect. Tukey post-hoc analysis revealed that self-report 3 ($M = 67.96$, $SD = 15.55$) was significantly better than all other reports. It was also the only report in which the classifiers significantly surpassed the baseline, $t(39) = 5.313$, $p < .001$, with Random Forests achieving a peak accuracy of 73.17% (kappa = .416).

Note that in addition to the effect of report time detected for curiosity, the average results obtained for boredom by restricting focus to a single self-report were also markedly higher than those obtained when all report times are classified together, as in the previous section. Overall, the results of this section seem to indicate that the relationship between gaze and affect varies over time. If this were true, we would expect that different gaze features would be more informative at different report times. Indeed, we examined the features chosen by WFS for each report, and found that there was considerable variability. Fig. 5 groups features into categories based on their AOI, and shows how the relevant features change along with time spent with MetaTutor. For example, the subgoals become highly relevant for predicting curiosity at report three, but otherwise are hardly chosen at all. We are not certain of the cause of this effect, however the changing importance of the features demonstrates that different patterns of behavior are indicative of the emotions over time.

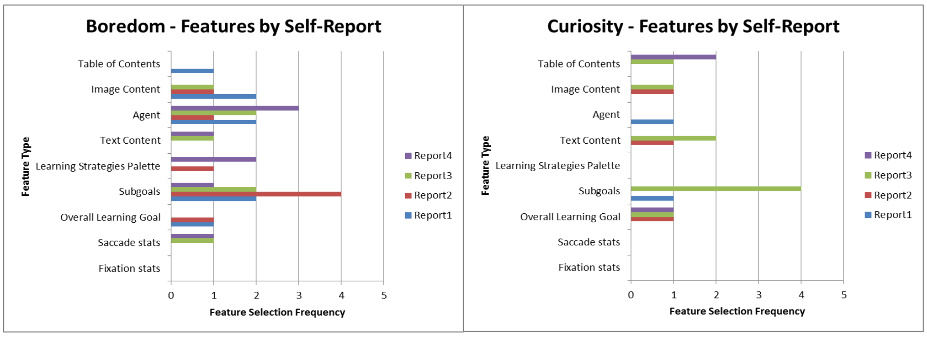


Fig. 5. The features found to be most predictive by wrapper feature selection change, depending on progress through MetaTutor

7 Conclusions and Future Work

The findings from this study demonstrate that eye gaze data alone is a useful tool for predicting boredom and curiosity in MetaTutor. The best results obtained, 69% (kappa = .33) for boredom and 73% (kappa = .42) for curiosity, are notable in the field of affect prediction, where near-perfect results are not the reality [13], and achieving higher accuracies often requires combining multiple sources of user information [24]. We also present empirical evidence to contradict the assumption that a short interval of a few seconds is always most appropriate when predicting affect. Finally, we have found that temporal information about a students' progress through MetaTutor can lead to increased accuracy, so the relationship between gaze and affect in MetaTutor may be dependent on timing.

In the future we plan to leverage additional data sources collected during the MetaTutor study in order to predict affect, such as Electrodermal Activity (EDA), since it is related to emotional arousal [3]. Once we are able to reliably detect student affect, we can leverage this information in order to develop interventions that will help increase task success, engagement, and user satisfaction.

Acknowledgments. We would like to thank Daria Bondareva for her work on the data processing and validation portion of this project. This research was supported by NSERC, SSHRC, NSF, and Microsoft Research.

References

1. AlZoubi, O., D'Mello, S., and Calvo, R. Detecting naturalistic expressions of nonbasic affect using physiological signals. (2012).
2. Anderson, J.R. and Gluck, K. What role do cognitive architectures play in intelligent tutoring systems. *Cognition & Instruction,* (2001), 227–262.
3. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., and Christopherson, R. Emotion sensors go to school. *AIED, July 6-10, Brighton, UK, IOS Press*, (2009), 17–24.
4. Azevedo, R., Harley, J., Trevors, G., et al. Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In *IHMLT*. Springer, 2013, 427–449.
5. Azevedo, R., Johnson, A., Chauncey, A., and Burkett, C. Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In *New Science of Learning*. Springer, 2010, 225–247.
6. Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R., and Bouchet, F. Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning. *AIED*, (2013), 229–238.
7. Conati, C. and Maclaren, H. Empirically building and evaluating a probabilistic model of user affect. *UMUAI 19*, 3 (2009), 267–303.
8. D Mello, S. and Graesser, A. Mind and body: Dialogue and posture for affect detection in learning environments. *FAIA 158*, (2007), 161.
9. D'Mello, S., Graesser, A., and Picard, R.W. Toward an affect-sensitive AutoTutor. *Intelligent Systems, IEEE 22*, 4 (2007), 53–61.
10. D'Mello, S., Olney, A., Williams, C., and Hays, P. Gaze tutor: A gaze-reactive intelligent tutoring system. *IJHCS 70*, 5 (2012), 377–398.
11. D'mello, S., Craig, S., Gholson, B., Franklin, S., Picard, R., and Graesser, A. Integrating affect sensors in an intelligent tutoring system. *Affective Interactions*, (2005), 7–13.

12. Field, A. *Discovering statistics using SPSS*. Sage publications, 2009.
13. Forbes-Riley, K., Litman, D., Friedberg, H., and Drummond, J. Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. *NAACL: Human Language Technologies*, (2012), 91–102.
14. Gutica and Conati. Student Emotions with an Edu-Game: A Detailed Analysis. (2013).
15. Harley, J.M., Bouchet, F., and Azevedo, R. Aligning and Comparing Data on Emotions Experienced during Learning with MetaTutor. *AIED*, (2013), 61–70.
16. Hussain, M.S. and Calvo, R.A. Multimodal affect detection from physiological and facial features during ITS interaction. *AIED*, Springer (2011), 472–474.
17. Kardan, S. and Conati, C. Exploring gaze data for determining user learning with an interactive simulation. Springer (2012), 126–138.
18. Klein, J., Moon, Y., and Picard, R.W. This computer responds to user frustration:: Theory, design, and results. *Interacting with computers 14*, 2 (2002), 119–140.
19. Kort, B., Reilly, R., Mostow, J., and Picard, R. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: adding human-provided emotional scaffolding to an automated reading tutor that listens. *ICMI*, (2002), 483.
20. Muldner, K., Atkinson, R., and Burleson, W. Investigating the utility of eye-tracking information on affect and reasoning for user modeling. *UMAP*, (2009), 138–149.
21. Pekrun, R., Goetz, T., Titz, W., and Perry, R.P. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist 37*, 2 (2002), 91–105.
22. Pekrun, R. Emotions as drivers of learning and cognitive development. In *New perspectives on affect and learning technologies*. Springer, 2011, 23–39.
23. Prendinger, H. and Ishizuka, M. The empathic companion: A character-based interface that addresses users'affective states. *APAI 19*, 3-4 (2005), 267–285.
24. Sabourin, J., Mott, B., and Lester, J.C. Modeling learner affect with theoretically grounded dynamic bayesian networks. In *ACII*. Springer, 2011, 286–295.
25. Sabourin, J., Shores, L., Mott, B., and Lester, J. Predicting student self-regulation strategies in game-based learning environments. (2012), 141–150.
26. Sibert, J.L., Gokturk, M., and Lavine, R.A. The reading assistant: eye gaze triggered auditory prompting for reading remediation. ACM Press (2000), 101–107.
27. Silvia, P.J. Interest—The curious emotion. *Current Directions in Psychological Science 17*, 1 (2008), 57–60.
28. Smilek, D., Carriere, J.S., and Cheyne, J.A. Out of Mind, Out of Sight Eye Blinking as Indicator and Embodiment of Mind Wandering. *Psych. Sci. 21*, 6 (2010), 786–789.
29. Wang, H., Chignell, M., and Ishizuka, M. Empathic tutoring software agents using real-time eye tracking. *ETRA*, (2006), 73–78.
30. Wang, N., Johnson, W.L., Mayer, R.E., Rizzo, P., Shaw, E., and Collins, H. The politeness effect: Pedagogical agents and learning outcomes. *IJHCS 66*, 2 (2008), 98–112.
31. Wang, W., Li, Z., Wang, Y., and Chen, F. Indexing cognitive workload based on pupillary response under luminance and emotional changes. *IUI*, (2013), 247–256.
32. Zimmerman, B.J. Self-efficacy: An essential motive to learn. *Contemporary educational psychology 25*, 1 (2000), 82–91.