Evaluating A Probabilistic Model of Student Affect

Cristina Conati[1] and Heather Maclaren[1]

[1] Dept. of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC, V6T 1Z4, Canada
{conati, maclaren}@cs.ubc.ca

Abstract. We present the empirical evaluation of a probabilistic model of student affect based on Dynamic Bayesian Networks and designed to detect multiple emotions. Most existing affective user models focus on recognizing a specific emotion or lower level measures of emotional arousal, and none of these models have been evaluated with real users. We discuss our study in terms of the accuracy of various model components that contribute to the assessment of student emotions. The results provide encouraging evidence on the effectiveness of our approach, as well as invaluable insights on how to improve the model's performance.

1 Introduction

Electronic games for education are learning environments that try to increase student motivation by embedding pedagogical activities in highly engaging, game-like interactions. Several studies have shown that these games are usually successful at increasing the level of student engagement, but they often fail to trigger learning [10] because students play the game without actively reasoning about the underlying instructional domain. To overcome this limitation, we are designing pedagogical agents that generate tailored interactions to improve student learning during game playing. In order not to interfere with the student's level of engagement, these agents should take into account the student's affective state (as well as their cognitive state) when determining when and how to intervene. However, understanding someone's emotions is hard, even for human beings. The difficulty is largely due to the high level of ambiguity in the mapping between emotional states, their *causes* and their *effects* [12].

One possible approach to tackling the challenge of recognizing user affect is to reduce the ambiguity in the modeling task, either by focusing on a specific emotion in a fairly constraining interaction (e.g. [9]) or by only recognizing emotion intensity and valence (e.g. [1]). In contrast, our goal is to devise a framework for affective modeling that pedagogical agents can use to detect multiple specific emotions in interactions in which this information can improve the effectiveness of the adaptive support provided. To handle the high level of uncertainty in this modeling task, the framework integrates in a Dynamic Bayesian Network (DBN [8]) information on both the *causes* of a student's emotional reactions and their *effects* on the student's bodily expressions.

Model construction is done as much as possible from data, integrated with relevant psychological theories of emotion and personality.

While the model structure and construction is described in previous publications [3,13], in this paper we focus on model evaluation. In particular, we focus on evaluating the *causal part* of the model. To our knowledge, whilst there have been user studies to evaluate sources of affective data (e.g., [2]), this is the first empirical evaluation of an affective user model, embedded in a real system and tested with real users.

We start by describing our general framework for affective modeling. We then summarize how we built the causal part of the model for Prime Climb, an educational game for number factorization. Finally we describe the user study, its results and the insights that it generated on how to improve the model's accuracy.

2 A DBN for Emotion Recognition

Fig. 1 shows two time slices of our DBN for affective modeling. The nodes represent classes of variables in the actual DBN, which combines evidence on both causes and effects of emotional reactions, to compensate for the fact that often evidence on causes or effects alone is insufficient to accurately assess the student's emotional state.

The part of the network above the nodes *Emotional States* represents the relations between possible causes and emotional states, as they are described in the OCC theory of emotions [11]. In this theory, emotions arise as a result of one's *appraisal* of the current situation in relation to one's goals. Thus, our DBN includes variables for *Goals* that a student may have during interaction with the game. Situations consist of the outcome of any event caused by either a student's or an agent's action (nodes *Student Action Outcome* and *Agent Action Outcome* in Fig. 1). Agent actions are represented as decision variables, indicating points where the agent must decide how to intervene in the interaction. The desirability of an event in relation to the student's

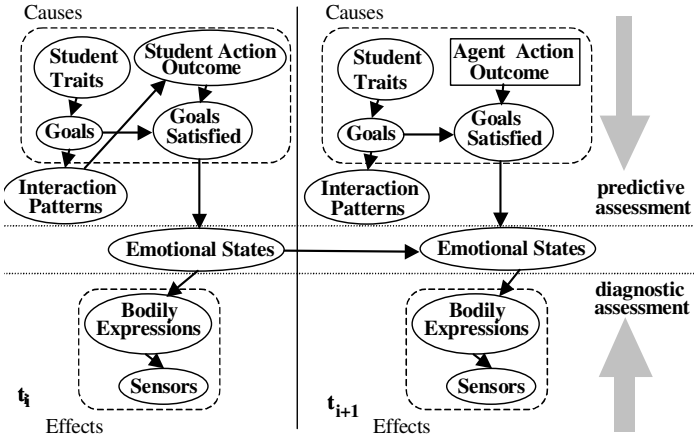

Fig. 1. Two time slices of our general affective model

goals is represented by the node class *Goals Satisfied,* which in turn influences the student's *Emotional States*.

Assessing student goals is non-trivial, especially when asking the student directly is not an option (as is the case in educational games). Thus, our DBN includes nodes to infer student goals from both *User Traits* that are known to influence goals (such as personality [7]) and *Interaction Patterns*.

The part of the network below *Emotional States* represents the interaction between emotional states, their observable effects on student behavior (*Bodily Expressions*) and sensors that can detect them. It is designed to modularly combine any available sensor information, to compensate for the fact that a single sensor can seldom reliably identify a specific emotional state.

In the next section, we show how we instantiated the causal part of the model to assess students' emotions during the interaction with the Prime Climb educational game. For details on the diagnostic part see [5].

3 Causal Model Construction for Prime Climb

Fig. 2 shows a screenshot of Prime Climb, a game designed to teach number factorization to 6th and 7th grade students. Two players must cooperate to climb a series of mountains that are divided in numbered sectors. Each player should move to a number that does not share any factors with her partner's number, otherwise she falls. Prime Climb provides two tools to help students: a *magnifying glass* to see a number's factorization, and a *help box* to communicate with the pedagogical agent we are building for the game. In addition to providing help when a student is playing with a partner, the agent engages its player in a "Practice Climb" during which it climbs with the student as a climbing instructor. The affective user model described here assesses the player's emotions during these practice climbs, and will eventually be integrated with a model of student learning [6] to inform the agent's pedagogical decisions.

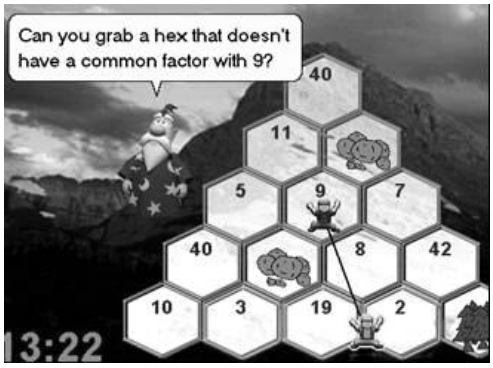

Fig. 2. Prime Climb interface

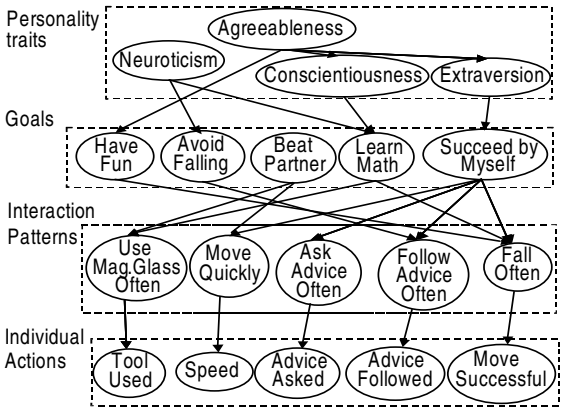

Fig. 3. Sub-network for goal assessment

We start by summarizing how we defined the sub-network that assesses students' goals. For more details on the process see [13]. Because all the variables in this sub-network are observable, we identified the variables and built the corresponding conditional probability tables (CPTs) using data collected through a Wizard of Oz study where students interacted with the game whilst an experimenter guided the pedagogical agent. The students took a pretest on factorization knowledge, a personality test based on the Five Factor personality theory [7], and a post-game questionnaire to express what goals they had during the interaction. The probabilistic dependencies among goals, personalities, interaction patterns and student actions were established through correlation analysis between the test results, the questionnaire results and student actions logged during the interactions.

Fig. 3 shows the resulting sub-network, incorporating both positive and negative correlations. The bottom level specifies how interaction patterns are recognized from the relative frequency of individual actions [13]. We intended to represent different degrees of personality type and goal priority by using multiple values in the corresponding nodes. However, we did not have enough data to populate the larger CPTs and resorted to binary nodes. Let's consider now the part of the network that represents the *appraisal mechanism* (i.e. how the mapping between student goals and game states influences student emotions). We currently represent in our DBN only 6 of the 22 emotions defined in the OCC model. They are *joy /distress* for the current state of the game, *pride/shame* of the student toward herself, and *admiration/reproach* toward the agent, modeled in the network by three two-valued nodes: *emotion for event*, *emotion for self* and *emotion for agent* (see Fig. 4).

The links and CPTs between *Goal* nodes, the outcome of student or agent actions and *Goal Satisfied* nodes, are currently based on subjective judgment. For some of these links, the connections are quite obvious. For instance, if the student has the goal *Avoid Falling*, a move resulting in a fall will lower the probability that the goal is achieved. Other links (e.g., those modeling which student actions cause a student to have fun or learn math) are less obvious, and could be built only through explicit

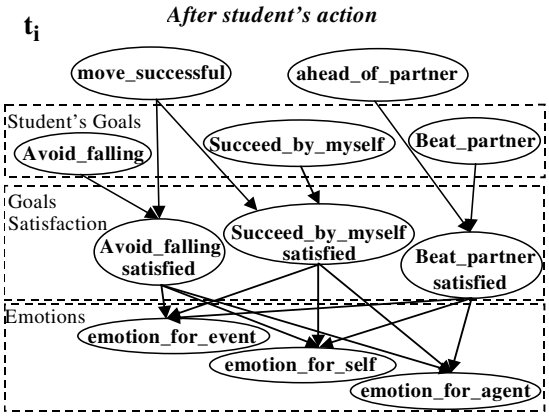

Fig. 4. Sample sub-network for appraisal

student interviews that we had no way to conduct during our studies. When we did not have good heuristics to create these links, we did not include them in the model. The links between *Goal Satisfied* nodes and the emotion nodes are defined as follows. We assume that the outcome of every agent or student action is subject to student appraisal. Thus, each *Goal Satisfied* node influences *emotion-for-event* in every slice. Whether a *Goal Satisfied* node influences *emotion-for-self* or *emotion-for-agent* in a given slice depends upon whether the slice was generated, respectively, by a student action (slice t_i in Fig. 4) or agent's action (not shown due to lack of space). The CPTs for emotion nodes are defined so that the probability of each positive emotion is proportional to the number of true *Goal Satisfied* nodes.

4 Evaluation

In order to gain an idea of how approximation due to lack of data affected the causal affective model we ran a study to produce an empirical evaluation of its accuracy. However, evaluating an affective user model directly is difficult. It requires assessing the students' actual emotions, which are ephemeral and can change multiple times during the interaction. Therefore it is not feasible to ask the students to describe them after game playing. Asking the students to describe them *during* the interaction, if not done properly, can significantly interfere with the very emotional states that we want to assess. Pilot testing various ways to try this second option showed that the least intrusive solution consisted of using two identical dialogue boxes [4]. One dialogue box (Fig. 5) is always available next to the game window for students to input their emotional states spontaneously. A similar dialogue box pops up if a student does not do this frequently enough, or if the model assesses that the student's emotional state has likely changed. Students were asked to report feelings toward the game and the agent only, as it was felt that our 11-year-old subjects would be too confused if asked

Fig. 5. The dialogue box presented to the students

to describe three separate feelings.

20 7th grade students participated in the study, run in a local school. They were told that they would be playing a game with a computer-based agent that was trying to understand their needs and help them play the game better. Therefore, the students were encouraged to provide their feelings whenever their emotions changed so that the agent could adapt its behavior. In reality, the agent was directed by an experimenter who was instructed to provide help if the student showed difficulties with the climbing task. Help was provided through a Wizard of Oz interface that allowed the experimenter to generate hints at different levels of detail. All of the experimenter's and student's actions were captured by the affective model, which was updated in real time to direct the appearance of the additional dialogue box, as described earlier. Students filled the same personality test and goal questionnaire used in previous studies. Log files of the interaction included the student's reported emotions and corresponding model assessment.

4.1 Results: Accuracy of Emotion Assessment

We start our data analysis by measuring how often the model's assessment agreed with the student's reported emotion. We translated the students' reports for each emotion pair (e.g. joy/distress) and the model's corresponding probabilistic assessment into 3 values; 'positive' (any report higher than 'neutral' in the dialogue box), 'negative' (any report lower than 'neutral') and 'neutral' itself. If the model's assessment was above a simple threshold then it was predicting a positive emotion, if not then it was predicting a negative emotion. We did not include a 'neutral' value in the model's emotion nodes because we did not have sufficient knowledge from previous studies to populate the corresponding CPTs.

Making a binary prediction from the model's assessment is guaranteed to disagree with any neutral reports given. However, we found that 25 student reports (53% and 35% of the neutral joy and admiration reports respectively) were neutral for both joy and admiration. If, as these reports indicate, the student had a low level of emotional arousal, then this state that can be easily picked up by biometric sensors in the diagno-

Table 1. Emotional belief accuracy using data from all 20 students

	Accuracy (%)		
Emotion	Mean	Std. Dev.	Data points
Joy	68.63	2.49	121
Distress	91.67	14.43	9
Combined J/D	80.15		
Admiration	20.61	3.39	97
Reproach	81.67	22.91	6
Combined A/R	51.14		

Table 2. Affective belief accuracy using data from 17 students with declared goals

	Accuracy (%)	
Emotion	Without Goals	With Goals
Joy	66.39	50.45
Distress	85.71	57.14
Combined J/D	76.05	53.78
Admiration	20.52	62.87
Reproach	75	75
Combined A/R	47.76	68.94

-stic part of the model [5]. This is a clear example of a situation where the observed evidence of a student's emotional state can inform the causal assessment of the model.

Using a threshold to classify the model's belief as positive or negative involves a trade-off between correctly classifying positive and negative emotions. We could argue that it will be more crucial for the pedagogical agent to accurately detect negative emotional states, but for the purpose of this evaluation we gave equal weight to positive and negative accuracy. Using this approach, threshold analysis showed that values between 0.6 and 0.7 produced the best overall results. We used the results at value 0.65, shown in Table 1, as the starting point for our data analysis.

The results were obtained from the model without using prior knowledge on individual students (i.e. the root personality nodes were initialized to 0.5 for every subject). For each emotion, we calculated the percentage of reports where the model agreed with the student. To determine whether any students had significantly different accuracy, we performed cross-validation to produce a measure of standard deviation. This measure is quite high for reproach and distress because far fewer data points were recorded for these negative emotions, but it is low for the other emotions, showing that the model produced similar performances for each student.

Table 1 shows that the combined accuracy for admiration/reproach is much lower than the combined accuracy for joy/distress. To determine to what extent these results are due to problems with the sub-network assessing student goals or with the sub-network modeling the appraisal process, we analyzed how the accuracy changed if we

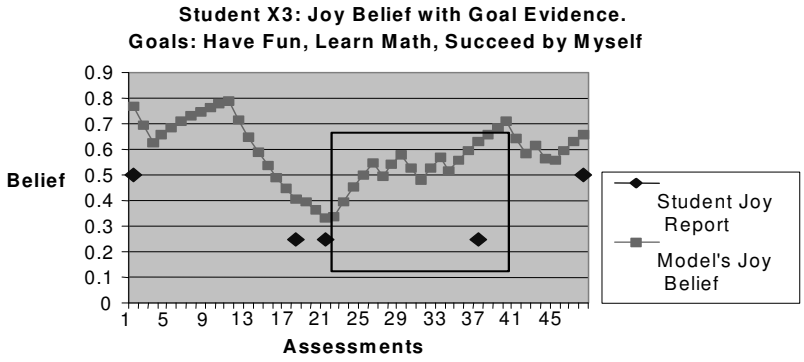

Fig. 6. A game session in which the student experienced frustration

added evidence on student goals into the model, simulating a situation in which the model assesses goals correctly.

Table 2 shows that, when we add evidence on student goals, the accuracy for admiration improves, but the accuracy for joy is reduced. To understand why, we took a closer look at the data for individual students. While the increase in accuracy for admiration was a general improvement for all students who reported this emotion, the decreases in accuracy for joy and distress were due to a small number of students for whom the model no longer gave a good performance. We have identified 2 reasons for this result:

Reason 1. As we mentioned in a previous section, we currently have no links connecting student actions to the satisfaction of the goals Have Fun and Learn Math because we did not have sufficient knowledge to build these links. However, in this study, 4 students reported that they only had goals of Have Fun or Learn Math (or both). For these students, the model's belief for joy only changed after agent actions. Since the agent acted infrequently, the model's joy belief changed very little from its initial value of 0.5. Thus, because of the 0.65 threshold, all student reports for joy/distress were classified as distress, and the model's accuracy for this emotion pair was reduced. Removing these 4 students from the data set improved the accuracy for detecting joy when goal evidence was used from 50% to 74%. An obvious fix for this problem is to add to the model the links that relate the goals Have Fun and Learn Math to student actions. We plan to run a study explicitly designed to gather the relevant information from student interviews after game playing.

Reason 2. Of the 7 distress reports collected, 4 were not classified correctly because they occurred in a particular game situation. The section of the graph within the rectangle in Fig. 6 shows the comparison between the model's assessment and the student's reported emotions (normalized between 0 and 1 for the sake of comparison) during one such occurrence. In this segment of the interaction, the student falls and then makes a rapid series of successful climbs to get back to the position that she fell from. She then falls again and repeats the process until eventually she solves the problem. This student has declared the goals Have Fun, Learn Math, and Succeed by My-

self but, for reason 1 above, only the latter goal influences the student's emotional state after a student action. Thus, each fall reduces the model's belief for joy because the student is not succeeding. Each successful move without the agent's help (i.e. in most of her moves) increases the model's belief for joy. However, apparently the model overestimated how quickly the student's level of joy recovered because of the successful moves. This was the case for all students whose reports of distress were misclassified. In order to fix this problem the model needs a long-term assessment of the student's overall mood that will influence the priorities of student goals. It also needs an indication of whether moves represent actual progress in the game, adding links that relate this to the satisfaction of the goal Have Fun. Finally, we can use personality information to distinguish between students who experience frustration in such a situation and those who are merely 'playing' (some students enjoy falling and do not care about succeeding).

The improvement in the accuracy of emotional assessment (after taking into account the problems just discussed) when goal evidence is included shows that the model was not always accurate in predicting student goals. Why then was the accuracy for joy and distress so high when goal evidence was not included? Without this information, the model's belief for each goal tended to stay close to its initial value of 0.5, indicating that it did not know whether the student had the goal or not. Because successful moves can satisfy three out of the five goals in the model (Succeed by Myself, Avoid Falling and Beat Partner) and all students moved successfully more often than they fell, the model's assessment for joy tended to stay above the threshold value of 0.65, leading to a high number of reports being classified as joy. Most of the 5 distress reports related to the frustrating situations described earlier were also classified correctly. This is because the model did not correctly assess the fact that all the students involved in these situations had the goal Succeed by Myself and therefore did not overestimate the rising of joy as it did in the presence of goal evidence. This behavior may suggest that we don't always need an accurate assessment of goals to have an acceptable model of student affect. However, we argue that knowing the exact causes of the student's affective states can help an intelligent agent to react to these states more effectively. Thus, the next stage of our analysis relates to understanding the model's performance in assessing goals and how to improve it. In particular we explore whether having information on personality and interaction patterns is enough to accurately determine a person's goals.

4.2 Results: Accuracy of Goal Assessment

Only 10 students completed the personality test in our study. Table 3 shows, for each goal, the percentage of these students for whom the declaration of that goal was correctly identified, and how these percentages change when personality information is used. A threshold of 0.6 was used to determine whether the model thought that a student had a particular goal, because goals will begin to substantially affect the assessment of student emotions at this level of belief. The results show that personality information improves the accuracy for only two of the five goals, Have Fun and Beat Partner. For the other goals, the results appear to indicate that the model's belief about

Table 3. The model's goal assessment accuracy when personality information is included.

	Accuracy (%)	
Goals	Without Personality	With Personality
Have Fun	30	100
Avoid Falling	70	70
Beat Partner	60	80
Learn Math	50	50
Succeed by Myself	40	40

these goals did not change. However, what actually happened is that in these cases the belief simply did not change enough to alter the models predictions using the threshold.

The model's belief about a student's goals is constructed from causal knowledge (personality traits) and evidence (student actions). Fig. 3 showed the actions identified as evidence for particular goals . When personality traits are used, they produce an initial bias towards a particular set of goals. Evidence collected during the game should then refine this bias, because personality traits alone cannot always accurately assess which goals a student has. However, currently the bias produced by personality information is stronger than the evidence coming from game actions. There are two reasons for this strong bias:

Reason 1. Unfortunately, some of the actions collected as evidence (e.g. asking the agent for advice) did not occur very frequently, even when the student declared the particular goal that the action was evidence for. One possible solution is to add to the model a goal prior for each of the covered goals. The priors would be produced by a short test before the game and only act as an initial influence since the model's goal assessments will be dynamically refined by evidence. Integration of the prior information with the information on personality and interaction patterns will require fictitious root goal nodes to be added the model.

Reason 2. Two of the personality traits that affect the three goals Learn Math, Avoid Falling, and Succeed by Myself (see Fig. 3) are Neuroticism and Extraversion. However, the significant correlations that are represented by the links connecting these goals and personality traits were based on very few data points. This has probably led to stronger correlations than would be found in the general population. Because evidence coming from interaction patterns is often not strong enough (see Reason 1 above), then the model is not able to recover from the bias that evidence on these two personality traits brings to the model assessment. An obvious fix to these problems is to collect more data to refine the links between personality and goals.

5 Discussion and Future Work

In this paper, we have discussed the evaluation of a probabilistic model of student affect that relies on DBNs to assess multiple student emotions during interaction with educational games. Although other researchers have started using this probabilistic approach to deal with the high level of uncertainty involved in recognizing multiple user emotions (e.g. [3,12]), so far there has been no empirical evaluation of the proposed models, or of any other existing affective user model for that matter.

The results presented show that if a student's goals can be correctly determined, then the affective model described can maintain a fairly accurate assessment of the student's current emotional state. Furthermore, we can increase this accuracy by implementing the solutions that we described to overcome the sources of error detected in the model structure and parameters. Accurate assessment of student goals, however, has been shown to be more problematic, which is not surprising given that what we are trying to do is basically *plan recognition*, which is one of AI's notoriously difficult problems. We reported two main sources of inaccuracy in goal assessment in our model, and presented suggestions on how to tackle them. However, it is unlikely that we will ever achieve consistently high accuracy in goal assessment for all students in all situations. This is where having a model that combines information on both causes and effects of emotional reaction can compensate for the fact that often evidence on causes or effects alone is insufficient to accurately assess the student's emotional state. Thus, we believe that our results provide encouraging evidence that confirms the potential of using DBNs to successfully model user affect in general.

In addition to collecting more data to refine the model as suggested by our data analysis, other improvements that we are planning include (1) investigating adding to the model varying degrees of goal priority and personality traits (2) combining the causal part of the model with a diagnostic model [5], that makes use of evidence from biometric sensors, to produce a model that integrates both causes and effects into a single emotional assessment.

References

1. Ball, G. and Breese, J. 1999. Modeling the Emotional State of Computer Users. *Workshop on 'Attitude, Personality and Emotions in User-Adapted Interaction', UM'99*, Canada.
2. Bosma, W. and André, E. 2004. Recognizing Emotions to Disambiguate Dialogue Acts. *International Conference on Intelligent User Interfaces 2004*. Madeira, Portugal.
3. Conati, C. 2002. Probabilistic Assessment of User's Emotions in Educational Games. *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI"*, **16**(7-8):555-575.
4. Conati, C. 2004. How to Evaluate models of User Affect? *Tutorial and Research Workshop on Affective Dialogue Systems*. Kloster Irsee, Germany.
5. Conati, C., Chabbal, R., and Maclaren, H. 2003. A Study on Using Biometric Sensors for Monitoring User Emotions in Educational Games. *Workshop on Modeling User Affect and Actions: Why, When and How. *UM'03, Int. conf. On User Modeling*. Johnstown, PA.

6. Conati C. and Zhao X. 2004. Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game. *International Conference on Intelligent User Interfaces 2004*. Madeira, Portugal.

7. Costa, P.T. and McCrae, R.R. 1992. Four Ways Five Factors are Basic. *Personality and Individual Differences* 1. **13**:653-665.

8. Dean, T. and Kanazawa, K. 1989. A Model for Reasoning about Persistence and Causation. *Computational Intelligence* **5**(3):142-150.

9. Healy, J. and Picard, R. 2000. SmartCar: Detecting Driver Stress. *15th Int. Conf. on Pattern Recognition*. Barcelona, Spain.

10. Klawe, M. 1998. When Does The User Of Computer Games And Other Interactive Multimedia Software Help Students Learn Mathematics? *NCTM Standards 2000 Technology Conference*, 1998. Arlington, VA.

11. Ortony, A., Clore, G.L., and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

12. Picard, R. 1995. *Affective computing*. Cambridge: MIT Press.

13. Zhou, X. and Conati, C. 2003. Inferring User Goals from Personality and Behavior in a Causal Model of User Affect. *International Conference on Intelligent User Interfaces 2003*. Miami, FL.