Data-Driven Refinement of a Probabilistic Model of User Affect

Cristina Conati and Heather Maclaren

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

Abstract. We present further developments in our work on using data from real users to build a probabilistic model of user affect based on Dynamic Bayesian Networks (DBNs) and designed to detect multiple emotions. We present analysis and solutions for inaccuracies identified by a previous evaluation; refining the model's appraisals of events to reflect more closely those of real users. Our findings lead us to challenge previously made assumptions and produce insights into directions for further improvement.

1 Introduction

The assessment of users' affect is increasingly recognized as an informative task when attempting to improve the effectiveness of interactive systems. Information on the user's affective state is particularly important when the user is focused on a highly engaging task where inappropriate system interventions may be especially disruptive, such as learning in simulated environments and educational games.

Educational games attempt to stimulate student learning by embedding pedagogical activities within a highly engaging, game like environment. We are working to improve the pedagogical effectiveness of these games by producing intelligent agents that monitor the student's learning progress and generate tailored interactions to improve learning during game playing. To avoid interfering with the student's level of engagement, these agents should take into account the student's affective state (in addition to her cognitive state) when determining when and how to intervene.

Assessment of emotions, particularly the multiple specific emotions that educational games can generate, is very difficult because the mapping between emotions, their causes, and their effects is highly ambiguous [10]. However, we believe that information on specific emotions may enable more precise and effective agent's interventions than a simpler assessment of arousal or valence (e.g.[1]), or stress [7]. To handle the high level of uncertainty in this modeling task, we have devised a framework for affective modeling that integrates in a Dynamic Bayesian Network (DBN) information on both the *causes* of a user's emotions and their *effects* on the user's behavior. Model construction is done as much as possible from data, integrated with relevant psychological theories of emotion and personality. The inherent difficulties

L. Ardissono, P. Brna, and A. Mitrovic (Eds.): UM 2005, LNAI 3538, pp. 40 – 49, 2005.

of this task include: the novel nature of the phenomena that we are trying to model, the limited existing knowledge of users' emotional reactions during system interaction, especially within the context of educational games, and the difficulty of observing variables that are key to the assessment of affect.

We have been using data collected in a series of studies (e.g. [3,11]) to construct a probabilistic model of the user's affective state that is based on the OCC model of emotions [9]. The data from our most recent study [4] was used to evaluate the model we have built so far. Although there have been evaluations using aggregated data [6] and evaluations of sources of affective data (e.g.[2]), to the best of our knowledge this is currently the only evaluation of an affective user model embedded in a real system and tested with individual users. Our results showed that if the user's goals could be correctly assessed then the model could produce reasonably accurate predictions of user affect, but also revealed some sources of inaccuracy that needed to be addressed. We recognize that the assessment of the user's goals must be improved before the model can be used autonomously within a real system. However, solutions for the other sources of inaccuracy within the model's emotional assessment will help clarify the full requirements of the goal assessment task.

In this paper we address previously identified inaccuracies within the model's mechanism of emotional appraisal. We then re-evaluate the refined model, producing insights into additional refinements that would produce further improvement.

2 The Affective User Model

Fig. 1 shows a high level representation of two time slices of our affective model. 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 [9]. 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 user may have during the interaction with an education game and its embedded pedagogical agent (for details on goal assessment see [11]). Situations consist of the outcome of any event caused by either a user's or an agent's action (nodes *User*

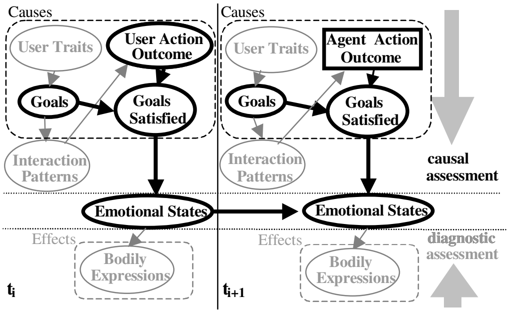

Fig. 1. Two slices of our general affective model

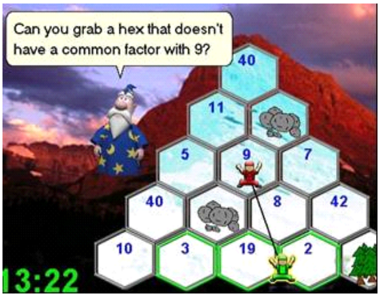
Can you grab a hex that doesn't have a common factor with 9?
40
11
5
9
7
40
8
42
10
3
19
2
13:22

Fig. 2. Prime Climb interface

Action Outcome and *Agent Action Outcome*). An event's desirability in relation to the user's goals is represented by *Goals Satisfied,* which in turn influences the user's *Emotional States*. The part of the network below the nodes *Emotional States* provides diagnostic assessment from bodily reactions known to correlate with emotions.

We have instantiated and evaluated the causal part of the model to assess players' emotions during the interaction with the Prime Climb educational game. In the rest of the paper we will focus on the refinement and evaluation of the appraisal part of this causal model (the bold nodes and links in Figure 1).

2.1 Causal Affective Assessment for Prime Climb

Figure 2 shows a screenshot of PrimeClimb, a game designed to teach number factorization to 6th and 7th grade students. In the game, 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, 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 pedagogical agent engages its player in a "Practice Climb" during which it climbs with the student as a climbing instructor.

The affective model described here assesses the student's emotions during these practice climbs. Figure 3 shows the appraisal part of this model created after the student makes a move. As the bottom part of the figure shows, we currently represent in our DBN 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 by three two-valued nodes: *emotion for game*, *emotion for self* and *emotion for agent.*

Let's now consider the workings of the part of the model that assesses the student's situation appraisal in Prime Climb. In this part of the model the links and Conditional Probability Tables (CPTs) between *Goal* nodes, the outcome of the student's or agent's action, and *Goal Satisfied* nodes were based on subjective judgment because our previous studies focused on collecting data to refine the model's assessment of student goals. For some links, the connections were quite obvious. For instance, if the student has the goal *Avoid Falling*, a move that results in a fall will lower the probability that the goal is achieved. For other goals, like *Have Fun* and *Learn Math*, the connections were not obvious and we did not have good heuristics to create the appraisal links. Thus we postponed including them in the model until we could collect data from which to determine an appropriate structure.

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-game* (*Joy* or *Distress*) in every slice. If a slice is generated by a student action then each *Goal Satisfied* node influences *emotion-for-self* (slice t_i in Fig. 3). If a slice is generated by an agent's intervention, then *emotion-for-agent* is influenced instead (slice not shown

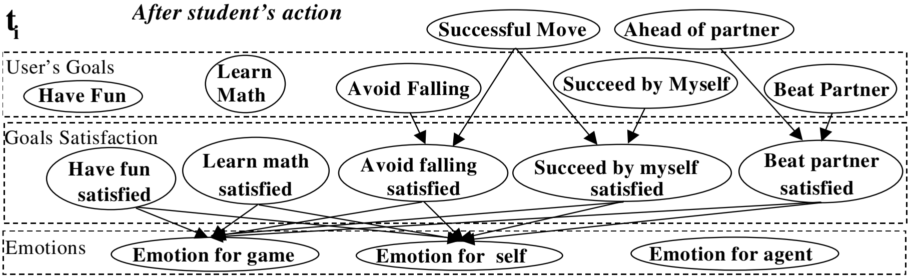

Fig. 3. Sample sub network for appraisal after student action

due to lack of space). We also assume that a student either has a goal or does not (i.e. we do not model goal priority) and that the student has the same goals throughout the game session. The CPTs for emotion nodes were defined so that the probability of the positive emotion is proportional to the number of true *Goal Satisfied* nodes.

When we evaluated the affective model that included this version of the appraisal component [4], we discovered two main sources of inaccuracy:

Source 1: *Joy* and *Distress* due to student actions. The absence of links (as shown in Fig. 3) between the outcome of a student's move and the satisfaction of goals *Have Fun* and *Learn Math* made the model underestimate the positive emotions towards the game for students that only had these goals. This reduced the model's accuracy for *Joy* from 74% to 50% and highlighted the need to collect data to create the missing links. The model also underestimated the negative emotions felt by some students when falling repeatedly and thus had low accuracy for *Distress* of 57%.

Source 2: *Admiration* and *Reproach* towards the agent. The subjective links between agent actions and goal satisfaction had caused the model to underestimate the students' positive feelings towards the agent. This produced an accuracy of 20.5% for *Admiration* and 75% for *Reproach*, further highlighting the need to collect data to refine the connections in the appraisal part of the model.

3 User Study

The general structure of this new study was similar to the previous one. Sixty-six 6th and 7th grade students from 3 local schools interacted with Prime Climb, and, during the interaction, were asked to report their feelings towards the game and towards the agent using simple dialogue boxes. However, while in the previous study the agent was directed in a Wizard of Oz fashion, in this study the agent was autonomous and based its interventions on a model of student learning [5]. While the model of student affect was dynamically updated during interaction, the pedagogical agent did not use it to direct its interventions. However, the assessments of the affective model were included in the log files, for comparison with the student's reported emotions.

As in the previous study, students completed a pre-test on number factorization, a post-questionnaire to indicate the goals they had during game playing, and a personality test. However, they also filled in two additional questionnaires, one on game events that could satisfy the goal *Have Fun* and one on events that could satisfy the goal *Learn Math*. Each questionnaire contained a list of statements of the type '*I learnt math/had fun when <event>*' which students rated using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). The events listed included:

For *Have Fun* – all student actions already in the model (a successful climb, a fall, using the magnifying glass, using the help box), *reaching the top of the mountain.*
For *Learn Math*

- all student actions already in the model (the same as above), *following the agent's advice,* and *encountering big numbers*.
- agent interventions already in the model that were intended to help the student learn math (reflect on reasons for success, reflect on reasons for failure), *think about common factors,* and *use the magnifying glass*.

The italicized items at the end of each list above had not been explicitly included in the model before, but were added based on anecdotal evidence suggesting that they may help to satisfy these goals. We did not ask students about agent actions that satisfied the goal *Have Fun* or other events that already satisfied other goals within the model due to limitations on time and to avoid students becoming fatigued.

4 Refinement of the Model's Causal Affective Assessment

Before discussing how we refined the model using data from the new study we describe how well the existing model performed on the new data set.

We measured the model's accuracy as the percentage of assessments that agreed with the students' reports for each emotion pair (e.g. *Joy/Distress*). If the model's corresponding assessment was above a simple threshold then it was predicting a positive emotion, if not then it was predicting a negative emotion. The threshold was determined using the data from our previous study [4].

Table 1 shows the accuracy obtained using three-fold cross-validation when the goals students declared in the questionnaire are used as evidence in the model; each iteration used one-third of the data as a test set. The results show that the inaccuracies discussed earlier still affect the model's performance on the new data set. The high variance for *Joy* is due to one test set containing some students who only had the goals *Have Fun* or *Learn Math*, thus the model underestimated their positive re-

Table 1. Emotional belief accuracy of the initial model for the new data set

Emotion	Accuracy (%)		
	Mean	Std. Dev.	Total data points
Joy	66.54	17.38	170
Distress	64.68	29.14	14
Combined J/D	65.61		
Admiration	43.22	12.53	127
Reproach	80.79	6.05	28
Combined A/R	62.00		

sponses. The high variance of *Distress* is due in part to the small number of data points, but it is also due to the model underestimating the negative feelings of some students who fell repeatedly. The low accuracy for *Admiration* and high accuracy for *Reproach* agree with the results of our previous study.

4.1 Assessment of Joy Due to Student Actions

The students' answers to the questionnaires indicated that all of the events related to student actions were relevant to some degree. We therefore scored all possible network structures using their log marginal likelihood [8], as we did for [11], in order to determine which events made a difference to the model's assessments. We found that (i) the outcome of the student's move influenced the satisfaction of the goal *Have Fun* and (ii) whether the student encountered a big number influenced the satisfaction of the goal *Learn Math*.

We included these findings in the model as follows. First, we added a node for the new event, *Big number,* and corresponding links to goal satisfaction nodes. We based our definition of a big number on the large numbers frequently incorrectly factorized in the students' pre-tests. Second, we used the study data to set the CPTs for the goal satisfaction nodes for *Have Fun* and *Learn Math.* Fig. 4 shows the revised time slice. Each new node and link is drawn using heavier lines.

4.2 Appraisal of Agent Actions

As mentioned earlier, the model's initial accuracy of assessing emotions towards the agent showed that we needed to revise and refine the existing links modeling how appraisal of the agent's actions affects players' emotions. Data analysis targeting this goal consisted of two stages.

Stage 1. First, we analyzed students' questionnaire items related to the influence of agent's actions on the goal *Learn Math*. We scored all possible network structures using their log marginal likelihood and found that our current structure received the highest score. Therefore our only refinement to the model based on these findings was to use the study data to refine the CPTs linking agent actions to the satisfaction of the goal *Learn Math*. However, a preliminary evaluation of these changes showed that the model was still underestimating students' admiration toward the agent. Thus, we moved to a second stage of data analysis.

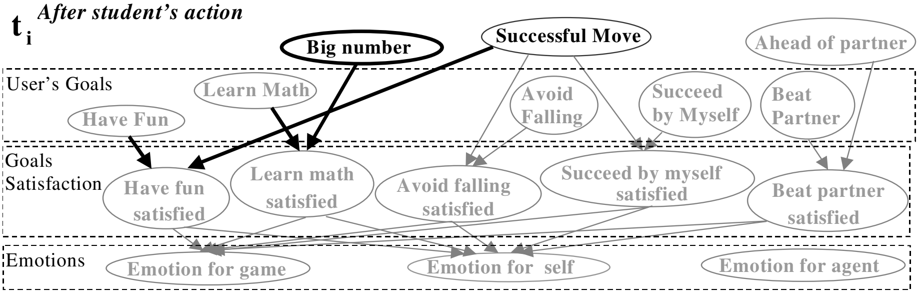

Fig. 4. Revised sub-network for appraisal after student action

Stage 2. We analyzed the log files of each student's session to identify situations in which students gave positive or negative reports towards the agent. The results are shown in Table 2. Congratulation by the agent (first row in Table 2) was already included in the original model as satisfying the goal *Have Fun*. Our data confirms that this action generates students' admiration, although it cannot tell whether this happens through the satisfaction of the goal *Have Fun*.

The second situation in Table 2 shows that students who are generally successful are usually either happy or neutral towards the agent, regardless of their goals. This suggests that the students' positive feelings toward the game will positively influence their attitude towards the agent. We translated this finding into the model by adding a link from the student's emotion towards the game in the previous time slice to the student's emotion towards the agent. This new link, and all the additions described below, can be seen in Figure 5.

The final two situations in Table 2 show reported feelings towards the agent when the student was falling and either received help or did not. Analysis of these situations revealed that approximately half of the students who reported reproach and half of the students who reported admiration when the agent intervened had declared the goal *Succeed By Myself.* This seems to indicate that, although some of the students may have wanted to succeed by themselves most of the time, when they began to fall they reduced the priority of this goal in favor of wanting help. This invalidates two of the choices previously made in the model implementation: (i) to ignore goal priority; (ii) to assume that goals are static during the interaction. Because we currently don't have enough data to model goal evolution in a principled way, we only addressed the implementation of multiple priority levels to model the relation between *Succeed By Myself* and wanting help. The model was changed as follows.

First, we added an additional goal, *Want Help*. The satisfaction of *Want Help* is dependent on two factors: the outcome of the student's move (i.e. a successful climb or a fall) and the agent's action. When the student falls, *Want Help* can only be satisfied if the agent provides help. If the agent congratulates the student, or does not perform any action, then this goal is not satisfied. If the student does not fall then satisfaction is neutral.

Second, we tried to determine which students' traits influenced their attitude towards receiving help during repeated falls. From our data, the only factor that seems to play a role is students' math knowledge. A Fisher test on the students' pre-test scores and whether they demonstrated that they wanted help showed a significant relationship (Fisher score = 0.029). Thus, a new node, representing prior math knowledge, was used to influence the priorities a student gives to the goals *Succeed By Myself* and *Want Help*. If the student has high knowledge, then satisfaction of *Want Help* is given

Table 2. Situations where students reported *Admiration* or *Reproach*

Situation	# Students reporting		
	Admiration	Neutral	Reproach
Student reaches mountain top, is congratulated by agent	12	13	2
Student is generally successful	26	19	4
Student falls frequently and agent intervenes	10	6	7
Student falls frequently and agent doesn't intervene	6	8	7

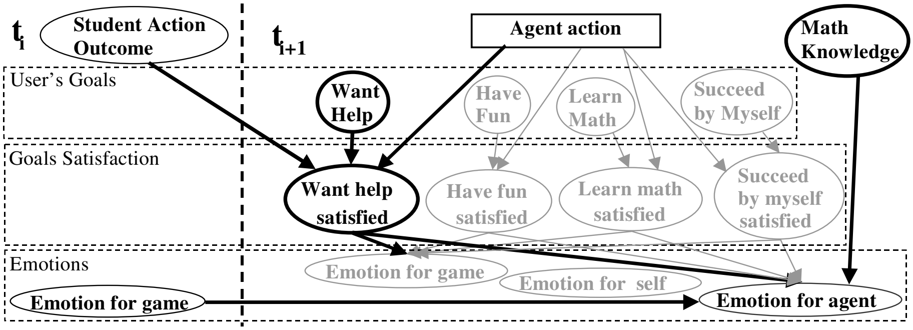

Fig. 5. Revised sub-network for appraisal after agent action

higher weight in the CPT defining the influence of goals satisfaction on emotion towards the agent. If the student has low knowledge, satisfaction of *Succeed By Myself* is given higher weight instead.

Third, the node representing the available agent's actions was refined to include the agent choosing not to intervene. All *Goal Satisfied* nodes other than *Succeed By Myself* and *Want Help* were given a neutral satisfaction for this new action. *Want Help* was discussed earlier, *Succeed By Myself* was given a small probability of satisfaction to reflect possible mild positive feelings towards the agent for not interrupting in general rather than at specific events.

4.3 Evaluation of the New Model

To evaluate the model changes discussed above, we replayed the event logs recorded during the study using a simulator that used the refined model. We added an additional 'no action' event after each student action that was not followed by an agent intervention. We performed cross-validation using the data from our current study; each iteration used two-thirds of the data to train the refined CPTs and one-third as a test set. Table 3 shows the results of the re-evaluation, when students' goals from the post-questionnaires are used as evidence in the model. To get evidence on the newly added goal *Want Help*, we relied on student answers to the questionnaire item *'I wanted help when I became stuck'*, originally used together with another item to assess the goal *Succeed By Myself.*

We start by discussing the accuracy results for *Admiration/Reproach*, because that will facilitate the discussion of *Joy/Distress*.

Accuracy of *Admiration/Reproach*. Table 3 shows that, although accuracy for *Admiration* improved considerably, accuracy for *Reproach* dropped off a comparable amount, bringing the combined accuracy to be slightly lower than the accuracy of the previous model. However, the high accuracy for *Reproach* in the previous model was a fortunate side effect of underestimating *Admiration*. Instead, an analysis of the model's assessment in relation to the interactions simulated from the log files shows that high accuracy for *Admiration* in the new model is mostly due to the added changes. The same analysis revealed that low accuracy for *Reproach* is mainly

Table 3. Emotional belief accuracy of the refined model

Emotion	Previous Accuracy (%)		Revised Accuracy (%)		Data points
	Mean	Std. Dev.	Mean	Std. Dev.	
Joy	66.54	17.38	76.26	1.75	170
Distress	64.68	29.14	71.30	40.48	14
Combined J/D	65.61		73.78		
Admiration	43.22	12.53	74.71	1.50	127
Reproach	80.79	6.05	38.23	19.23	28
Combined A/R	62.00		56.47		

due to two factors. First, goals declared by students at the end of a game session did not seem to match their goals throughout the game. Some students did not declare the goal *Want Help*, but their reports showed that they wanted help when they began to fall. Other students declared the goal but then did not want help. This is additional evidence that goal's priority can change during the interaction, and shows that the model is sensitive to these changes, confirming that in order to improve the model's accuracy we will have to lift the current model's assumption of static goals. Second, using only previous math knowledge to help assess each student's attitude toward wanting help incorrectly modeled some of the students. There appear to be other factors that should be taken into account, such as personality traits. We collected personality data during the study but encountered difficulties due to the general integrity of the students when describing their personality. We are investigating other methods for obtaining more reliable personality measurement.

Accuracy of *Joy/Distress*. As we can see from Table 3, the accuracy for *Joy* and *Distress* increased to about 76% and 71% respectively in the new model. The increase in *Joy* accuracy is mostly due to the changes discussed in Section 4. However, we should note that the impact of these changes is partially reduced by the goal fluctuation issues discussed above. Recall that the model's appraisal of agent actions also affects the assessment of *Joy* and *Distress* toward the game (Figure 5). From log file analysis, we saw that fluctuations of the goal *Want Help* made the model overestimate the negative impact of episodes of not receiving help for another group of 8 students who reported this goal, did not receive help when they were falling, but still reported joy toward the game and neutral or positive feelings toward the agent. It appears that, while we are correctly modeling the priority that these students give to the *satisfaction* of receiving help (thus the improved accuracy for admiration), we are overestimating the importance that they give to this goal *not being satisfied*. Thus, as it was the case for *Admiration/Reproach*, there appear to be other student traits that, if modeled, could further improve model accuracy.

The refinements made to assess *Admiration/Reproach* are the main reason for the improvement in *Distress*, because they correctly classified the *Distress* reports given by a student who was falling repeatedly, had the goal *Want Help* and did not receive help. These few correctly classified reports have high impact because of the limited number of *Distress* reports in the dataset (as the high variation for *Distress* shows). Note that this same student did not report *Reproach* during the same falling episodes, so he does not improve the model's *Reproach* accuracy.

5 Summary and Future Work

Building a user model of affect from real data is very difficult; the novel nature of the phenomena that we are trying to model, the limited existing knowledge of emotional reactions during system interaction, especially within the context of educational games, and the difficulty of observing key variables all contribute to the inherent complexity of the task.

In this paper, we have addressed sources of inaccuracy found within our model of user affect during a previous evaluation by refining the model's appraisal of both student and agent actions. We used data collected from real users to revise the relationship between game events and the satisfaction of two goals, *Have Fun* and *Learn Math*. We also used the data to analyze students' attitudes towards the agent and determined the common situations in which they changed. This analysis led to the introduction of a new goal, *Want Help*, the appraisal of the agent not giving help, and the first steps towards accommodating students giving different priorities to goals.

Our analysis has challenged two assumptions that were made during model construction; firstly that the set of goals the user is trying to achieve remains the same throughout the game session, secondly that we can make assessments using these goals without modeling goal priority. As part of our future work on revision of the model's goal assessment we intend to construct a clearer picture of how user's goals fluctuate during game sessions. We can then use this information to further improve the model's emotional assessment.

References

1. Ball, G. and Breese, J.: Modeling the Emotional State of Computer Users. Workshop on 'Attitude, Personality and Emotions in User-Adapted Interaction', UM'99, Canada (1999)
2. Bosma, W. and André, E.: Recognizing Emotions to Disambiguate Dialogue Acts. International Conference on Intelligent User Interfaces (IUI 2004). Madeira, Portugal (2004)
3. Conati, C.: Probabilistic Assessment of User's Emotions in Educational Games. Journal of Applied Artificial Intelligence 16(7-8), special issue: "Merging Cognition and Affect in HCI", (2002) 555-575
4. Conati, C. and Maclaren, H.: Evaluating A Probabilistic Model of Student Affect. Proceedings of the 7th Int. Conference on Intelligent Tutoring Systems, Maceio, Brazil (2004)
5. Conati, C. and Zhao, X.: Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game. IUI 2004. Madeira, Portugal (2004)
6. Gratch, J. and Marsella, S.: Evaluating the Modeling and Use of Emotion in Virtual Humans, 3rd Int. Jnt. Cnf. on Autonomous Agents and Multiagent Systems, New York (2004)
7. Healy, J. and Picard, R.: SmartCar: Detecting Driver Stress. 15th Int. Conf. on Pattern Recognition. Barcelona, Spain (2000)
8. Heckerman, D.: A Tutorial on Learning with Bayesian Networks, in Jordan, M. (ed.): Learning in Graphical Models (1998)
9. Ortony, A., Clore, G.L., and Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
10. Picard, R.: Affective Computing. Cambridge: MIT Press (1995)
11. Zhou, X. and Conati, C.: Inferring User Goals from Personality and Behavior in a Causal Model of User Affect. Int. Conference on Intelligent User Interfaces. Miami, FL (2003)