Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features

ABSTRACT

Affect sensitivity is of the utmost importance for a robot companion to be able to display socially intelligent behaviour, a key requirement for sustaining long-term interactions with humans. This paper explores a naturalistic scenario in which children play chess with the iCat, a robot companion. A person-independent, Bayesian approach to detect the user's engagement with the iCat robot is presented. Our framework models both causes and effects of engagement: features related to the user's non-verbal behaviour, the task and the companion's affective reactions are identified to predict the children's level of engagement. An experiment was carried out to train and validate our model. Results show that our approach based on multimodal integration of task and social interaction-based features outperforms those based solely on non-verbal behaviour or contextual information (94.79 % vs. 93.75% and 78.13%).

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*; J.4 [**Computer Applications**]: Social and Behavioural Sciences

General Terms

Algorithms, Human Factors, Design, Theory

Keywords

Affect recognition, human-robot interaction, non-verbal expressive behaviour, contextual information

1. INTRODUCTION

Many existing prototypes of robots still lack important social and affective skills. Their ability to display a socially acceptable behaviour is still limited and this prevents them from engaging in a truly natural interaction with users. Robot companions are an example of robots that may benefit from the integration of affective and social abilities [7] [10].

A prerequisite of companioninship is the ability to sustain long-term interactions. Affect sensitivity, i.e., the ability to understand the user's affective states and expressions, is of the utmost importance for a robot companion to be able to display socially intelligent behaviour, a key requirement for sustaining long-term interactions with humans. Robot companions should be sensitive to multimodal behavioural cues displayed by the user and the context in which the interaction with the user takes place in order to infer useful information regarding the ongoing situation. While affect recognition has been extensively addressed in the literature [21], many issues related to the design of a module for affect recognition to be integrated in a human-robot interaction framework still have to be investigated.

Designing an affect recognition framework for robot companions presents several challenges. A question of primary importance is what type of affective states and expressions a companion should be sensitive to. To answer this question we face a design issue: affective states and expressions that a companion may want to detect depend on many aspects related to the specific interaction scenario [8], for example whether the user is interacting with the companion or the companion is just monitoring the user's behaviour, the distance between user and companion (e.g., face-to-face or long-range interaction), whether the user is involved in some tasks with the companion or not, what is important for a companion to know in order to engage in something more than a short interaction with the user. For example, a socially intelligent companion would not try to recognise smiles when the user is in another part of the room or to detect fear when it needs to know whether the user is enjoying interacting with it. A companion should therefore be able to detect scenario-dependent affective states and cues.

Contrary to the majority of affect recognition systems reported in the literature, an affect recognition module for robot companions should be trained using naturalistic and spontaneous expressions and integrate contextual information [1]. The generation of an affective state during the interaction with a companion can be influenced by many different variables. Examples include the user's personality, gender, preferences, history and goals, the task, the presence of other people, the events unfolding in the environment, the type of behaviour displayed by the companion, etc. This suggests that possible causes of affective states should be taken into consideration as contextual information in the design of an affect recognition system for robot companions.

In this paper we propose a person-independent, Bayesian approach to detect the user's engagement with a game companion that takes into consideration causes and effects of engagement. Task and social interaction-based features are selected as source of information for the prediction of the level of engagement of children playing chess with an iCat robot. Experimental results show that the multimodal integration of contextual information with non-verbal cues displayed by the user improves the recognition of the user's engagement with the iCat robot. The present study addresses several challenges from the perspective of affect recognition for a robot companion. First, it focuses on a naturalistic interaction, where the user's states and expressions are spontaneous and scenario-dependent. Second, it considers multimodal information, which includes task-related features and social interaction cues, displayed both by the user and the companion. Finally, it addresses the issue of robustness in real-world scenarios: contextual features related to the task and the iCat's expressive behaviour, whose integration proved successful in the detection of engagement, represent valuable information that an affect recognition system could rely on in case of poor illumination conditions and noisy backgrounds.

The paper is organised as follows. The next Section provides an overview of previous work on recognition of spontaneous affective states in human-companion interaction, as well as previous attempts to infer affect by taking into consideration contextual information. Section 3 describes our interaction scenario and the proposed framework for the modelling of user engagement. Section 4 presents experimental results and reports the data collection process and the methodology used to train and evaluate our model. Finally, Section 5 summarises the characteristics of the proposed approach and the results.

2. RELATED WORK

A limited number of studies in affect recognition research have addressed the recognition of scenario-dependent, spontaneous affective states emerging during the interaction with an artificial companion. The system proposed by Kapoor et al. [11], for example, allows for the automatic prediction of frustration of students interacting with a learning companion by using multimodal non-verbal cues such as facial expressions, head movement, posture, skin conductance and mouse pressure data. Peters et al. [17] modelled the user's interest and engagement with a virtual agent displaying shared attention behaviour, by using eye gaze and head direction information.

While some efforts have been reported in the literature, works on affect recognition abilities for a specific interaction scenario that take into account contextual information are still not numerous. Kapoor and Picard [12] proposed an approach for the detection of interest in a learning environment by combining non-verbal cues and information about the learner's task (e.g., level of difficulty and state of the game). Conati et al. [9] designed Prime Climb, an educational game where an intelligent pedagogical agent helps users succeed in a math game. Inspired by the OCC model of cognitive appraisal [16], they combined information about the user's knowledge with a model for the prediction of multiple emotions experienced by users in order to determine how the agent could intervene so as to maximise the trade-off between student learning and engagement. Malta et al. [14] proposed a system for the multimodal estimation of a driver's irritation that exploits information about the driving context.

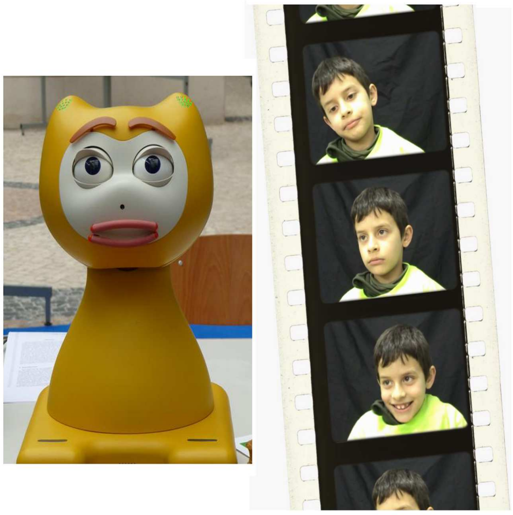

Figure 1: User interacting with the iCat robot.

Some studies showed how the use of contextual information can improve vision-based recognition of non-verbal behaviour. Morency et al. [15], for example, proposed a context-based recognition framework that integrates information from human participants engaged in a conversation to improve visual gesture recognition.

3. DETECTING ENGAGEMENT WITH A GAME COMPANION

3.1 Interaction scenario

The interaction scenario consists of a social robot, the iCat [20], that acts as the opponent of a human player (i.e., a child) in a chess match played on an electronic chessboard (Figure 1). The iCat generates affective behaviour that is influenced by the state of the game, and is reflected on the robot's facial expressions. By interpreting the affective reactions displayed by the iCat, children may acquire additional information to better understand the game.

The robot's affective system includes two main components: mood and affective reactions. Mood is a background affective state that is always present with low intensity. Affective reactions are displayed after every move played by the user and have the duration of approximately 3 seconds. The reactions are computed based on an anticipatory system: the iCat generates an expectation of the user's next move, and based on the mismatch between this expectation and the evaluation of the move that the user actually makes, one out of nine "sensations" is activated and a corresponding affective facial expression is displayed by the iCat [2] [3].

Sensation	Animation
Stronger Reward	Excited
Expected Reward	Confirm
Weaker Reward	Happy
Unexpected Reward	Arrogant
Negligible	Think
Unexpected Punishment	Shocked
Weaker Punishment	Apologise
Expected Punishment	Angry
Stronger Punishment	Scared

Table 1: Mapping between the iCat's sensations and the displayed facial animations.

Table 1 shows the mapping between the nine sensations and the corresponding facial animations displayed by the robot. These animations are part of the iCat software library (see [6] for an investigation on how users perceive the iCat's affective expressions). The choice of the mapping between iCat sensations and facial expressions was based on the meaning of the sensations. For example, "stronger reward" means that the iCat experienced a much better "reward" than it was expecting, and therefore the corresponding displayed animation is "excitement".

3.2 User engagement

It was reasoned that user engagement with the iCat would be a key consideration in this specific interaction scenario. Our rationale is that the detection of user engagement is an important requirement for a companion to behave in a more social and empathic manner and crucial to establishing a long-term interaction with the user. The user's engagement with the iCat was chosen to describe the level of social interaction established between them. Engagement has been defined as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and continuing the interaction" [18]. We regard engagement with the iCat as being characterised by an affective and attention component (see [17] for a similar view on engagement in human-agent interaction): the user is considered as engaged with the iCat if she is willing to interact and maintain the interaction with it. This relates, for example, to the amount of time the user looks at the iCat, regardless of whether the iCat is talking or displaying an affective reaction or not. We consider engagement with the iCat to also imply the existence of an active component in the behaviour displayed by the user, for example when the user is "aroused socially" and shows a high action tendency. Finally, engagement with the iCat may include, although not necessarily, the expression of a positive feeling by the user.

3.3 Framework for the modelling of user engagement

Our framework for the modelling of user engagement is designed so as to include both causes and effects of engagement. In our interaction scenario, the user plays chess while interacting with the iCat robot. It is expected that the user's engagement with the companion is both influenced by the task the user is involved in and the social interaction with the iCat. It is also expected that the level of engagement of the user with the iCat affects the generation of expressive non-verbal behaviour in the user. Therefore, in our scenario user engagement is modelled using a number of task and social interaction-based features:

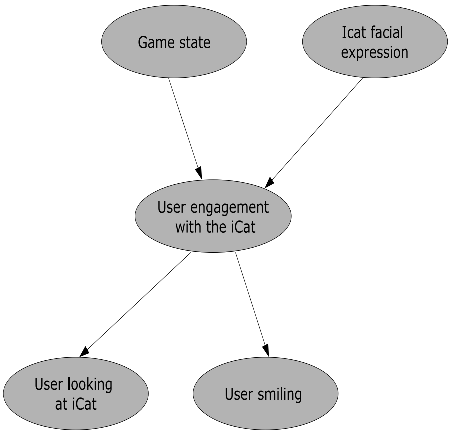

Figure 2: Cause-effect relationships between user engagement with the iCat and task and social interaction-based features in our scenario.

User behaviour: Expressive non-verbal behaviour has been previously linked to user engagement with a companion, namely eye gaze and smiles [17] [4]. Thus, in our framework, we consider as effects of user engagement with the iCat the following features:

- User looking at the iCat
- User smiling

Contextual information: User engagement with the iCat can be modelled by several sources of contextual information. In fact, affect can be the result of several events, situations and user characteristics. Previous findings suggested that user engagement with the iCat correlate with the game state and the presence of affective facial expressions displayed by the iCat [5]. Thus, in our scenario, we model user engagement with two different levels of contextual information:

- Game state
- iCat displaying an affective reaction

A Bayesian network is used to represent user engagement, task and social interaction-based features, and their probabilistic relationships. Figure 2 shows the identified cause-effect relationships in our interaction scenario.

4. EXPERIMENTAL RESULTS

4.1 Data collection

In this Section we explain how we collected the experimental data used to train and evaluate our model. An experiment was performed in two different locations, a primary school where once a week children have two hours of chess lessons, and a chess club where children are more experienced and practice chess more than two hours per week. 8 children (6 males and 2 females, average age 8.5) took part in this study. Every participant played two different exercises, one with low and one with medium difficulty. By using more than one level of difficulty we expected to elicit different types of expressive behaviours in the children. The exercises were proposed by a chess instructor who was familiar with each student's chess skills. The children's goal in the experiment was to try to win against the iCat. In each game the robot begins the interaction by inviting the user to play. The user is always in control of the white pieces and is always the first to play. After each move is made by the user, the iCat asks her to play its move as it does not have any grasping mechanism with which to move the chess pieces itself. Each interaction ends when the user completes both exercises by winning, losing or withdrawing.

All the interactions were recorded by three video cameras: one capturing the children's face, one capturing a lateral view of the children and one capturing the iCat's behaviour. The videos recorded with the frontal camera were annotated in terms of user engagement with the iCat by three annotators. The annotation was based on the behaviour displayed by the children and the situation of the game. A number of video segments were identified starting from the 16 collected videos (two for each participant) using ANVIL, a free video annotation tool [13]. Each video segment had a duration of approximately 7 seconds. After agreeing on the meaning of each label to describe the user's engagement with the iCat, annotators could choose one out of three options: *engaged with the iCat*, *not engaged with the iCat* and *cannot say.* Each annotator associated labels with each video segment working separately. The results of the three individual annotation processes from each annotator were then compared for each video segment: a label was selected to describe the level of engagement of the user in a video segment when it was chosen by two or three of the annotators. In case each of the annotators chose a different label, the video segment was labelled as *cannot say.* After the annotation process, 96 video segments (12 for each participant, 48 labelled as *engaged with the iCat* and 48 as *not engaged with the iCat*) were selected as the samples to be used to train and test our model.

For each exercise played by the children, a log file was also automatically created. After each move played by the user, a new entry is written in the log. Each entry contains different types of information, including the time since the beginning of the interaction (for synchronisation purposes with the video data), the game state and the iCat's sensations. From the data saved in the log files the contextual information employed in our model were derived.

4.2 Task and social interaction-based features

A number of task and social interaction-based features were computed for each of the selected 96 video segments.

4.2.1 Non-verbal behaviour

As described in Section 3.3, eye gaze and smiles of the user were considered as indicators of the user's engagement with the iCat. Figure 1 shows some examples of expressive nonverbal behaviour displayed by the user. Two coders annotated the portions of video segments in which these behaviours were displayed by the children. For each video segment each behaviour was assigned a value. This value was computed as the average number of frames over which a specific behaviour was displayed in the video segment.

4.2.2 Contextual features

Two levels of contextual information were used to model the user's engagement with the iCat: the game state and the effect of the display of an affective facial expression by the iCat.

Game state: The game state represents the condition of advantage/disadvantage of the user in the game. The more the value of the game state is positive, the more the user is in a condition of advantage with respect to the iCat, the more it is negative, the more the iCat is winning. To compute the game state value we used the chess engine from Tom Kerrigan's Simple Chess Program (TSCP) [1]. The term chess engine refers to a chess playing system that does not have a graphical interface, since it is only the "thinking" part of a chess program. Chess programs consider chess moves as a game tree. From the current board position, the Minimax algorithm [19] generates a tree with all the possible moves, taking into account the best moves played by the opponent. This evaluation continues until it reaches a final "leaf" position, which is evaluated and returned. To evaluate the end nodes TSCP uses a simple evaluation function that takes into account four distinct values: (1) material score, by simply summing the piece types multiplied by piece values (e.g. a queen is worth 900, whereas a pawn is worth 100); (2) pawn structure, which considers the placement of the pawns in the chessboard (e.g., by evaluating passed pawns, double pawns, etc.); (3) piece scoring, that evaluates the placement of each piece (e.g., a queen on the middle of the board is more effective than a queen on the corner); (4) king safety, which considers the pawn shelter around the king. This evaluation function returns the sum of these four aspects; values range between -3000 and 3000. For example, 3000 is a checkmate position, 100 means a pawn up in advantage and 0 a neutral position.

iCat displaying an affective reaction: This is a feature related to the behaviour displayed by the iCat during the game and the interaction with the user. Each affective facial expression is a direct consequence of the situation of the game and is the main channel through which the iCat can communicate an affective message to the user, hence interact socially with her.

We are interested in considering the effect that the display of an affective reaction may have on the user's engagement with the iCat. For this purpose, a metric that takes into consideration the temporal interval the engagement of the user may be affected by the iCat's facial behaviour was defined. During the game, as soon as the user makes a move, the iCat looks at the chessboard and then generates an affective facial expression. Each of these two animations last approximately 2.5 seconds each. Given this information, we

[1] http://www.tckerrigan.com/Chess/TSCP

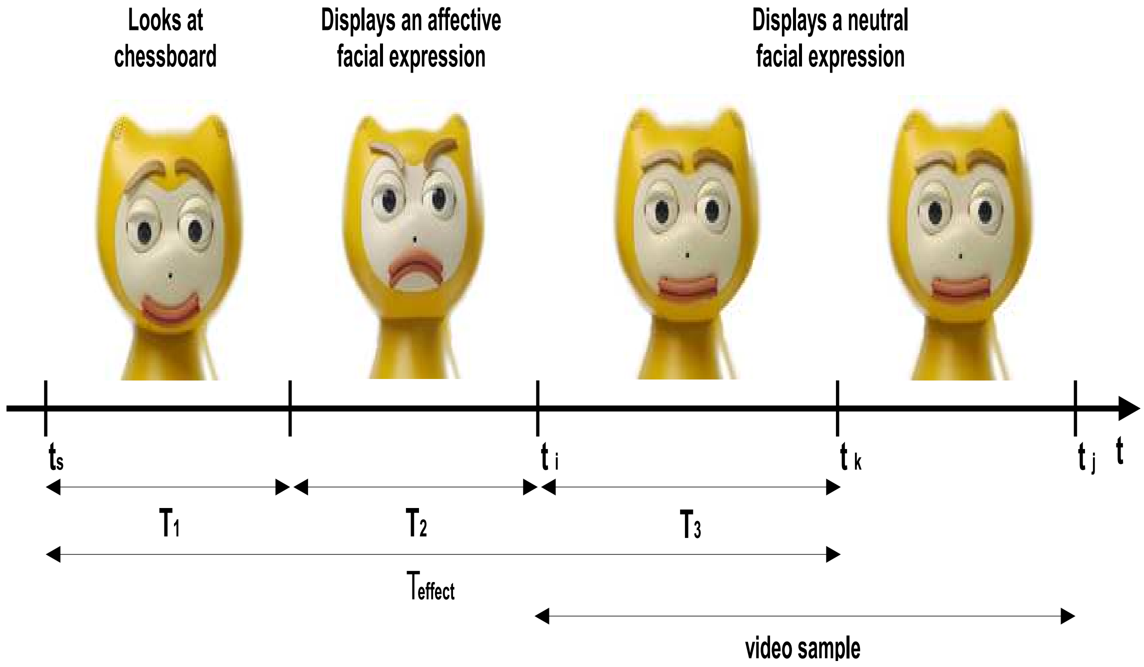
Looks at chessboard
Displays an affective facial expression
Displays a neutral facial expression
ts
ti
tk
tj
t
T1
T2
T3
Teffect
video sample

Figure 3: Effect of the iCat's affective reaction on the user's engagement. At time t_s the user makes a move. Given a video sample in our corpus ranging from time t_i to time t_j and given T_{effect}, the user's engagement with the iCat in that sample is considered as affected by the iCat's affective reaction from time t_i to time t_k.

regard the temporal interval in which the user can be affected by the iCat's affective reaction as formed by three main components (Equation 1):

$$T_{effect} = T_1 + T_2 + T_3 \qquad (1)$$

where T_1 is the time taken to the iCat to look at the chessboard after the user's move, T_2 is the duration of the affective facial expression displayed by the iCat, and T_3 is the duration of the effect of the iCat's expression on the user after the end of the displayed animation. We hypothesise that the effect of the affective reaction displayed by the iCat on the user's engagement begins as soon as the iCat starts looking at the chessboard. T_3 was chosen so that T_{effect} would be smaller than the minimum distance observed in our video corpus between two subsequent facial expressions displayed by the iCat, in order to keep differentiated the effects of two different affective reactions. Given that this distance is approximately of 12 seconds, T_3 was assigned a value so that T_{effect} would be smaller than 12 seconds: in our case we set T_3 to 3 seconds, so that T_{effect} is 8 seconds. We regard 8 seconds as a reasonable temporal interval where we can consider an effect on the user's engagement, as it is more likely to observe a change in the user's behaviour when or immediately after the iCat displays a facial expression.

When the user makes her last move (i.e., when the user or the iCat wins, loses or draws), the iCat, after looking at the chessboard, generates an affective facial expression that lasts more than those displayed during the game (i.e., approximately 4 seconds), hence in these circumstances we consider $T_{effect} = 9.5$ seconds.

Given this temporal interval, each video segment of the corpus was assigned a value that represents how many frames, on average, the user's behaviour in that video segment has been affected by the iCat's facial behaviour. Figure 3 shows how the effect of the iCat's affective reaction is taken into account.

4.2.3 Multimodal fusion

The multimodal fusion step aggregates the user's behavioural features and the contextual information to improve the recognition of the user's engagement with the iCat. For each sample of our corpus, information about the eye gaze and smiles of the user, the game state and the effect of the affective reaction displayed by the iCat were concatenated: these vectors of fused information are used as inputs for our Bayesian classifier for the prediction of the user's engagement with the iCat.

4.3 Evaluation

4.3.1 Methodology

A Bayesian network was used to model user engagement and its probabilistic relationship with the task and social interaction-based features. Experiments were performed to discriminate the user's level of engagement with the iCat. 48 samples labelled as *engaged with the iCat* and 48 as *not engaged with the iCat* were used in this evaluation phase.

Recognition approach	**Recognition rate**	**ROC area**
Non-verbal behaviour	93.75%	0.95
Contextual information	78.13%	0.78
Multimodal	94.79%	0.96

Table 2: Recognition rates and ROC area values for the *user behaviour only*, the *contextual information only* and the *multimodal* approach (non-verbal behaviour + contextual information).

In order to train our model, a "leave-one-subject-out" cross-validation was performed: the data was divided into 8 different subsets, each of them consisting of 12 samples of one of the sujects. At each step of the process 84 samples (corresponding to 7 out of 8 subjects) were used for training and 12 samples (i.e., 1 out of 8 subjects) for test. This means that at each step samples of the same subject are not both in the training and the test set, thus allowing our method to perform in a subject-independent way. To evaluate the model this process was repeated 8 times, so that each time the samples of one of the subjects were used as the test set.

4.3.2 Results

The experimental evaluation assessed the performance of our multimodal framework using task and social interaction-based features and compared it with the performance of a classifier based solely on user non-verbal features and one based solely on contextual information. The metrics used to assess the performance of the classifiers are the recognition rates and the area of the ROC (Receiver Operating Characteristic) curve. Table 2 groups recognition rates and ROC area values for the three approaches. Results show that user engagement with the iCat is well discriminated in all three cases, with the best performance achieved by the multimodal classifier (94.79% vs. 93.75% achieved by the classifier based on the user's non-verbal behaviour and 78.13% by the classifier based on contextual information). This shows that the multimodal integration of task and social interaction-based features improves the recognition of user engagement with the iCat with respect to when single channels of information (non-verbal behaviour of the user and contextual information) are used. Results also show that the classifier based on the non-verbal behaviour displayed by the user is more successful than the classifier trained with the contextual information.

5. CONCLUSION

In this paper we modelled a naturalistic scenario in which children play chess with a game companion. Some of the challenges in affect recognition research were first addressed from a design perspective: the affective state and features that the game companion should be sensitive to were identified by taking into account what valuable information could be extracted from the specific scenario of interaction.

A person-independent Bayesian approach for the modelling of the user engagement with a game companion was presented. Our framework models both causes and effects of engagement: features related to the user's non-verbal behaviour, the task and the companion's behaviour are identified to predict the user's level of engagement with the companion. In particular, results show that our approach based on the multimodal fusion of task and social interaction-based features outperforms those based solely on non-verbal behaviour or contextual information (94.79 % vs. 93.75% and 78.13%). Furthermore, the classifier based on the non-verbal behaviour displayed by the user proved more successful than the classifier trained with the contextual information. Nevertheless, these results show that contextual information could be successfully used to predict the user's level of engagement with the iCat during a chess game and represent a valuable resource in case of noisy or missing data from the vision channel, which is not unlikely to happen under some real-world conditions.

6. REFERENCES

[1] Anonymised.

[2] Anonymised.

[3] Anonymised.

[4] Anonymised.

[5] Anonymised.

[6] C. Bartneck, J. Reichenbach, and A. Breemen. In your face, robot! The influence of a character's embodiment on how users perceive its emotional expressions. In *Proceedings of the Design and Emotion Conference 2004*, pages 32–51, Ankara, Turkey, 2004.

[7] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, July 2003.

[8] G. Castellano and P. W. McOwan. Analysis of affective cues in human-robot interaction: A multi-level approach. In *10th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 258–261, London, 2009.

[9] C. Conati and H. Maclaren. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, January 2009.

[10] K. Dautenhahn. Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.

[11] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[12] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *ACM International Conference on Multimedia*, pages 677–682, 2005.

[13] M. Kipp. Spatiotemporal coding in ANVIL. In E. L. R. A. (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

[14] L. Malta, C. Miyajima, and K. Takeda. Multimodal estimation of a driver's affective state. In *Workshop on Affective Interaction in Natural Environments (AFFINE), ACM International Conference on Multimodal Interfaces (ICMI'08)*, Chania, Crete, Greece, 2008.

[15] L.-P. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *ACM International Conference on Multimodal*

Interfaces (ICMI'08), pages 181–188, Chania, Crete, Greece, 2008.

[16] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, July 1988.

[17] C. Peters, S. Asteriadis, K. Karpouzis, and E. de Sevin. Towards a real-time gaze-based shared attention for a virtual agent. In *Workshop on Affective Interaction in Natural Environments (AFFINE), ACM International Conference on Multimodal Interfaces (ICMI'08)*, Chania, Crete, Greece, 2008.

[18] I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication.* Weidler, Berlin, 2007.

[19] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Pearson Education, 2003.

[20] A. van Breemen, X. Yan, and B. Meerbeek. iCat: An animated user-interface robot with personality. In *AAMAS '05: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 143–144, New York, NY, USA, 2005. ACM.

[21] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.