International Journal of Humanoid Robotics
Vol. 10, No. 1 (2013) 1350010 (23 pages)

DOI: 10.1142/S0219843613500102

World Scientific
www.worldscientific.com

MULTIMODAL AFFECT MODELING AND RECOGNITION FOR EMPATHIC ROBOT COMPANIONS

GINEVRA CASTELLANO

School of Electronic, Electrical and Computer Engineering,
University of Birmingham, Birmingham,
United Kingdom
g.castellano@bham.ac.uk

IOLANDA LEITE, ANDRÉ PEREIRA,
CARLOS MARTINHO and ANA PAIVA

INESC-ID and Instituto Superior Técnico,
Technical University of Lisbon,
Porto Salvo, Portugal

PETER W. MCOWAN

School of Electronic Engineering and Computer Science,
Queen Mary University of London, London,
United Kingdom

Received 5 July 2012
Accepted 21 January 2013
Published 2 April 2013

Affect recognition for socially perceptive robots relies on representative data. While many of the existing affective corpora and databases contain posed and decontextualized affective expressions, affect resources for designing an affect recognition system in naturalistic human−robot interaction (HRI) must include context-rich expressions that emerge in the same scenario of the final application. In this paper, we propose a context-based approach to the collection and modeling of representative data for building an affect-sensitive robotic game companion. To illustrate our approach we present the key features of the Inter-ACT (INTEracting with Robots−Affect Context Task) corpus, an affective and contextually rich multimodal video corpus containing affective expressions of children playing chess with an iCat robot. We show how this corpus can be successfully used to train a context-sensitive affect recognition system (a valence detector) for a robotic game companion. Finally, we demonstrate how the integration of the affect recognition system in a modular platform for adaptive HRI makes the interaction with the robot more engaging.

Keywords: Affect recognition; non-verbal behavior; context-sensitivity; human−robot interaction; social robotics.

1. Introduction

As robots are increasingly being viewed as social entities to be integrated in our daily lives,[1] providing them with social perceptive abilities seems a necessary requirement for enabling more natural interaction with human users.[2] For example, affect sensitivity, i.e. the ability to recognize people's affective states and expressions, is of the utmost importance for a robot to be able to display socially intelligent behavior,[3] a key requirement for sustaining long-term interactions with humans.[4]

While affect recognition has been extensively addressed in the literature,[5] many issues related to the design of a module for affect recognition to be integrated in a human–robot interaction (HRI) framework still have to be investigated.

The design of an affect recognition system for socially perceptive robots, such as robotic companions, requires representative data. Many of the existing corpora and databases of affective expressions contain posed data collected in scenarios which differ from that of the final application.[5] These often include portrayals of prototypical emotions expressed by adults rather than application-dependent states. Moreover, while many of the most recent databases contain multimodal data, the availability of contextual information is still not frequent.

Nevertheless, naturalistic HRI requires affect recognition systems to be trained and validated with contextualized affective expressions, i.e. expressions that emerge in the same interaction scenario of the final application.[4,6] In addition, representative data for automatic inference of the user's affect in HRI should include not only information about the user's behavior, but also information about the task that the user and the robot are involved in and the behavior generated by the robot. In fact, especially in face-to-face HRI, the robot and the user mutually influence each other in a continuous cause and effect cycle: the behavior of the robot may elicit a response from the user and, similarly, the behavior and actions of the latter may trigger the generation of an appropriate response from the robot.[2]

In this paper, we propose an approach for the collection of naturalistic affect resources to build an affect recognition system for a robotic game companion for young children. The paper is divided in three main parts. First, we propose a methodology for the collection and modeling of context-sensitive affect data in a naturalistic, adaptive HRI scenario. To illustrate our approach, we discuss the key features of the Inter-ACT (INTEracting with Robots–Affect Context Task) corpus,[7] an affective and contextually rich multimodal video corpus containing affective expressions of children playing chess with an iCat robot.[8,9] The Inter-ACT corpus contains videos from multiple view-points that allow for the interaction to be captured from different perspectives and includes synchronized contextual information about the game and the iCat's behavior. The corpus is intended to be a comprehensive repository of naturalistic and contextualized, task-dependent data in an educational game scenario with a robot companion. It is unique in its genre, as it includes contextulized affective expressions of children, rather than adults, suitable to train an affect recognition system for use in child–robot interaction. This is an

important contribution of our work, as children's expressions differ from those of adults and their automatic analysis presents more challenges. Secondly, we show how this corpus can be successfully used to train a context-sensitive affect recognition system (a valence detector) for a robotic game companion. Finally, we demonstrate how the integration of the affect recognition system in a modular platform for adaptive HRI makes the interaction with the robot more engaging.

The paper is organized as follows. The next section discusses the challenges in collecting affective corpora for HRI applications and provides an overview of previous work on automatic affect recognition in HRI and on the use of context in affect recognition frameworks. Section 3 provides an overview of the showcased HRI scenario, while Sec. 4 describes the data collection and annotation process of the Inter-ACT corpus, and Sec. 5 the traning and evaluation of the valence detector. Section 6 provides an overview of the architecture of the robotic game companion and presents methodology and results of an evaluation experiment conducted in a primary school. Finally, Sec. 7 summarizes and discusses the main results.

2. Background

2.1. *Affective corpora: Requirements for HRI*

The design of affect-sensitive robotic companions requires research on affective corpora to be taken beyond the state of the art. In the following, we review some of the challenges in the collection of representative data for building an affect recognition system for a robotic companion.

(1) **Scenario-related states**
Robotic companions require the design of affect recognition systems with the ability to go beyond the recognition of prototypical emotions, and to allow for more variegated affective signals conveying more subtle, scenario-related states such as, for example, boredom, frustration, interest, willingness to interact, engagement, etc., to be captured.[4]

(2) **Spontaneous versus acted expressions**
While examples of naturalistic databases are gradually increasing in the literature,[10–12] the design of many existing affect recognition systems was largely based on databases of acted affective expressions.[5] While acted affective expressions, contrary to spontaneous expressions, can be defined precisely, allow for the recording of several affective expressions for the same individual, and can be characterized by very high quality, they often reflect stereotypes and exaggerated expressions, not genuine affective states, and they are often decontextualized.[13]

(3) **Multimodal affective expressions**
Another important issue for affect-sensitive artificial companions is the need for a multimodal affect recognition system. It is expected that a companion is endowed with the ability to analyze different types of affective expressions,

depending on the specific interaction scenario. On the other hand, fusing different affective cues can allow for a better understanding of the affective message communicated by the user to be achieved. While unimodal systems (mainly based on facial expression or speech analysis) have been deeply investigated, studies taking into account the multimodal nature of the affective communication process are still not numerous.[5]

(4) **Context-rich descriptions**
HRI applications require systems trained with data including contextual descriptions synchronized with other modalities.[14] For example, information about the user, their artificial interactant, the environment, the task they are involved in, etc., becomes necessary to complement affective behavioral data.

(5) **Contextualized affective expressions**
Affect expression depends on context. Most of the affective video corpora and databases available in the literature contain expressions recorded in contexts that are not specific to a particular application.[5] An exception is represented by the CAL database by Afzal and Robinson,[6] which contains affective expressions collected in a computer-based learning environment. HRI requires contextualized affective user expressions for system training and validation, i.e. expressions collected in the same scenario of the final application.

2.2. *Affect recognition in HRI*

Recent advances in automatic affect recognition show that human affective states can successfully be predicted using a variety of affective cues in several HRI applications.[5]

A few computational approaches for affect recognition have also been proposed in the HRI and social robotics fields. Kulic and Croft, for example, developed an HMM-based system capable of estimating valence and arousal elicited by viewing robot motions using physiological data such as heart rate, skin conductance and corrugator muscle activity.[15] Rich and colleagues[16] proposed an approach for the automatic recognition of engagement between a human user and a humanoid robot. Their approach is based on the recognition of connection events such as directed gaze, mutual facial gaze, conversational adjacency pairs and backchannels. Liu *et al.*[17] developed an affect inference mechanism based on physiological data for real-time detection of affective states of children with autism spectrum disorder interacting with a robot. Mower *et al.* presented an approach to detect user engagement with a robot using physiological data.[18]

Previous work in the scenario investigated in this paper showed that children's engagement with a robotic game companion can be successfully inferred using task and social interaction-based features[19] and expressive body postures data.[20]

2.3. *Context-sensitive affect recognition*

Of late there has been an increasing interest towards the role of context in research on multimodal interfaces and multimedia applications.[21]

Context has been identified as a key requirement for meaningful content interpretation in vision-based recognition systems.[22] Morency *et al.*,[23] for example, proposed a context-based recognition framework that integrates information from human participants engaged in a conversation to improve visual gesture recognition. They proposed the idea of encoding dictionary, a technique for contextual feature representation that models different relationships between a contextual feature and visual gestures.

On the other hand, affective states in HRI can be influenced by many different factors. Examples include the user's personality, gender, preferences, underlying mood, history and goals, the task, the presence of other people, the events unfolding in the environment, the type of behavior displayed by the interactant, etc. All this can be referred to as context. This suggests that context can be used as an additional source of data to improve the performance of an affect recognition system, when other modalities are not sufficient or lead to a non-meaningful interpretation.

While some efforts have been reported in the literature, only a limited number of studies have addressed the problem of context-sensitive affect recognition. Kapoor *et al.*,[14] for example, 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). Peters *et al.*[24] modeled the user's interest and engagement with a virtual agent displaying shared attention behavior, by using contextualized eye gaze and head direction information. Sabourin and colleagues[25] investigated the automatic prediction of learner affect using dynamic Bayesian networks modeling personality attributes, appraisal variables and student activity in a learning environment. Malta *et al.*[26] proposed a system for the multimodal estimation of a driver's irritation that exploits information about the driving context. Martinez and Yannakakis[27] proposed a method for the fusion of physiological signals and game-related information for automatic affect recognition in a game scenario. Their approach uses frequent sequence mining to extract sequential features that combine events across different user input modalities.

Previous work conducted by the authors in the iCat scenario showcased in this paper showed that game context can be used to discriminate user affect[28] and that children's engagement with the robot can be recognized using a combination of task and social interaction-based features.[19]

3. Scenario

The interaction scenario consists of a social robot, the iCat,[8] that acts as the opponent of a human player in a chess game. In this scenario, the iCat robot plays the role of a game companion for children using an electronic chessboard. The user sits in front of the chessboard, which is placed between the user and the iCat robot (Fig. 1).

While playing with the iCat, children receive feedback on their moves through the robot's facial expressions, which are generated by an affective system influenced

Fig. 1. (Color online) User interacting with iCat in a primary school.

by the state of the game, and confirmation signals, such as small utterances and nodding gestures. The iCat's affective system is self-oriented, which means that when the user makes a good move, the iCat displays a negative facial expression, and when the user makes a bad move, it expresses positive reactions. By interpreting the affective reactions displayed by the iCat, children can acquire additional information to better understand the game.

Previous studies in this scenario showed that, after repeated interactions with the robot, children started realizing that the robot's behavior did not take into account the emotions they experienced, and social presence decreased over time.[29] To overcome these limitations, in this paper, we describe the collection and modeling of affective representative data needed to train an affect recognition system for the iCat robot. We anticipate that affect sensing will allow the robot to adapt to the user's behavior in an appropriate way throughout the game, thus leading to the establishment of interactions that are more engaging and more successful over extended periods of time, which is an important requirement for companionship.

4. The Inter-ACT Corpus

The Inter-ACT corpus consists of 156 six-second "thin-slices" of the interaction between children and an iCat robot that play chess. Each slice of the interaction is described by multimodal data: a frontal video capturing the face and the upper body of the children, a lateral video capturing their lateral posture and full-body movements, a video capturing the iCat, and a set of synchronised contextual features that describe the events of the game and the behavior displayed by the robot. Videos and contextual data together provide a comprehensive description of the ongoing interaction.

4.1. *Data collection*

4.1.1. *Subjects, scenario and setup*

The data collection procedure was performed in two different locations, a primary school where every week children have 2 h of chess lessons, and a chess club where children are more experienced and practice chess more frequently. Eight children (six male and two female, average age 8.5) took part in the data collection procedure.

Every participant was asked to play two different exercises, one with low and one with medium difficulty, chosen by a chess instructor who was familiar with each student's chess skills. By adopting two different levels of difficulty we expected the children to display a broader range of expressive behaviors.

In each exercise the robot begins the interaction by inviting the user to play. After each move is made by the user, the iCat asks them to make its move as it does not have any grasping mechanism to move the chess pieces by itself. Each interaction ends when the user completes both exercises by winning, losing or withdrawing. The duration of each exercise varied depending on the specific participant, with exercises lasting up to 15 min.

All the exercises were recorded with four video cameras: two capturing the frontal view, one the lateral view of the children and one the iCat. To capture the frontal view we used a firewire camera (15 fps, 1024×768 spatial resolution) and a DV camera (25 fps, 720×576 spatial resolution). For the lateral view, we used a DV camera (25 fps, 720×576 spatial resolution). A standard 25 fps webcam was used to capture the behavior displayed by the iCat. Figure 2 shows some examples of frames from the frontal, lateral and iCat's view.

Each video of the Inter-ACT corpus was segmented by an expert coder. Segmentation was performed starting from the frontal videos in order to include coherent samples of behavior: the corpus includes a balanced number of samples displaying a range of different expressions.

4.1.2. *Contextual information*

During the video recordings of our corpus, contextual information about the game and the iCat's behavior was logged in real-time via a game engine built on top of the chess engine from Tom Kerrigan's Simple Chess Program (TSCP).[a] After each move made by the user on the electronic chessboard, the chess evaluation function returns a new value, updated according to the current state of the game. Based on the history of evaluation values, the game engine automatically extracts contextual features.

Every log entry contains a timestamp to enable the synchronization with the corresponding video files containing the users' non-verbal behaviors. There is one log file associated with each exercise, which contains an entry for every move played by the user, the consequent iCat's move and a set of contextual information retrieved

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

Fig. 2. (Color online) Examples of frames from the frontal, the lateral and the iCat's view in the Inter-ACT corpus.

after each move played by the user. Thus each video of the Inter-ACT corpus includes the following synchronized contextual information.

Game state. A value that represents the condition of advantage/disadvantage of the user in the game. This value is obtained by the same chess evaluation function that the iCat uses to plan its own moves, but from the user's perspective. 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 and vice versa.

Game evolution. The difference between the current and the previous value of the game state. A positive value for game evolution indicates that the user is improving in the game, while a negative value means that the user's condition is getting worse with respect to the previous move.

Captured pieces. If there were any captured pieces either by the user or by the iCat, this value indicates the type of piece that was taken.

Game phase. A value that represents the phase of the game (*game starts, after user's move, after iCat's move, draw, user gives up, iCat gives up, user wins, user loses*).

Level of difficulty. This value indicates the level of difficulty of the chess exercise.

User emotivector. It refers to the result of the mismatch between expectation and actual outcome of the user's progress in the game. After each move made by

the user, the chess evaluation function returns a new value, updated according to the current state of the game. The robot's emotivector system[30] is an anticipatory system that captures this value and, by using the history of evaluation values, computes an expected value for the chess evaluation function associated with the user's moves. Based on the mismatch between the expected value and the actual value, the system generates a set of affective signals describing a sentiment of reward, punishment or neutral effect for the user.[9]

For example, if after three moves in the chess game the user has already captured an iCat's piece, they might be expecting to keep the advantage in the game (i.e. expecting a reward) after the iCat's next move. Therefore, if the iCat makes a move that is worse than the one the user was expecting (e.g. by putting its queen in a very dangerous position), the generated affective signal will be a "stronger reward", which means that the state of the game is better than what the user was probably expecting.

iCat's facial expressions. Based on the affective signals generated by the emotivector system described above, the iCat provides feedback to the user by displaying an affective facial expression.[9] 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.

4.2. *Affect annotation*

The Inter-ACT corpus contains affective labels that describe the user's affect in each "thin-slice" of the interaction.

Off-line analysis of videos recorded during several interactions showed that children display prototypical emotional expressions only occasionally. On the other hand, we believe that the robotic game companion would benefit from the ability to detect states emerging during the game and the social interaction with the iCat robot, such as the valence of the affect experienced during the game and the level of interest towards the robot:

(1) **Valence of affect**
The valence of the affect experienced by the user was chosen to measure the degree to which the user's affect is positive or negative.[31] This dimensional description of affect appears to be adequate for the purpose of describing the overall feeling that the user is experiencing throughout the game and the interaction with the robot.

(2) **Interest towards the robot**
The level of interest towards the robot relates to the amount of time the user pays attention to it.[24] It can be considered as a social emotion, as it provides information about the attitude of the user towards the robot. In human face-to-face interaction, detecting whether someone is interested or not is a very important social capability, which can also be of extreme importance to achieve natural HRI.

Affect annotation was performed by a group of three expert coders. The annotation was based on the frontal videos and was performed by asking the coders to focus on the behavior displayed by the children, without access to any information about the context. For each video the coders were requested to rate the level of valence of affect (*positive, negative* or *neutral*) and the level of interest towards the robot (*high interest, low interest* or *medium interest*). The annotation was based on the definitions of valence of affect and interest towards the robot provided above. The coders were also asked to provide their confidence level (from 1 to 10) for each annotation.

Inter-coder agreement was measured with the Fleiss' kappa statistics: results showed an overall fair to moderate agreement (Fleiss' kappa = 0.37 for valence; Fleiss' kappa = 0.43 for interest towards the robot). A label for valence and interest towards the robot for a specific video of the corpus was selected when at least two out of three coders agreed. In case of disagreement amongst all coders, the confidence level provided by the coders was taken into account to determine the final label to assign to each video.

In this work, we focus on the valence experienced by the user throughout the game. The next section shows how we use the part of the corpus annotated in terms of user valence to train an automatic valence detector to be integrated in our robotic companion.

5. Valence Detector: Design and Evaluation

Previous work in the iCat scenario highlighted that some user expressions are more frequent than others and help discriminate among positive and negative affect.

In terms of user expressions, the behaviors displayed by the children that are mainly affected by the valence of affect are the eye gaze and the smiles: when the affect is positive, children tend to look at the iCat and smile more than when the affect is negative.[4]

As far as contextual information is concerned, previous work highlighted a key role of game state and game evolution to discriminate between positive and negative affect (i.e., the user's affect tends to be more positive than negative when the user is winning or improving in the game).[28]

These findings highlight which are the features that may be more relevant to the target recognition task, thus suggesting to focus the design of the valence detector only on a limited subset of behavioral and contextual features. Therefore, in addition to game state (*GameState*) and game evolution (*GameEvolution*), described in Sec. 4.1.2, we extracted the following user behavioral features:

Smiles. We used faceAPI, a real-time face tracking toolkit from Seeing Machines,[b] to track head movements and salient facial points. Facial landmarks are provided in pixel coordinates frame (2D) or head coordinates frame (3D).

[b] http://www.seeingmachines.com/.

Smile indicators were automatically extracted from portions of the Inter-ACT corpus containing smile behavior in order to build a smile detector. These include (1) geometrical indicators extracted from facial landmarks in pixel coordinates, such as the ratio of the lips bounding box and the ratio of the bounding box defined by the eyebrows and lips corners and (2) lips corners in head coordinates.

In order to capture the local changes of the above indicators, we compute a *local behavior baseline* at each frame by averaging each indicator over the previous N seconds, and we subtract the local baseline to each indicator at each frame. Empirical tests showed good results for $N = 0.27$ s.

A smile detector based on SVMs was trained using 552 sample frames from the Inter-ACT corpus, to classify frames as smiling or not smiling. Selected samples included smiling and not smiling behavior of all eight children. For training and testing the smile detector, we used the LibSVM library[32] and a Radial Basis Function (RBF) kernel. The recognition performance using a "leave-one-subject-out" cross-validation approach is 96.10%.

The smile detector was used to extract the probability of smile for each frame of the videos of Inter-ACT corpus. For each video, the values of the probability of smile were averaged over the whole video duration (6 s) to obtain the final smile feature used to train the valence detector (*AvgProbSmile*).

Eye gaze. We performed manual annotation of eye gaze features. Three binary eye gaze features were extracted for each video, depending on the amount of time the user looked at the iCat during the 6 s of the video: *LookingAtIcatHigh* (when the user looks at the iCat for 3 s or more), *LookingAtIcatMedium* (when the user looks at the iCat for less than 3 s), *NotLookingAtIcat* (when the user does not look at the iCat at all).

An SVM classifier with RBF kernel provided by LibSVM[32] was employed for valence classification experiments, in order to predict three different outputs for valence of affect (*positive, negative* or *neutral*). The classifier was trained and tested with 113 samples (see Table 1 for a summary of the features used in the training and testing phase) of the Inter-ACT corpus. The remaining ones were disregarded due to problematic tracking. This occurs, for example, when children move away from being in a reasonable range of the camera. A "leave-one-subject-out" cross-validation approach was followed during the training and testing phase, leading to a recognition performance of 63% for predicting three different outputs for valence of affect (three labels: *positive*, *negative* or *neutral*).

Table 1. Behavioral and contextual features.

Feature	Category
AvgProbSmile	Behavior
LookingAtIcatHigh	Behavior
LookingAtIcatMedium	Behavior
NotLookingAtIcat	Behavior
GameState	Context
GameEvolution	Context

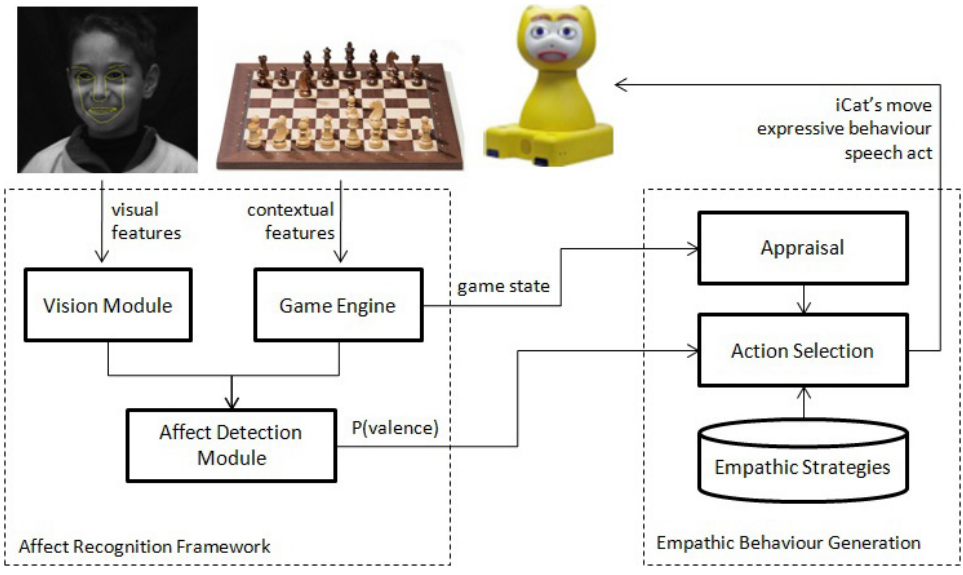

Fig. 3. (Color online) iCat system architecture.

In order to evaluate the effects of the robot's ability to perceive children's affect, we integrated the valence detector into a novel platform for affect-sensitive, adaptive HRI (Fig. 3) and performed a field experiment in a primary school, described in the next section.

6. Experimental Evaluation

In this section, we describe the modular architecture of the iCat system and provide an overview of the field experiment that we conducted and a summary of the results.

6.1. *System architecture*

The platform integrates an array of sensors in a modular client-server architecture that includes a vision module, a game engine, an affect detection module (the valence detector), an appraisal module and an action selection module (Fig. 3). After every move made by the user, the user's affective state is inferred by the affect detection module based on behavioral indicators provided by the vision module and contextual indicators extracted by the game engine. Information about the user's affective state is used to trigger the generation of empathic interventions by the robot.

In the following we describe the modules that compose the iCat's platform:

(1) **Vision module.** A standard Logitech webcam, positioned in front of the user (Fig. 1), captures the non-verbal behavior displayed by the children during the game and the interaction with the robot. The system performs tracking of head movements and salient facial points via faceAPI and extracts information about

users' gaze direction and probability of smile. After a calibration phase, the system estimates the gaze direction of the user based on head direction and rotation data. Furthermore, geometrical facial features extracted from the tracked facial points are used to detect user's smiles using SVMs. Averaged eye gaze and smile features are continuously sent in input to the affect detection module.

(2) **Game engine.** As described in Sec. 4.1.2, the game engine is built on top of the chess engine from TSCP. After each move made by the user on the electronic chessboard, the chess evaluation function returns a new value, updated according to the current state of the game. Based on the history of evaluation values, the game engine automatically extracts game-related contextual features.

(3) **Affect detection module.** The affect detection module consists of the SVM-based valence detector described in Sec. 5. It continuously receives synchronized features from the vision module and the game engine and provides as output probability values for valence (i.e., whether the user is more likely to be experiencing a positive, neutral or negative affect).

(4) **Appraisal module.** This module appraises the situation of the game and provides as output information on the robot's affective state. Information on the robot's affective state triggered by the game are given in input to the action selection module to generate an appropriate affective facial expression.

(5) **Action selection module.** This module generates affective facial expressions as a consequence of the robot's affective state and, when the user is not experiencing a positive affect, selects an empathic strategy for the robot to display.

6.2. *Methodology*

In order to test the iCat system and evaluate the effects of the robot's ability to perceive children's affect, we conducted a field experiment in a real classroom environment.

6.2.1. *Experimental setting*

The study was conducted in a Portuguese elementary school where children play chess two hours per week as part of their school curriculum. The experimental setting comprised the robot, an electronic chessboard, a computer where all the processing takes place and a webcam that captures the children's expressions to be analyzed by the affect detection module (Fig. 1). Two video cameras were also used to record all the interactions with the robot from a frontal and lateral perspective. The setting was installed in the room where children have their weekly chess lessons. The objective was to integrate the robot in the natural environment where children usually play chess.

6.2.2. *Procedure*

A total of 26 children, with ages between 8 and 10 years old, participated in the experiment. Participants were randomly assigned to two different conditions,

corresponding to two different parametrizations of the robot's behavior:

(1) **Neutral.** The robot does not exhibit any empathic behavior. It simply comments the moves in a neutral way (e.g., "you played well", "bad move", etc.).
(2) **Empathic.** When the user is not experiencing a positive affect, the iCat randomly selects one of the available empathic strategies (i.e., providing encouraging comments, letting the user play again if their move was not good, offering help and intentionally playing a bad move[33]).

In addition to differences in the empathic behavior, the robot's affective feedback is also different in the two conditions. While in the empathic condition the robot's affective behavior is user-oriented (i.e., the robot shows happiness if the user makes good moves and sadness if the user makes bad moves), in the neutral condition the robot's behavior is self-oriented (i.e., the robot shows sadness if the user makes good moves, etc.). This is reflected in the robot's facial expressions displayed after every user's move. Our final sample consisted of 13 children in the empathic condition and 13 in the neutral condition.

Participants were guided to the room where the setting was installed and were instructed to sit in front of the robot and play a chess exercise. The exercise was the same for all the participants, and was suggested by the chess instructor so that the difficulty was appropriate for the children. Two experimenters were in the room controlling the experiment and observing the interaction. Each child played, on average, 15 min with the robot. After that period, depending on the state of the game, the iCat either gave up (if it was in disadvantage) or proposed a draw (if the child was loosing or if none of the players had advantage), arguing that it had to play with another user.

6.2.3. *Data collection*

During the experiment for each human−robot interaction we automatically collected a set of questionnaires reporting the children's levels of engagement with the robot, perceived help and self-validation, for a total of 26 ratings for each variable.

After playing with the robot, participants were taken to another room where they were asked to fill in a questionnaire. Note that children from the age group taken into consideration are sufficiently developed to answer questions with some consistency.[34]

Children were asked to rate their level of engagement with the robot, perceived help and self-validation throughout the interaction using a 5-point Likert scale, where 1 meant "totally disagree" and 5 meant "totally agree". Engagement is a metric that has been extensively used both in HRI and human-agent interaction and has been defined from several perspectives.[35,36] For example, Sidner *et al.* defined engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection".[36] The questionnaire regarding engagement in our study is based on the questions used by Sidner *et al.* to evaluate users' responses towards a robot capable of using social capabilities to attract users' attention. As in

the questionnaire used by Leite *et al.*[37] help measures how the robot provided guidance and other forms of aid to the users, whereas self-validation measures the degree of reassuring, encouraging and helping the other to maintain a positive self-image.

6.3. *Results*

We performed a statistical analysis on the average ratings of engagement, help and self-validation collected from the questionnaires. We verified that the distribution of the data was not normal by applying the Kolmogorov−Smirnov test, so non-parametric tests were applied.

Mann−Whitney tests were performed to assess the significance of the differences in the ratings of engagement, help and self-validation observed in the neutral and the empathic conditions (Figs. 4−6). The tests showed that the empathic condition differed significantly from the neutral condition in terms of engagement ($U = 46$ $r = -0.4$ $p < 0.05$), help ($U = 33$ $r = -0.53$ $p < 0.05$) and self-validation ($U = 45$ $r = -0.37$ $p < 0.05$). These results suggest that participants in the empathic condition significantly found the robot more engaging and helpful than participants in the neutral condition and provided significantly higher scores for self-validation.

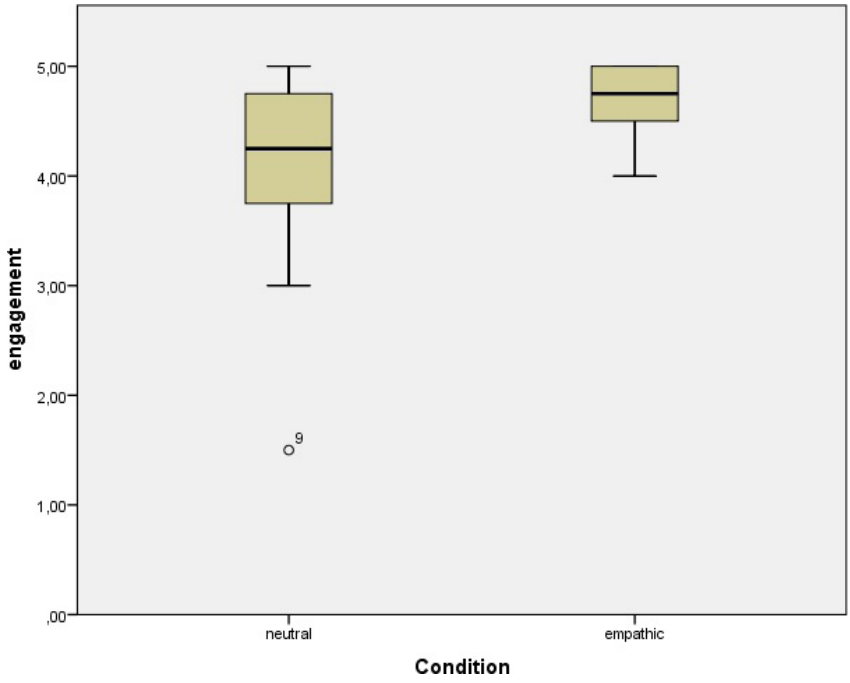

Fig. 4. (Color online) Boxplot charts for "engagement".

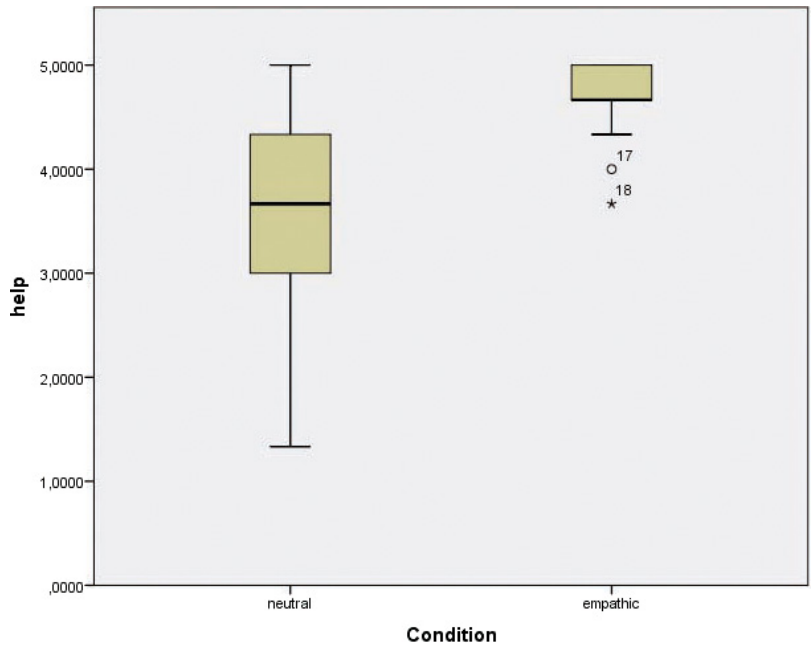

Fig. 5. (Color online) Boxplot charts for "help".

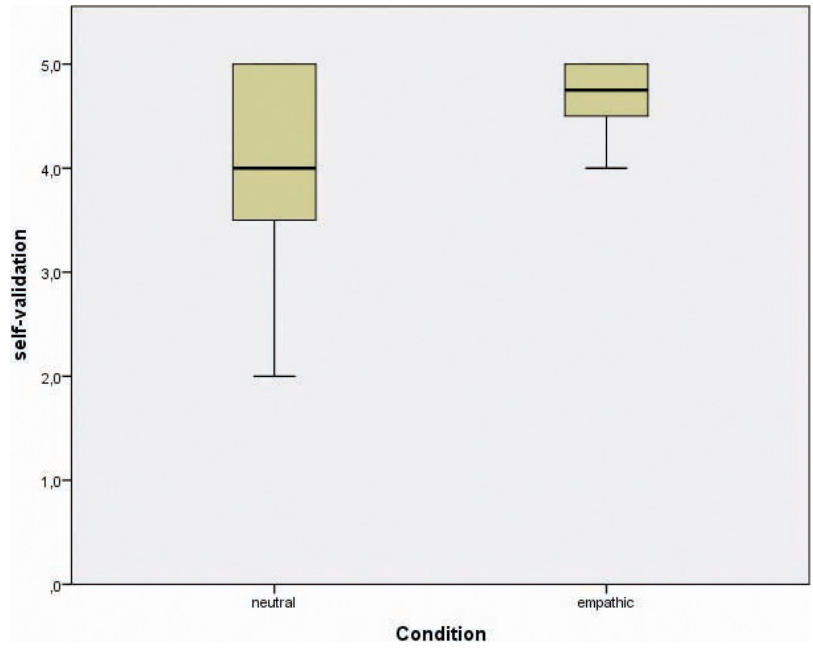

Fig. 6. (Color online) Boxplot charts for "self-validation".

7. Conclusion

In this paper, we proposed a context-based approach to the collection and modeling of affective data in a naturalistic HRI scenario. To describe the proposed approach, we discussed the key characteristics of the Inter-ACT corpus, a multimodal video corpus of affective expressions emerging during the interaction with a robotic game companion. The corpus is a comprehensive repository of visual data and synchronized contextual information designed to train and evaluate an affect recognition system for a robotic game companion.

The design of the Inter-ACT corpus addressed some of the latest challenges in the collection of representative data for affect recognition in naturalistic HRI[38,39]: the corpus contains multimodal, spontaneous affective expressions of young children, scenario-related states and descriptors that carry information about the social and the game context. We demonstrated how this corpus can be successfully used to train a context-sensitive affect recognition system for a robotic game companion that works in real world settings.

The integration of the valence detector into a novel platform for affect-sensitive, adaptive HRI showed how the robot's ability to perceive children's affect has an effect on their perception of the robot. The results showed that children perceived the robot as more engaging and helpful and also provided higher ratings in terms of self-validation.

These findings confirm the results of an ethnographic study on the perception of empathic behavior conducted in the same scenario, which showed that the robot's empathic behavior, generated as a response to the user's affect, affected positively how children perceived the robot.[33] These results show an improvement over previous studies conducted with our robotic companion without the affect sensitivity ability, which showed that the robot's social presence decreased over time.[29] The results provide support for our scenario-centred affect modeling and recognition approach for building a robotic game companion for children, and are confirmed by initial findings of studies conducted over several weeks in a primary school.[40]

Future work will also focus on the integration of interest and engagement[41] detectors in our companion's framework. We anticipate that affect sensing will lead to the establishment of empathic and personalized interactions[42,43] that are more natural and engaging, which is an important requirement for companionship over extended periods of time.[44,45]

Acknowledgments

This work was partly supported by the EU FP7 ICT-215554 project LIREC (LIving with Robots and intEractive Companions) and by national funds through FCT-Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011 and the PIDDAC Program funds.

References

1. F. Tanaka, A. Cicourel and J. R. Movellan, Socialization between toddlers and robots at an early childhood education center, *Proc. National Academy of Science*, Vol. 194, No. 46 (2007), pp. 17954–17958.
2. C. Breazeal, Role of expressive behaviour for robots that learn from people, *Philos. Trans. Royal Soc. B* **364** (2009) 3527–3538.
3. K. Dautenhahn, Socially intelligent robots: Dimensions of human-robot interaction, *Philos. Trans. Royal Soc. B: Biol. Sci.* **362**(1480) (2007) 679–704.
4. G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva and P. McOwan, Affect recognition for interactive companions: Challenges and design in real-world scenarios, *J. Multimodal User Interfaces*, **3**(1–2) (2010) 89–98.
5. Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. on Pattern Analy. Mach. Intell.* **31**(1) (2009) 39–58.
6. S. Afzal and P. Robinson, Natural affect data — collection and annotation in learning context, *Proc. 3rd Int. Conf. on Affective Computing and Intelligent Interaction (ACII 2009)* (IEEE, 2009), pp. 22–28.
7. G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva and P. W. McOwan, Inter-ACT: An affective and contextually rich multimodal video corpus for studying interaction with robots, *Proc. ACM Int. Conf. on Multimedia* (ACM, 2010), pp. 1031–1034.
8. A. van Breemen, X. Yan and B. Meerbeek, iCat: An animated user-interface robot with personality, *AAMAS '05: Proc. Fourth Int. Joint Conf. on Autonomous Agents and Multiagent Systems* (ACM, New York, NY, USA, 2005), pp. 143–144.
9. I. Leite, A. Pereira, C. Martinho and A. Paiva, Are emotional robots more fun to play with? *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE Int. Symposium on (Aug. 2008)*, pp. 77–82.
10. G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schrder, The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent, *IEEE Trans. on Affect. Comput.* **3** (2012) 5–17.
11. M. Pantic, M. Valstar, R. R. and L. Maat, Web-based database for facial expression analysis, *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME '05)* (July 2005), pp. 317–321.
12. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir and K. Karpouzis, The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data, *Proc. 2nd Intl. Conf. on Affective Computing and Intelligent Interaction, ser. ACII '07* (Springer-Verlag, Berlin, Heidelberg, 2007) pp. 488–500.
13. K. Scherer and T. Banziger, On the use of actor portrayals in research on emotional expression, *Blueprint for Affective Computing: A Sourcebook*, eds. K. R. Scherer, T. Bänziger and E. B. Roesch (Oxford University Press, Oxford, England, 2010).
14. A. Kapoor and R. W. Picard, Multimodal affect recognition in learning environments, *Proc. ACM Int. Conf. on Multimedia* (2005), pp. 677–682.
15. D. Kulic and E. A. Croft, Affective state estimation for human-robot interaction, *IEEE Trans. on Robotics* **23**(25) (2007), pp. 991–1000.
16. C. Rich, B. Ponsler, A. Holroyd and C. L. Sidner, Recognizing engagement in human-robot interaction, *HRI '10: Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction* (ACM, New York, NY, USA, 2010), pp. 375–382.
17. C. Liu, K. Conn, N. Sarkar and W. Stone, Online affect detection and robot behavior adaptation for intervention of children with autism, *IEEE Trans. Robotics* **24**(4) (2008) 883–896.

18. E. Mower, D. Feil-Seifer, M. Mataric and S. Narayanan, Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements, *Proc. 16th IEEE Int. Workshop on Robot and Human Interactive Communication, RO-MAN '07* (IEEE, 2007), pp. 1125−1130.
19. G. Castellano, A. Pereira, I. Leite, A. Paiva and P. W. McOwan, Detecting user engagement with a robot companion using task and social interaction-based features, *Int. Conf. Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI'09)* (ACM Press, Cambridge, MA, USA, 2009), pp. 119−126.
20. J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan and A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, *ACM/IEEE Int. Conf. Human-Robot Interaction* (ACM, Lausanne, Switzerland, 2011).
21. R. Jain and P. Sinha, Content without context is meaningless, *Proc. ACM Int. Conf. on Multimedia* (ACM, 2010), pp. 1259−1268.
22. H. Aghajan, R. Braspenning, Y. Ivanov, L.-P. Morency, A. Nijholt, M. Pantic and M.-H. Yang, Use of context in vision processing: An introduction to the ucvp 2009 workshop, *Proc. Workshop on Use of Context in Vision Processing, ser. UCVP '09* (ACM, New York, NY, USA, 2009), pp. 1:1−1:3.
23. L.-P. Morency, I. de Kok and J. Gratch, Context-based recognition during human interactions: Automatic feature selection and encoding dictionary, *ACM Int. Conf. on Multimodal Interfaces (ICMI'08)* (Chania, Crete, Greece, 2008), pp. 181−188.
24. C. Peters, S. Asteriadis and K. Karpouzis, Investigating shared attention with a virtual agent using a gaze-based interface, *J. Multimodal User Interfaces* **3**(1−2) (2010) 119−130.
25. J. Sabourin, B. Mott and J. Lester, Modeling learner affect with theoreticallly grounded dynamic bayesian networks, *Proc. 4th Int. Conf. on Affective Computing and Intelligent Interaction (ACII'11)* (Springer, 2011).
26. L. Malta, C. Miyajima and K. Takeda, Multimodal estimation of a driver's affective state, *Workshop on Affective Interaction in Natural Environments (AFFINE), ACM Int. Conf. on Multimodal Interfaces (ICMI'08)* (Chania, Crete, Greece, 2008).
27. H. P. Martinez and G. N. Yannakakis, Mining multimodal sequential patterns: A case study on affect detection, *Proc. 13th Int. Conf. on Multimodal Interaction (ICMI'11)* (ACM, New York, NY, USA, 2011).
28. G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva and P. McOwan, It's all in the game: Towards an affect sensitive and context aware game companion, *Proc. 3rd Int. Conf. on Affective Computing and Intelligent Interaction (ACII 2009)* (IEEE, 2009), pp. 29−36.
29. I. Leite, C. Martinho, A. Pereira and A. Paiva, As time goes by: Long-term evaluation of social presence in robotic companions, *IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)* (IEEE, Toyama, Japan, 2009).
30. C. Martinho and A. Paiva, Using anticipation to create believable behaviour, *American Association for Artificial Intelligence Technical Conference* (Boston, July 2006), pp. 1−6.
31. J. A. Russell, A circumplex model of affect, *J. Personality Social Psychol.* **39** (1980) 1161−1178.
32. Y.-W. Chen and C.-J. Lin, Combining SVMs with various feature selection strategies, *Feature Extraction, Foundations and Applications*, eds. I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh (Springer, 2006).
33. I. Leite, G. Castellano, A. Pereira, C. Martinho and A. Paiva, Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world

settings, *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)* (ACM, Boston, MA, USA, 2012).
34. N. Borgers, E. de Leeuw and J. Hox, Children as respondents in survey research: Cognitive development and response quality, *Bulletin de Methodologie Sociologique* **66**(1) (2000) 60.
35. I. Poggi, *Mind, Hands, Hace and Body. A Goal and Belief View of Multimodal Communication* (Weidler, Berlin, 2007).
36. C. L. Sidner, C. D. Kidd, C. H. Lee and N. B. Lesh, Where to look: A study of human-robot engagement, *IUI '04: Proc. 9th Int. Conf. on Intelligent User Interfaces* (ACM, Funchal, Madeira, Portugal, New York, NY, USA, 2004), pp. 78–84.
37. I. Leite, S. Mascarenhas, A. Pereira, C. Martinho, R. Prada and A. Paiva, Why can't we be friends? An empathic game companion for long-term interaction, *IVA'2010*. Lecture Notes of Computer Science, Vol. 6356 (Springer), pp. 315–321.
38. G. Castellano, R. Aylett, A. Paiva and P. W. McOwan, Affect recognition for interactive companions, *Proc. Workshop on Affective Interaction in Natural Environments: Real-Time Affect Analysis and Interpretation for Virtual Agents and Robots (AFFINE'08) ICMI'08* (Chania, Crete, Greece, 2008).
39. G. Castellano and C. Peters, Socially perceptive robots: Challenges and concerns, *Interaction Studies* (John Benjamins Publishing Company, 2010).
40. I. Leite, G. Castellano, A. Pereira, C. Martinho and A. Paiva, Long-term interactions with empathic robots: Evaluating perceived support in children, *Proc. Int. Conf. on Social Robotics.* (Chengdu, China, 2012).
41. G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva and P. W. McOwan, Detecting engagement in HRI: An exploration of social and task-based context, *Proc. IEEE/ASE Int. Conf. on Social Computing (SocialCom'12)* (Amsterdam, The Netherlands, 2012).
42. L. D. Riek, P. Paul and P. Robinson when my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry, *J. Multimodal User Interfaces* **3**(1) (2010) 99–108.
43. A. Tapus and M. J. Mataric Socially assistive robots: The link between personality, empathy, physiological signals, and task performance, *AAAI Spring Symp.* (Palo Alto, Stanford, USA, March, 2008).
44. R. Aylett, G. Castellano, B. Raducanu, A. Paiva and M. Hanheide Long-term socially perceptive and interactive robot companions: Challenges and future perspectives, *Proc. Int. Conf. Multimodal Interaction (ICMI'11)* (Alicante, Spain, 2011).
45. P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris and V. Enescu Long-term human-robot interaction with young users, *Proc. IEEE/ACM HRI-2011 Workshop on Robots Interacting with Children* (Lausanne, Switzerland, 2011).

Dr. Ginevra Castellano is a senior researcher at the University of Birmingham, United Kingdom. Her research interests lie at the crossroads of affective behavioral computing and social robotics, and include the automatic analysis and recognition of human non-verbal behavior for adaptive human-computer and human-robot interaction. She has published more than 50 research articles on these topics. She is the coordinator of the EU FP7 EMOTE (EMbOdied-perceptive Tutors for Empathy-based learning) project (2012−2015). She is also co-investigator of the EU FP7 ILearnRW (Integrated Intelligent Learning Environment for Reading and Writing) project (2012−2015). She is a member of the management board of the HUMAINE Association; co-founder and co-chair of the AFFINE (Affective Interaction in Natural Environments) workshops; co-editor of special issues of the Journal on Multimodal User Interfaces and the ACM Transactions on Interactive Intelligent Systems; Program Committee member of several international conferences, including ACM/IEEE HRI, ACM ICMI, IEEE SocialCom, AAMAS, IUI, ACII.

Iolanda Leite is a Ph.D. candidate at Instituto Superior Técnico, Technical University of Lisbon. She works as a research assistant at EU-funded project LIREC (Living with Robots and IntEractive Companions, FP7). Previously, she was also involved in MINDRACES (from Reactive to Anticipatory Cognitive Systems, FP6). Her research interests concern with the role of empathy and adaptive behavior in long-term human-robot interaction. She has authored and co-authored over 20 peer-reviewed scientific papers. Her work was published at conferences such as the ACM/IEE HRI (Int. Conf. On Human-Robot Interaction), IEEE RO-MAN (Int. Symposium in Robot and Human Interactive Communication), ACM Multimedia, ACM UMAP (User Modeling, Adaption, and Personalization) and ACII (Affective Computing and Intelligent Interaction).

André Pereira is a Ph.D. student at Instituto Superior Técnico, Technical University of Lisbon. He started his academic adventure by finishing his Master thesis at GAIPS (Intelligent Agents and Synthetic characters group). In his Master thesis he developed and studied a scenario where a social robot dotted with emotional behavior could play chess with human opponents. In his Ph.D. he is still studying how to create more believable and enjoyable board game opponents. He authored or co-authored more than 20 scientific papers, in areas such as human-robot interaction, digital games, intelligent virtual agents and multi-agent systems. Currently, he is working on his dissertation and as a research assistant at EU-funded project LIREC (Living with Robots and IntEractive Companions, FP7).

Carlos Martinho received his Ph.D. degree in Computer Science and Engineering from Instituto Superior Técnico, Technical University of Lisbon. He is currently an Assistant Professor in the Computer Science and Engineering Department of IST and a Senior Researcher in the Intelligent Agents and Synthetic Character Group at INESC-ID. His research interests include autonomous agents, synthetic characters, social robotics and user adaptation, with a focus on the creation of believable behavior. He has co-authored over 80 papers, served on the program committee of international conferences as ACII, IVA and AAMAS, and as a reviewer for international journals as Computational Intelligence, JAAMAS and TAFFC. He is currently working in EU FP7 project LIREC, developing long-term synthetic companions, in EU FP7 project SIREN, developing intelligent interactive software to support teaching conflict resolution skills to children, and in UTAustin-Portugal cooperation agreement project INVITE, developing autonomous characters with social identity awareness.

Professor Ana Paiva is a research group leader of GAIPS at INESC-ID and an Associate Professor at Instituto Superior Técnico, Technical University of Lisbon. She is well known in the area of Artificial Intelligence Applied to Education, Intelligent Agents and Affective Computing. Her research is focused on the affective elements in the interactions between users and computers with concrete applications in robotics, education and entertainment. She served as a member of numerous international conference and workshops. She has (co)authored over 100 publications in refereed journals, conferences and books. She was a founding member of the Kaleidoscope Network of Excellence SIG on Narrative and Learning Environments, and has been very active in the area of Synthetic Characters and Intelligent Agents. She coordinated the participation of INESC-ID in several European projects, such as the IDEALS (funded under the Telematics program), NIMIS (an I3-ESE project), DiViLab and Safira (IST- 5th Framework), where she was the prime contractor, VICTEC, COLDEX, MindRaces and E-Circus (in the 6th framework), LIREC (in the 7th Framework), SIREN, Ecute and SEMIRA (under the Complexity Net area).

Peter McOwan is currently Professor of Computer Science in the School of Electronic Engineering and Computer Science at Queen Mary, University of London. His research interests are in visual perception, mathematical models for visual processing, in particular motion, cognitive science and biologically inspired hardware and software. He has authored more than 100 papers in these areas. He recently served on the Program Committee for ACII2009, CVPR 2009 and IEEE Artificial Life and is a member of the editorial board of the Journal on Multimodal User Interfaces. Current research projects include LIREC, an EU FP7 IP, developing long term synthetic companions, an EPSRC programme grant CHI+MED investigating design to reduce human errors in medical software and an EPSRC PPE CS4fn, an outreach project to enthuse schools about computer science research.