Inter-ACT: An Affective and Contextually Rich Multimodal Video Corpus for Studying Interaction with Robots

Ginevra Castellano
Dept. of Computer Science
School of EECS
Queen Mary University of
London, United Kingdom
ginevra@dcs.qmul.ac.uk

Iolanda Leite, André Pereira, Carlos Martinho, Ana Paiva
INESC-ID
Instituto Superior Técnico
Porto Salvo, Portugal

Peter W. McOwan
Dept. of Computer Science
School of EECS
Queen Mary University of
London, United Kingdom

ABSTRACT

The Inter-ACT (INTEracting with Robots - Affect Context Task) corpus is an affective and contextually rich multimodal video corpus containing affective expressions of children playing chess with an iCat robot. It contains videos that capture the interaction from different perspectives and includes synchronised contextual information about the game and the behaviour displayed by the robot. The Inter-ACT corpus is mainly intended to be a comprehensive repository of naturalistic and contextualised, task-dependent data for the training and evaluation of an affect recognition system in an educational game scenario. The richness of contextual data that captures the whole human-robot interaction cycle, together with the fact that the corpus was collected in the same interaction scenario of the target application, make the Inter-ACT corpus unique in its genre.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.5.2 [**Pattern Recognition**]: Design Methodology; J.4 [**Computer Applications**]: Social and Behavioural Sciences

General Terms

Algorithms, Human Factors, Design, Theory

Keywords

Affect recognition, human-robot interaction, multimodal video corpus, context awareness, application-dependent design

1. INTRODUCTION

As robots are increasingly being viewed as social entities to be integrated in our daily lives [11], social perceptive abilities seem a necessary requirement for enabling more natural interaction with human users. These include recognising people's affective expressions and states, and accounting for the events unfolding in the environment.

The design of a context-sensitive affect recognition system for socially perceptive robots relies on representative data. Most of the existing corpora and databases of affective expressions include posed data collected in scenarios which differ from that of the final application [12]. Moreover, while many of the most recent databases contain multimodal data, the availability of contextual information is still not frequent.

Nevertheless, naturalistic human-machine interaction requires an affect recognition system to be trained and validated with contextualised affective expressions, that is, expressions that emerge in the same interaction scenario of the target application [1] [3]. In addition, representative data for automatic inference of the user's affect in human-robot interaction should include not only information about the user's behaviour, but also information about the task that the user and the robot are involved with and the behaviour generated by the robot itself. In fact, in human-robot interaction, the robot and the user mutually influence each other in a continuous cause and effect cycle: the behaviour of the robot may elicit a response from the user and, similarly, the behaviour and actions of the latter may trigger the generation of an appropriate response from the robot. Thus it becomes necessary for a corpus of affective expressions to capture as much information as possible about the interaction cycle.

In this paper we present the Inter-ACT (INTEracting with Robots - Affect Context Task) corpus, an affective and contextually rich multimodal video corpus including affective expressions of children playing chess with the Philips iCat robot [6]. The Inter-ACT corpus contains videos from multiple view-points that allow for the interaction to be captured from different perspectives and includes synchronised contextual information about the game and the iCat's behaviour. These could be used to derive contextualised visual features.

The Inter-ACT corpus is intended to be a comprehensive repository of naturalistic and contextualised, task-dependent data in an educational game scenario. It is one of the first to be collected in the same interaction scenario of the final application. Moreover, it is unique in its genre, as it combines multiple views of the user, as well as synchronised task-dependent contextual information.

Our corpus is mainly intended to provide representative

MM'10, October 25–29, 2010, Firenze, Italy.

data for the training and validation of an automatic system for the inference of the user's affect in a specific application. However, its use also extends to encompass the study of dynamics in human-robot interaction and it represents a methodology for the collection of representative data in adaptive human-robot interaction scenarios.

2. RELATED WORK

Most of the affective video corpora and databases available in the literature contain acted expressions recorded in contexts that are not specific to a particular application [12]. While naturalistic databases are gradually increasing [8] and some examples of databases including context-specific expressions have been reported [4], repositories of data collected in the same scenario of the target application are still uncommon. An exception is represented by the CAL database by Afzal and Robinson [1], which contains affective expressions collected in a computer-based learning environment.

Although still not numerous, some efforts on the design of affect recognition systems for a specific interaction scenario that take into account contextual information have been reported in the literature [12]. For example, Kapoor and Picard [5] designed a system for the detection of interest in a learning environment that combines non-verbal cues and information about the learner's task (e.g., level of difficulty and state of the game).

New human-computer and human-robot interaction applications require the definition of corpora including contextual information about the task and the loop the user and their artificial interactor are involved in. The Inter-ACT corpus represents an attempt to provide a comprehensive repository of contextually rich video data in a specific human-robot interaction scenario.

3. 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 series of synchronised contextual features that describe the events of the game and the behaviour displayed by the robot. Videos and contextual data together provide a comprehensive description of the ongoing interaction.

3.1 Data collection

3.1.1 Subjects, scenario and setup

The data collection procedure was performed in two different locations, a primary school where every week children have two hours of chess lessons, and a chess club where children are more experienced and practice chess more frequently. 8 children (6 male and 2 female, average age 8.5) took part in the data collection procedure (Figure 1).

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 behaviours.

Figure 1: The children that took part in the data collection procedure.

In each exercise the robot begins the interaction by inviting the user to play. After each move is played by the user, the iCat asks them to play 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 minutes at most.

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, 1024X768 spatial resolution) and a DV camera (25 fps, 720X576 spatial resolution). For the lateral view, we used a DV camera (25 fps, 720X576 spatial resolution). A standard 25 fps webcam was used to capture the behaviour displayed by the iCat. Figure 2 shows some examples of frames from the frontal and lateral view.

3.1.2 Contextual information

Humans can infer the state of others just by direct observation of emotional cues, such as facial expressions, and/or by taking into consideration the others' situation and the events unfolding in the environment. In the immersive context of a chess game, the information about the game can be extremely useful to understand what players might be thinking or feeling. In a previous study [2], we showed that in a chess match contextual information is successful to discriminate the valence (positive or negative) of the players' affective state and their level of engagement with the iCat.

During the video recordings of our corpus, contextual information about the game and the iCat's behaviour was also logged in real-time. Every log entry contains a timestamp to enable the synchronisation with the corresponding video files containing the users' non-verbal behaviours. 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 series of contextual information retrieved after each move played by the user. Thus each video of the

Figure 2: Examples of frames from the frontal and the lateral view in the Inter-ACT corpus.

Inter-ACT corpus includes the following synchronised 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 viceversa.

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.

iCat's facial expressions: after every move played by the user, the iCat evaluates the new state of the game, updates its affective state and provides feedback to the user by displaying a facial expression. The robot's affective state is computed using the *emotivector* system by [7], an anticipatory mechanism that generates an affective signal based on the mismatch between an expected and a sensed value. In our case, the input for the emotivector system is the game state (from the iCat's perspective). The expected value is obtained by the history of the previous game state values, and the sensed value is the current game state. The nine possible outcomes of the emotivector system are described in Table 1. Each outcome is then mapped onto a different facial expression to be displayed by the iCat. For example, if after a few moves in the game the iCat has already captured an opponent's piece, it might be expecting to maintain the advantage in the game (in the emotivector terminology, expecting a "reward"). In this situation, if the user makes a move that is worse than the one the iCat was expecting (e.g., by putting their queen in a very dangerous position), the generated affective signal will be a "stronger reward", and the robot will display a facial expression of excitement. For more details on the iCat's affective system please see [6].

The information about the iCat's facial expressions during the interaction might be relevant to understand certain aspects of mimicry, empathy or engagement on the user's side. For example, during the recordings, we noticed that sometimes users mimic the facial expressions displayed by the robot.

From the contextual features extracted in real-time, we also derived other features that provide additional information about the situation of the game from the user's perspective:

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.

Table 1: Description of the nine possible outcomes of the emotivector system, and the corresponding animations in the iCat's embodiment.

Sensation	Description	Animation
Stronger Reward	Better than expected	Excited
Expected Reward	As good as expected	Confirm
Weaker Reward	Not as good as expected	Happy
Unexpected Reward	Good, not expected	Arrogant
Negligible	Nothing changed	Think
Unexpected Punishment	Bad, not expected	Shocked
Weaker Punishment	Not as bad as expected	Apologise
Expected Punishment	Bad, as expected	Angry
Stronger Punishment	Worse than expected	Scared

User sensations: calculated using the same mechanism used by the iCat to generate its affective reactions, but taking into account the user's game state. This feature attempts to predict the user's possible sensations of the events happening in the game. Consider the situation described above to illustrate the iCat's affective behaviour, but from the perspective of the user: after three moves in the game the user has lost one piece, so they might be expecting the iCat to keep the advantage (i.e., expecting a "punishment"). If the user plays a very bad move, they might experience something closer to a "stronger punishment" sensation.

In addition to providing useful information about the game and the behaviour of the robot, the above information can be of valuable help when attempting to automatically derive contextualised visual features such as facial expressions and eye gaze. Thus, behavioural and task-related features would not only be used separately, but fused together for the automatic prediction of the user's affect. For example, we may want to know if the user looked at the robot after a certain move or after the iCat displayed a specific behaviour.

4. ANNOTATION

The Inter-ACT corpus includes 312 six-second videos (156 frontal and 156 lateral videos) of the children's behaviour pre-segmented by an expert coder. Pre-segmentation was performed starting from the frontal videos in order to include coherent samples of behaviour: the corpus includes samples displaying full expressions, no expressions and blends of expressions. There are roughly an equal number of samples belonging to each category.

4.1 Affective labels

The Inter-ACT corpus is provided with affective labels that describe each "thin-slice" of the interaction. In the educational game scenario with the iCat robot, we are inter-

	Kappa (1st group)	Kappa (2nd group)
Valence of feeling (2 labels)	0.27	0.36
Valence of feeling (3 labels)	0.27	0.41
Interest towards the iCat (2 labels)	0.23	0.36
Interest towards the iCat (3 labels)	0.15	0.23

Table 2: Measures of inter-coder agreement for the videos of the Inter-ACT corpus.

ested in capturing the user's states that are related with the game and the social interaction with the iCat. The *valence of the feeling* experienced by the user during the game was chosen to describe the degree to which the user's affect is positive or negative [10]. In addition, the user's *interest towards the iCat* was identified to describe in what measure the user pays attention to the iCat over time [9].

26 students and researchers (10 male and 16 female, average age: 21.9) were recruited for an annotation experiment. The coders were divided in two groups of 13 people each and each group was assigned 78 videos (frontal view) to label. The coders were asked to assess the affective components of *valence of the feeling* and *interest towards the iCat* in each video based on the behaviour displayed by the children, without any information about the context. Specifically, the annotators were asked to label each video in terms of the above affective components in two annotation steps. Regarding the *valence of the feeling*, they were required to choose among positive or negative in the first step, and positive, negative or neutral in the second step. Regarding the *interest towards the iCat*, the annotators could choose among high interest or low interest in the first step, and high interest, low interest or medium interest in the second step. The annotators were provided with a clear description of each label.

Inter-coder agreement was measured with the Fleiss' kappa statistics: results show an overall fair agreement for the affective labels (see Table 2).

5. CONCLUSIONS

This paper presented the Inter-ACT video corpus, a multimodal video corpus of affective expressions emerging in a human-robot interaction scenario. The corpus is a comprehensive repository of visual data and synchronised contextual information about the task and the interaction between the user and their artificial interactor. It is intended to provide representative data for the training and evaluation of an automatic affect recognition system for the iCat robot in the presented scenario. Currently the corpus is protected for privacy reasons, due to the presence of children in the recordings.

Annotation of affect by non expert coders showed an overall fair inter-coder agreement. Future studies will compare these results with those of an annotation performed by expert coders.

6. ACKNOWLEDGEMENTS

This work was supported by the EU FP7 ICT-215554 project LIREC (LIving with Robots and intEractive Companions).

7. REFERENCES

[1] S. Afzal and P. Robinson. Natural affect data - collection and annotation in learning context. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, pages 22–28. IEEE, 2009.

[2] 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. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, pages 29–36. IEEE, 2009.

[3] 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. *Journal on Multimodal User Interfaces*, 3(1–2):89–98, 2010.

[4] E. Douglas-Cowie, R. Cowie, and M. Schröder. A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*, 2000.

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

[6] I. Leite, A. Pereira, C. Martinho, and A. Paiva. Are emotional robots more fun to play with? In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 77–82, Aug. 2008.

[7] C. Martinho and A. Paiva. Using anticipation to create believable behaviour. In *American Association for Artificial Intelligence Technical Conference*, pages 1–6, Boston, July 2006.

[8] M. Pantic, M. Valstar, R. R., and L. Maat. Web-based database for facial expression analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pages 317–321, July 2005.

[9] C. Peters, S. Asteriadis, and K. Karpouzis. Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces*, 3(1-2):119–130, 2010.

[10] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[11] F. Tanaka, A. Cicourel, and J. R. Movellan. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Science*, 194(46):17954–17958, 2007.

[12] 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.