Deliverable 3.2

"Framework for the perception of user actions"

Contract number**: FP7-215554   LIREC**

LIving with Robots and intEractive Companions

Start date of the project: 1st March 2008 Duration: 54 months

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement* n° 215554.

Identification sheet

Project ref. no.	**FP7-215554**
Project acronym	LIREC
Status & version	"D3.2"
Contractual date of delivery	30th of April 2009
Actual date of delivery	19th of June 2009
Deliverable number	D3.2
Deliverable title	"Framework for the perception of user actions"
Nature	Report
Dissemination Level	PU Public
WP contributing to the deliverable	WP3
WP / Task responsible	WP3/T3.1.2
Editor	Ana Paiva
Editor address	INESC-ID / Instituto Superior Técnico Av. Prof. Dr. Cavaco Silva, TagusPark 2780-990 Porto Salvo, Portugal
Author(s) (alphabetically)	Ginevra Castellano (QM), Ricardo Chaves (INESC-ID), Pedro Cuba (INESC-ID), João Dias (INESC-ID), Mário Rui Gomes (INESC-ID), Mariusz Janiak (WRUT), Iolanda Leite (INESC-ID), Carlos Martinho (INESC-ID), David Matos (INESC-ID), Peter McOwan (QM), Robert Muszynski (WRUT), Luís Oliveira (INESC-ID), Ana Paiva (INESC-ID), Thomas Pellegrini (INESC-ID), André Pereira (INESC-ID), Krzysztof Tchon (WRUT), Andreas Wichert (INESC-ID), Marek Wnuk (WRUT). Checkers: Peter McOwan (QM)
EC Project Officer	Pierre-Paul Sondag
Keywords	Perceiving user actions, affect sensitivity, facial expression, body expression, speech recognition, locating the user, contextual information.
Abstract (for dissemination)	This document describes the work that LIREC partners have been conducting towards the development of a joint framework for the perception of the user actions using several interaction modalities. These include user's face and body expression, limited speech recognition and contextual information of the interaction.

CONTENTS

1 Introduction

One of the main requirements for an artificial companion to sustain a long-term interaction with humans is the ability to display social, affective behaviour (Breazeal, 2003). Social capabilities are necessary for all those applications in which a robot or a virtual agent need to interact with humans as a companion, a partner or a friend (Dautenhahn, 2007). To accomplish this, interactive companions must be capable of sensing, processing and interpreting information about the user's actions and the context in which those actions take place, so that they can plan and generate appropriate responses.

In deliverable D3.1 we reported a set of initial experiments related to communication between users and companions. The purpose of such experiments was to exploit further techniques and modalities to use in the project. Given that LIREC companions will be developed and tested in real world scenarios for long periods of time, the technology has to be robust, but at the same time non-intrusive for users. As such, the directions concerning technology to perceive user activity include markerless vision systems aware of user body pose and facial expressions, limited speech recognition capabilities, as well as contextual information that may be scenario dependent.

This deliverable describes the work that LIREC partners have been conducting towards the development of a joint framework for the perception of the user actions using several interaction modalities. We use the term *user actions* not only to describe concrete activities performed by the user (e.g. user left the room, user played a move at the chessboard, etc.), but also to address more abstract activities related to the user's mental or affective states (e.g., user is engaged with the companion).

Understanding the user's affective or mental states from her verbal and non-verbal behaviour is of vital importance for the companion to establish and maintain meaningful social relations with users. The ability to attribute affective or mental states to the user can be referred to as *affect sensitivity.* Affect sensitivity concerns to the way social affective cues conveyed by people's behaviour can be used to infer behavioural states. These span from basic emotions (such as joy, anger, sadness, etc.) to more complex affective and mental states such as interest, boredom or frustration.

This document is organized as follows. In Section 2 we present the initial technology developed in LIREC to interpret some user's verbal and non-verbal behaviours in terms of facial expression, body expression and speech. After that, in Section 3, we show how such technology is being applied to implement user location and proximity distance behaviours in two different mobile robots. Section 4 describes a multi-level approach for the analysis of user affective cues which takes into account the variety of interaction modalities. A case study in the "MyFriend" scenario is also reported: an experiment that investigates which non-verbal behaviours displayed by the users and which contextual information are more relevant to discriminate among some user states. We intend to exploit the findings of this study to carry out a rigorous design of an affect recognition system for a game companion. Finally, Section 5 contains some conclusions and future work directions.

2 Technology to perceive user’s verbal and non-verbal behaviour

The mail goal of this phase in WP3 is the development of a framework to perceive user actions and states by combining different modalities of human communication and expressivity such as facial expressions, body gestures and speech recognition. Another important aspect of this WP is the design of an affect recognition system capable of identifying some of the user’s affective states by employing the modalities indicated earlier, as well as contextual information of the interaction between the companion and the user.

Different software modules for the automatic detection and analysis of the user’s verbal and non-verbal behaviour, which are the result of the contribution of different partners in the project, are currently under test in the LIREC scenarios. This section describes not only modules that are being developed from scratch, but also some open source modules that will be integrated in the framework.

2.1 Face Expression

In this section we present two different modules that rely on user’s facial expressions. The first one is FacET library and allows for automatic extraction of some facial features; the second one is being developed with the purpose of recognizing different users.

2.1.1 Facial Features Detection

WRUT developed FacET (Face Expression Tracker), a library for the automatic extraction of facial features that is cross-platform and runs on Windows and Linux. Based on functions provided by the OpenCV library (Intel) and on the work by Namysl (Namysl, 2008), FacET (available at: https://lirec.ict.pwr.wroc.pl/svn/FacET/) allows for the automatic detection of multiple faces and extraction of 14 facial features:

- left eyebrow bend angle (top)
- left eyebrow declination angle (side)
- right eyebrow bend angle (top)
- right eyebrow declination angle (side)
- distance between the left eyelids (relative to the eyeball sub region)
- distance between left pupil and eyebrow top (relative to the eye sub region)
- distance between the right eyelids (relative to the eyeball sub region)
- distance between right pupil and eyebrow top (relative to the eye sub region)
- aspect ratio of the lips bounding box
- Y position of the left corner of the lips (relative to the lips bounding box)
- Y position of the right corner of the lips (relative to the lips bounding box)
- number of horizontal wrinkles in the center of the forehead
- nostrils baseline width (relative to the face width)
- area of the visible teeth (relative to the lips bounding box)

Automatic extraction of the facial features is performed through the following main steps: (1) image acquisition, (2) face detection, (3) face sub-regions extraction (eyes, mouth, nose, forehead), (4) facial features parameterisation (see Figure 1).

Face detection can be performed both using Haar classifiers, which allow for a very precise (but time consuming) region of interest (ROI) definition, and using skin colour detection (Huang & Chiang, 2006), which is much faster, but requires that there are no objects of the skin-like colour in the field of view in order to avoid erroneous hits.

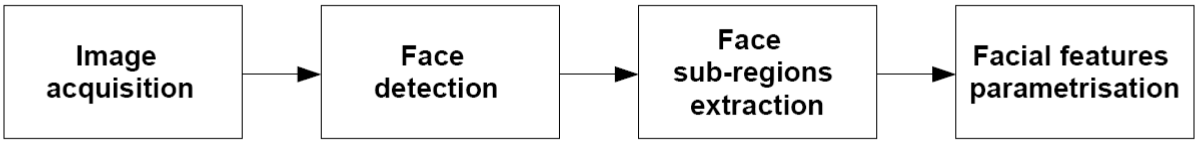


Figure 1. Steps for the automatic extraction of facial features.

Detection of the face sub-regions has been implemented both with Haar classifiers and a simple but effective method of horizontal and vertical projections (Chen *et al*., 2005).

In the last processing step the facial features are extracted in appropriate sub-regions and the numerical parameters are computed (see Figure 2). Several image segmentation methods have been used:

- dual (hysteresis) thresholding (eyebrows, teeth)
- template matching (eyes)
- morphological opening and propagation (eyebrows, lips)
- Hough transform (eyebrows, pupils)
- Haar classifier (eyes, lips)

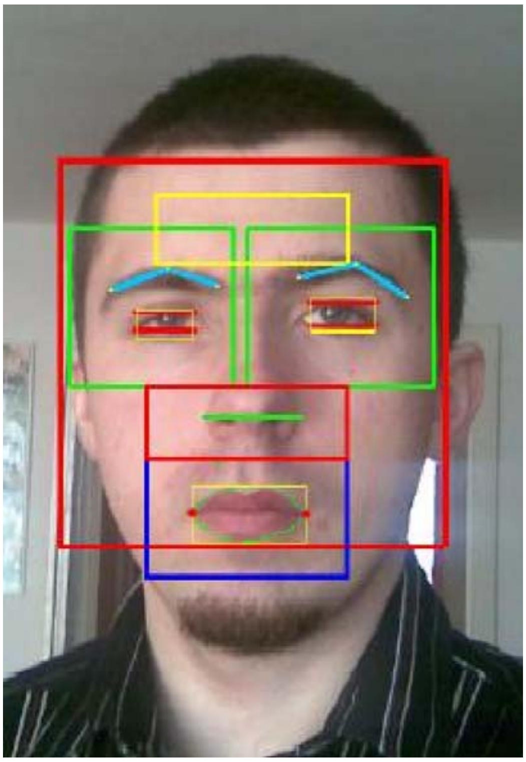


Figure 2. Automatically extracted facial features (from Namysl, 2008).

The FacET library is currently under test in the LIREC scenarios, where its usability has to be evaluated with respect to several issues deriving from the presence of different levels of interaction (e.g., face-to-face, short or long range interaction with the companion), the presence of a high variety of users interacting with the companions (e.g., children or adults), the need to work in real-world environments (where, for example, changes in the illumination and in the background may occur) and on mobile robotic platforms.

2.1.2 User identification

FOAM developed a prototype system that allows for a simple face identification to be performed. The system has two modes of operation. The first one is calibration, where each user must show their face to the camera separately and can be assigned with a different ID. The second mode of operation is face identification (based on the face detection code provided by the OpenCV library), where each visible detected face is given an ID based on its closest match (within a given error threshold) with the faces recorded during calibration. Image differencing is performed to compare each detected face with those recorded during the calibration (see Figure 3 and Figure 4). The system runs on Windows and Linux, supports the YARP network interface and is currently integrated in “MyFriend” scenario.

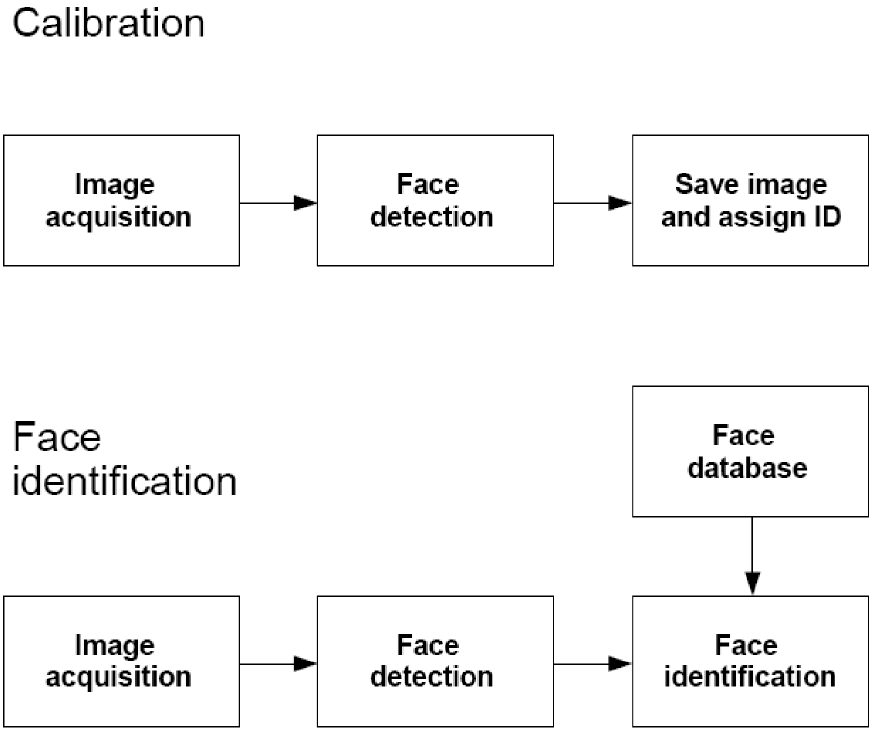


Figure 3. The calibration and face identification modes of operation.

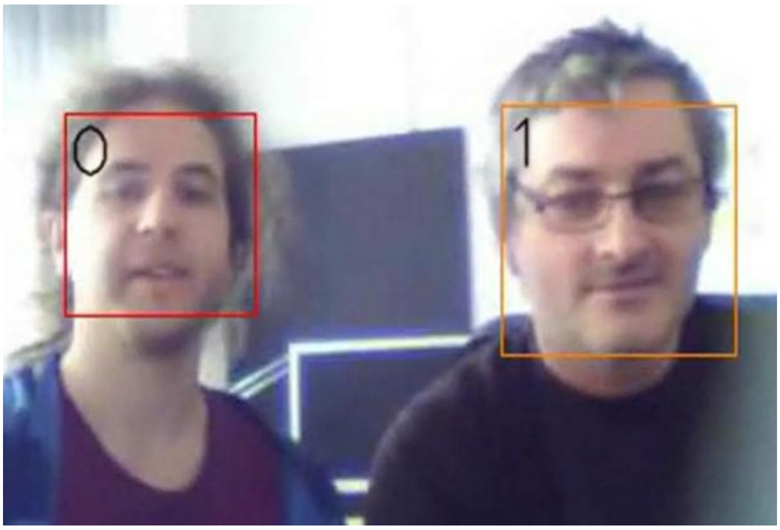


Figure 4. Prototype system for face identification: the boxes drawn around the faces display the IDs.

2.2 Body Expression

QMUL developed a prototype system that allows for the real-time prediction of whether the user is staying still, approaching the camera or withdrawing.

The system works with a frontal view of the user and is based on the face detection code provided by the OpenCV library. For each frame the face is detected and the area of the face bounding box is computed. The values of the area are stored and used for the prediction of the type of movement (staying still, approaching the camera or withdrawing). The area of the face bounding box in the current frame is compared with the values of the area in a temporal window preceding it.

If the area in the current frame does not change much from the area in the first frame of the window then the user is regarded as staying still. If the area in the current frame is greater than the average area over the temporal window preceding it, then the user is predicted to be approaching the camera, otherwise to be withdrawing from it (see Figure 5 and Figure 6).

This system runs on Windows and Linux and is currently under test in the “Spirit of the Building” scenario (see section 3.1).

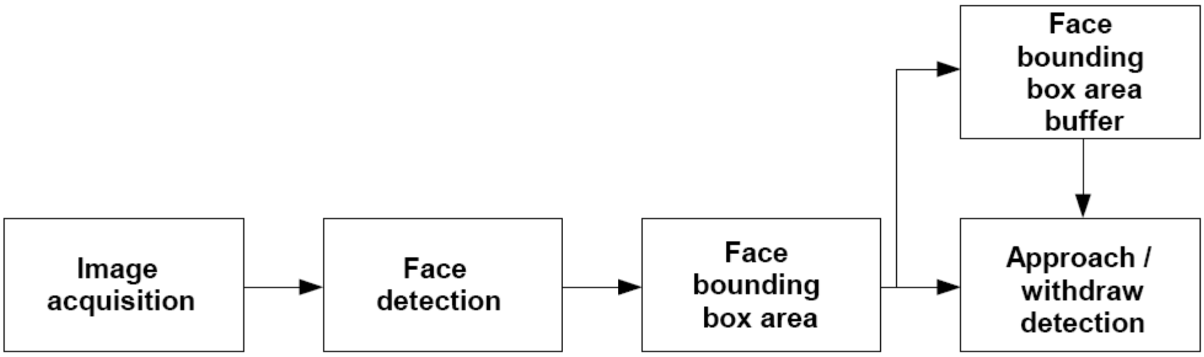
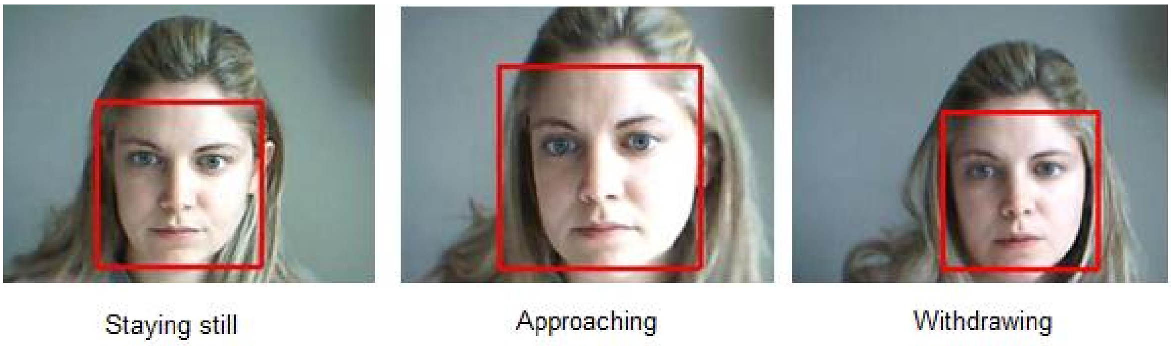


Figure 5. Main components of the system for approach / withdraw detection.



Staying still Approaching Withdrawing

Figure 6. From left to right, user is staying still, approaching the camera and withdrawing.

The Watson (Head Tracking and Gesture Recognition Library) library may represent a valuable resource if a specific interaction scenario requires the companion to track the user's head orientation and detect the user's head gestures: Watson, in fact, recognises two head gestures using support vector machines (SVMs), head nods and head shakes.

The OpenCV library provides many algorithms that can be employed to perform motion analysis, from segmentation techniques to tracking algorithms (e.g., Camshift algorithm, motion templates, etc.). These represent a useful tool for analysing global indicators of human movement, such as the presence of people in a room, presence of movement, degree of contraction/expansion of movement (e.g., using bounding box of human silhouette, etc.), and are also applicable to gesture recognition.

2.3 Speech

The deployment of speech recognizers in real-life situations is limited by the sensitivity of these systems due to four major factors:

- Inter-speaker variability;
- Intra-speaker variability;
- Channel variability;
- Background noise.

The inter-speaker variability accounts for the different ways that a sentence can be uttered by different speakers while the intra-speaker variability is related with variations in pronunciation by the same speaker. The other two factors are related to signal acquisition conditions.

2.3.1 Standardization of speech signal acquisition

In order to assure a similar performance of the speech recognition systems in the different scenarios, some sort of common standardized hardware should be specified for the acquisition of the speech signal. The major aspects of channel variability are:

- The analog-to-digital converter characteristics;
- The anti-aliasing filter;
- The microphone characteristics;
- The distance from the microphone to the speakers mouth.

The first two factors can be minimized by specifying the characteristics of an audio adapter that can be used in different operating systems and on a wide range of hardware. Our proposal is to use an external usb audio converter. This option has the advantages of being more immune to digital noise, can be easily interchangeable and has a very small factor.

The selection of a commercially available system has the advantage of being easily replicable, and we propose the use of the Andrea PureAudio interface (Figure 7), that has the following specifications:

- stereo microphone input,
- stereo audio output,
- 8 KHz to 48 KHz sampling rate,
- 16 bit resolution,
- 20-20000 Hz bandwidth,
- power requirements: 4.5-5.5 V DC/ 120 mA

Figure 7. The Andrea PureAudio USB audio interface.

Regarding the selection of the microphone, since we cannot specify a common average distance from the speaker, the best option is to use a device that reduces the interference of ambient noises. The major noise reduction technologies are the spectral subtraction, where a secondary microphone captures the environment noise to be removed by signal processing techniques, and beamforming that uses multiple microphones to increase its directionality.

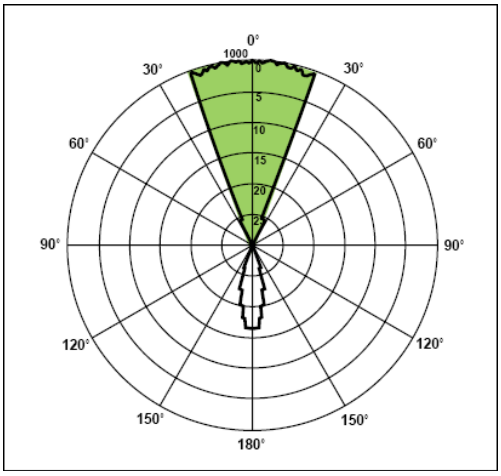


Figure 8. Dual microphone array polar pattern.

Given that the signal processing necessary to implement spectral subtraction technique degrades the performance of automatic speech recognizers, our option was to use a commercially available dual microphone array from the same manufacturer of the PureAudio USB audio interface. The performance of a speech recognition system with the Andrea

Superbeam microphone array (see Figure 9) has a performance comparable to the use of a headset microphone (results displayed on Figure 10).

Figure 9.The Andrea Superbeam microphone array.

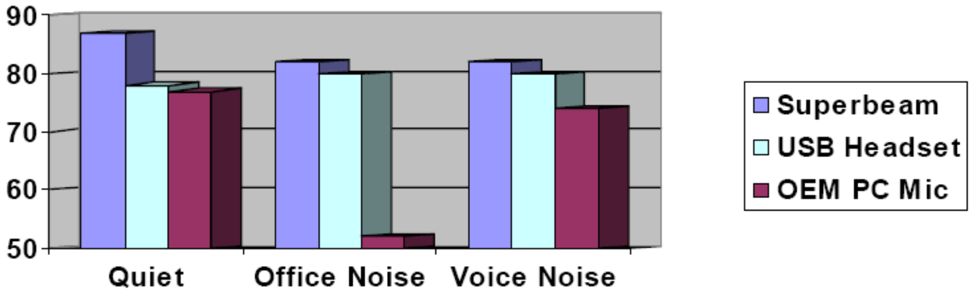


Figure 10. Results of speech recognion accuracy in a open cubicle office scenario (%).

2.3.2 English ASR System

In this section we evaluate how well can state-of-the-art speech recognition systems deal with intra and inter-speaker variability. For this purpose we will test two available systems under controlled conditions. The English ASR systems described in this section are specific to American English, although they can be used with British English with lower performances expected.

2.3.2.1 System selection

Two state-of-the-art automatic speech recognition (ASR) systems have been compared in terms of performance: our in-house system Audimus (Meinedo *et al.*, 2003), and the latest open source version of the Sphinx system (Sphinx-4), distributed by Carnegie Mellon University with no restriction against commercial use or redistribution (Walker *et al.*, 2004).

Audimus and Sphinx use two different core techniques: an hybrid Artificial Neural Network / Hidden Markov Model (ANN/HMM) scheme for the first one, and HMM of Gaussian Mixture Models (HMM/GMM) for the second. On one hand, the main advantage of hybrid ANN/HMM is that classification networks are usually considered better pattern classifiers than Gaussian mixtures approaches. Additionally, an appealing characteristic of the hybrid systems is that they are very flexible in terms of merging multiple input streams. On the other hand, one of the most significant limitations of hybrid systems is related to the lack of flexibility and increased difficulty when context-dependent phone modelling is desired.

On a practical point of view, one advantage of Sphinx is its modularity. Sphinx-4 supports a wide variety of tasks. Various implementations of the decoder are provided to support tasks that range from small vocabulary tasks (digit recognition, command-and-control applications) with the use of context free grammars, to large vocabulary tasks such as broadcast news transcription (HUB-4), with the use of stochastic language models (N-grams). Ready-to-use language models and context-dependent phone-like acoustic models are also available.

2.3.2.2 Corpora

Performance has been measured on the most difficult and general task of large vocabulary speech recognition: the HUB-4 task or the transcription of broadcast news.

Important work has been done to clean and normalize both transcriptions and newspapers texts used to train the language models. A better normalization of numbers, dates, URLs, acronyms has been performed. Tools from NIST and MIT have been adapted to build Perl modules that normalize this training material (both audio transcripts and newspaper texts). These modules can be very easily adapted to the specific conventions of a new corpus.

2.3.2.2.1 Training corpora for Audimus

The speech material for acoustic model training we used are the 1996 (LDC97S44) and 1997 (LDC98S71) English Broadcast News Speech corpus (HUB-4), comprised of respectively 73 and 67 hours of manually transcribed speech, coming from ABC, CNN and CSPAN television networks and NPR and PRI radio networks. The training sets of HUB-4 were used to train a set of 39 monophonic acoustic models, plus two non-speech models (one for silence and one for breath). A new phone for breath has been recently added and achieved around 2% relative (0.5% absolute) Word Error Rate (WER) improvement. Units that modelize the context (di-phones and tri-phones) are being currently trained, and that is expected to give around 2-3% absolute improvement over monophone units (Abad & Neto, 2008).

For language modelling, the orthographic transcriptions of the two corpus mentioned here-above (850k and 790k words) have been used for the HUB-4 language model (LM) part, along with an additional corpus of transcriptions, named CSR'96 (LDC98T31) that totalizes 149M words. The second part of the language model was trained on two newspaper and newswire text corpus: the North American News Text Corpus (LDC95T21 - 505M words), and the LDC98T30 corpus with 462M words. One LM per source (New York Times, Associated Press, Wall-Street Journal, REUFF, REUTE, LATWP) was built.

Up to nine LM were linearly interpolated with optimization of the weights on a subset (~3%, 22.3k words) of the HUB-4 1997 training corpus used as development corpus. This subset comprises the latest transcribed shows of the corpus. As expected, the biggest interpolation weights are those of the HUB-4 part of the language model (between 0.16 and 0.31), the newspaper part weights are all around 0.05 except for the NYT source of LDC98T30 with a 0.11 weight. A pruning at 1e-9 threshold has been processed to obtain the final LM.

The final language model is a 4-gram LM, with Kneser-Ney modified smoothing, comprised of 64k words (or 1-grams), 12 M 2-grams, 5.8M 3-grams and 4.5M 4-grams.

The vocabulary consists of all the words contained in the HUB4 training set plus the most frequent words in the broadcast news texts and Newspapers texts. These words were selected until the resulting vocabulary had 64k words. The pronunciations were extracted from the same public domain lexicon. For words not included in this lexicon, a rule-based grapheme-to-phone conversion system was used. The multiple-pronunciation lexicon included 70k entries.

2.3.2.2.2 Training corpora for Sphinx-4

Although software is provided to train acoustic models for the Sphinx decoder, already trained acoustic models are available for the HUB-4 task. The models are context-dependent triphones, with 8 Gaussians per mixture. The language model is a 3-gram LM, comprised of 64k words, 9.4M 2-grams and 13.5M 3-grams. The documentation about the data used to train both AMs and LMs is insufficient. It mentions "a variety of permitted sources, including broadcast news for the LM.", and some LDC HUB-4 corpus for AM, without mentioning which one(s).

2.3.2.3 Evaluation and results

Table 1 gives the Word Error Rates (WER) for both systems and for 3 distinct evaluation corpus: Eval 97, 98 and 03, which correspond respectively to LDC02S11, LDC00S86 and LDC07S10.

For Eval 97, the global WER is 27.5% but 19.7% on clean planned speech (the so-called F0 acoustic condition). It can be seen that Audimus outperforms Sphinx. The main reason may be the high Out-Of-Vocabulary (OOV) rates for Sphinx: 4.7% on Eval 97 and 1.9% on Eval 03. In comparison, the OOV rates for Audimus are 0.7% and 0.9% on the same evaluation data. Experiments crossing LM and AM resources were conducted: a test with Sphinx with our vocabulary and LM (the same we use with Audimus) showed worse performance. In fact, the AM should be re-trained when using a new vocabulary to have a fair comparison.

	Eval 97	Eval 98	Eval 03
Audimus	27.5 (F0: 19.7%)	26.9	26.5
Sphinx-4	30.5 (F0: 21.8%)		29.3

Table 1. Comparison of the Word Error Rates

2.3.2.4 Summary of work on English ASR

Performances of Sphinx are about 3% higher than those mentioned by the Sphinx developers on HUB-4 clean speech: a 18.8% WER on the evaluation data LDC00S88 is referenced on the Sphinx documentation. We purchased this corpus to manage complementary comparisons with Audimus. More expertise with the Sphinx system is needed to achieve better results with this system. No hard decision can be taken about the choice between both systems at the moment, although Audimus seems to outperform Sphinx for the HUB-4 task. Performance achieved with Audimus can still be improved, in particular the use of context dependent acoustic models (diphones and triphones) is ongoing work.

3 Locating the user

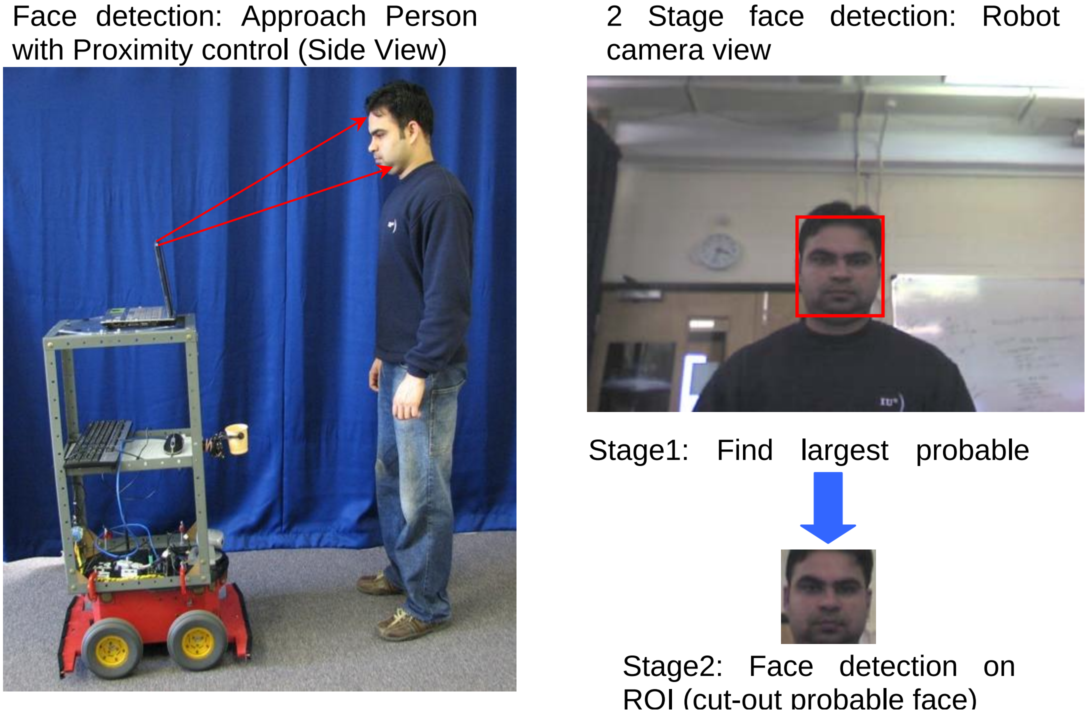
Locating the user is the starting point for the interaction between the companion and the user. As most of the LIREC scenarios include mobile robots, user proximity is a necessary functionality, so they can further perceive the user actions and affective states. In this section, we introduce two different approaches that are being implemented in different LIREC robots. Even though they are being developed separately, the two approaches complement each other and can be easily integrated given the modularity of the three layer architecture.

3.1 User proximity control: application in Spirit of the Building Scenario

Ideally, if a user wants to engage in face to face interaction, she would face towards an interaction partner suggesting that a user is keen to interact. We use this cue to firstly find the user's face in the environment using face detection and then to drive the robot closer to the user to facilitate interaction. The face detection implementation developed by HWU is specifically designed to determine user proximity and location to enhance interaction with the user. The face detection has been implemented with the OpenCV vision library which works in real time and has the following functionalities:

- Reduces false detection by using two different face detection stages. In first stage it determines the largest probable face of an image and cuts the ROI (probable face in the image). Then, in the second stage, it re-performs face detection on the sub image.
- Calculates the approximate angles from camera focal point to detected face (X,Y degrees) – it can be used to turn the robot/character relative to the detected face.
- Calculates the approximate distance of face from camera using the difference between Face bounding Box and the image size (currently it works with 640*480 resolution with a threshold value of 297000 pixels to detect faces in a range of 80cm to 100cm).
- Detects only one face at a time; in case of multiple users, it will detect the largest/closest face to the camera.
- Can be used for user proximity distance sensing and control.

This application was used during the demo at the review in HWU 22rd April 2009 to control the robot motion and stop the robot at an appropriate distance from user while approaching the user (see Figure 11). The robot in the demo performed tasks such as finding a person (using the face detection mechanism) in a defined area avoiding obstacles and moving towards a person to greet her using text-to-speech.


Face detection: Approach Person with Proximity control (Side View)

2 Stage face detection: Robot camera view

Stage1: Find largest probable

Stage2: Face detection on ROI (cut-out probable face)

Figure 11. User location by face detection in HWU mobile robot

3.2 Vision and Sound Fusion in FLASH

FLASH (Flexible LIREC Autonomous Social Helper) is the new robotic companion that has been developed by WRUT and will be used in RobotHouse scenarios. The social robot FLASH is to be equipped with a series of sensor devices, including a set of microphones and a pan-tilt-zoom camera. These devices will be utilised to perceive the environment, herein to locate objects, persons, recognise faces. Below the FLASH active vision system component used for person location is described in more detail.

The FLASH person location mechanism makes use of two perception channels: visual and sonic. It is supported by a face recognition system and a sound source location system. The visual face recognition system is capable to detect and locate human faces in camera images. The sound location system plays an auxiliary role in the person location procedure – it provides information on sound sources placement. The fusion of vision and sonic data enhances the process efficiency. Since for typical systems the angle of hearing is much larger than the angle of view, such multimodal person location approach increases the system operational scope.

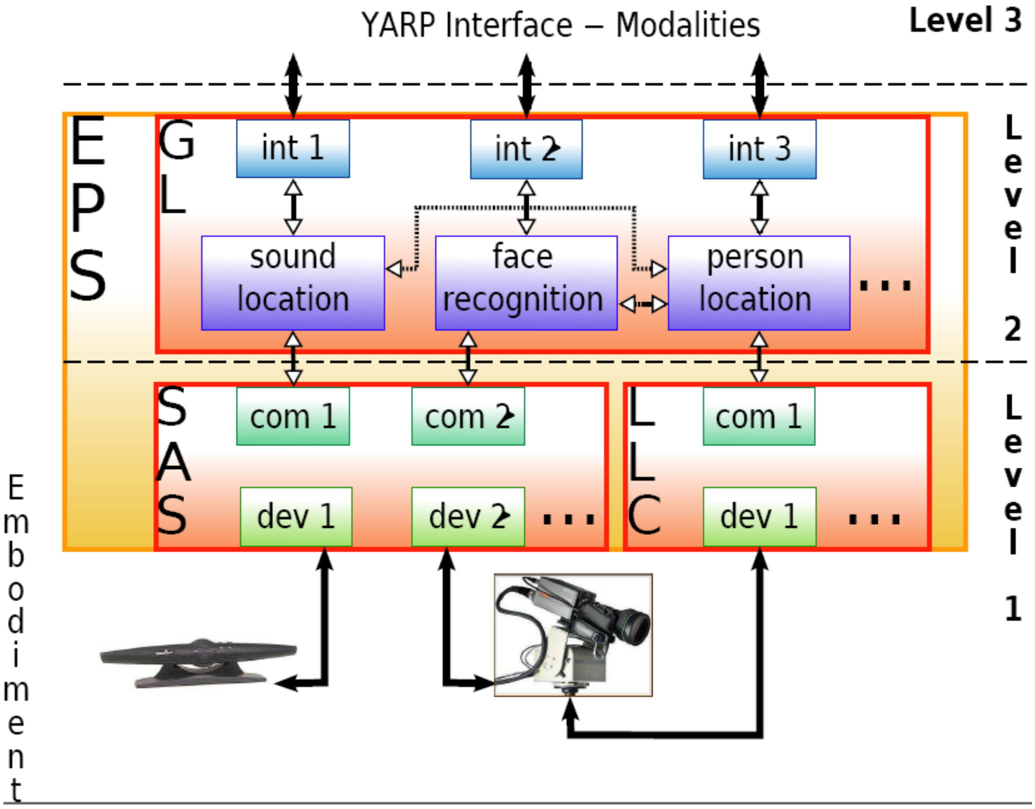


Figure 12. Structure of FLASH active vision system.

The general structure of the FLASH active vision system is shown in Figure 12. The system is a part of the robot environment perception system (EPS) and is designed as a modular, scalable unit for perceiving persons and objects. It confirms the three layered integration architecture for LIREC robotic companions, introduced in the deliverable D9.1, and constitutes its level 1 and 2 layers. In the system all the robot hardware devices can be accessed uniformly with use of level 1 sensor access system (SAS) and low level controller (LLC) units. Functionality of modalities is provided by the generalisation level (GL) and can be utilised in level 3 via the YARP interface.

For the person location modality two other modalities are employed: the face recognition and the sound source location. Cooperation between the person location, sound source location and face recognition systems is envisioned as follows:

1. The visual face recognition system continually detects and locates human faces in the camera field of view;
2. Every time a sound source is detected by the sound source location module the sound direction is transferred to PTU via the person location module;
3. The PTU sets the pan of the camera;
4. The camera is aimed at the detected face (see point 1) and zoomed to provide the required resolution of the face region (for face recognition, eyes tracking and facial expression coding).

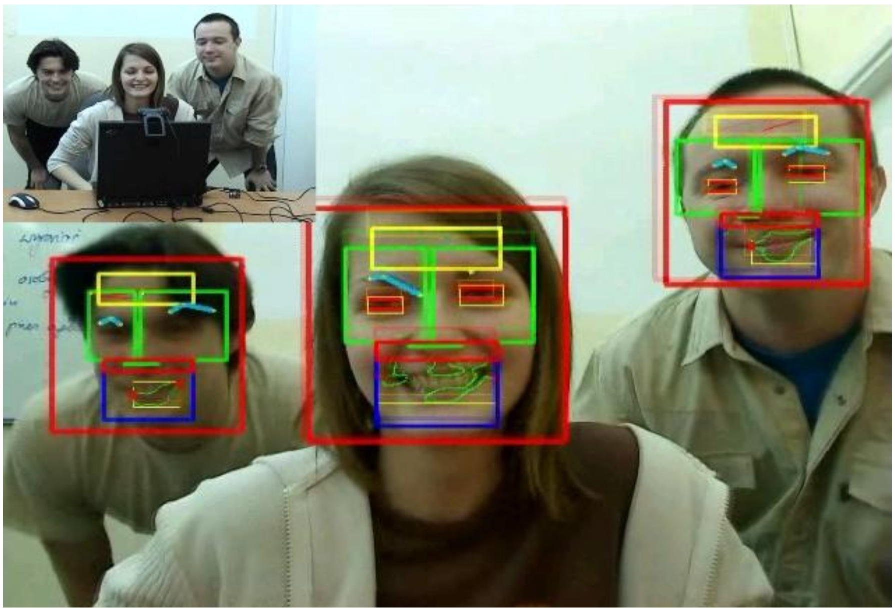


Figure 13. Exemplary results of face location and coding.

The FLASH prototype vision system for face location and facial expression coding (FacET) is based on the OpenCV library. It performs face detection using Haar classifiers and other (simpler but less accurate) methods. Particular functionalities (e.g. facial expression coding, eyes tracking) require certain, minimal resolution of the face region, therefore proper image zooming is essential. The results of the vision system working are illustrated in Figure 13. The FLASH prototype sound system for sound source location utilises the time delay of signal arrival algorithm. It calculates signal direction using the generalised cross-correlation method. To make computational process more efficient, the fast Hartley transform has been applied. The results of sound source localisation system work are illustrated in Figure 14.

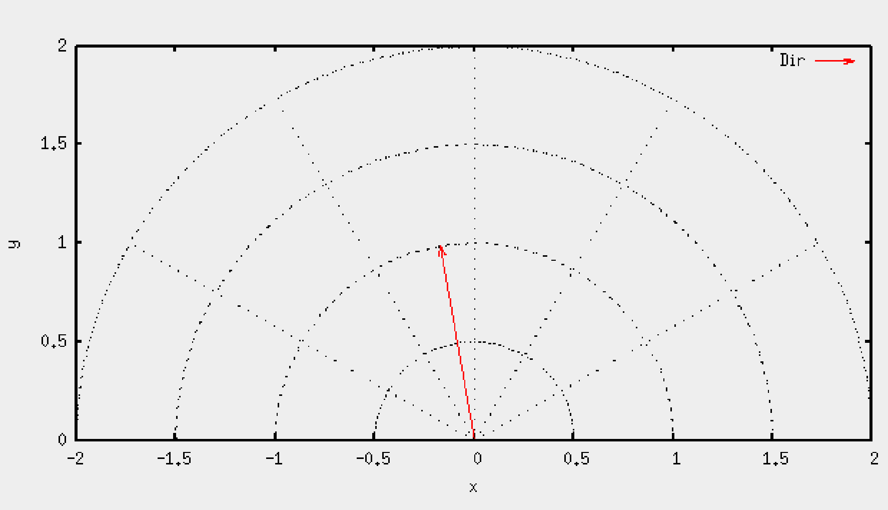


Figure 14. Exemplary results of sound source location.

4 Perceiving the user's expressions and states: a multi-level approach

The design of affect sensitive interactive companions requires the research on affect recognition to be taken beyond the state of the art. First, while the automatic recognition of more complex states has started to receive some attention only lately (Zeng *et al.*, 2009), artificial companions require an affective framework in which affect sensitivity goes beyond the ability to recognise prototypical emotions, and allows for more variegated affective signals conveying more subtle states to be captured. The design of an artificial companion would also require the development of affect recognition systems that are trained and tested with spontaneous, multimodal affective expressions.

Furthermore, the framework for perceiving the user should be able to work in the user's own settings, and this means that it should be robust to real world conditions and personalised. For these reasons, an affect recognition system for artificial companions should be designed according to a specific interaction scenario.

Affect recognition systems for socially intelligent companions must vary according to the context of interaction. In a given interaction scenario, the distance between user and companion defines the level of interaction between them and, consequently, it influences the ability and the need for the companion to detect and interpret different affective cues and states. For this reason, we propose to use a multi-level approach to analyse affective cues in human-companion interaction, in which different non-verbal behaviours of the user are analysed depending on the distance at which user and companion interact (Castellano & McOwan, 2009).

Short-range interaction

The user and the companion interact face-to-face. Examples of non-verbal cues that a companion may be sensitive to during face-to-face interactions are facial expressions, eye gaze, head gestures and orientation, posture, expressive movements. An example of system that works in a short-range interaction scenario is described in the work by Kapoor et al. (Kapoor et al., 2007), in which a user's frustration during the interaction with a learning companion is detected using multimodal non-verbal behaviour. In LIREC, analysis of cues typical of short-range interaction is applicable to the "MyFriend" scenario (e.g., children playing chess with the iCat robot).

Medium-range interaction

The user and the companion do not interact face-to-face, but the user is in the range of the companion. The main focus, at this level, is on global indicators, such as expressive body movements, that are reported to be effective for affect discrimination (Castellano et al., 2008b; Camurri *et al.*, 2003), and simple gestures and actions such as waving, approaching, withdrawing. These indicators may be indicative of the user's willingness to interact and may be appropriate to consider, for example, in the LIREC "Spirit of the Building" scenario.

Long-range interaction

The user is in the same environment, but not in the range of the companion. Coarse cues such as the presence of people in a room may be of help for the companion to determine whether it is required or not by the user. In LIREC, this level of analysis may be applicable to the "Robot House" scenario.

Note that the specifications of the verbal and non-verbal behaviours that our companions will be sensitive to, as well as of the systems to analyse such behaviours, will be the result of an iterative design process strictly depending on the interaction scenario and currently ongoing.

In the remaining of this Section we present a first experiment carried out in the "MyFriend" scenario that investigates how non-verbal behaviours displayed by the user and contextual information can model some user affective states. This experiment aims to provide the foundations for the design of an affect recognition system suitable for this specific scenario. We will use the results of this experiment to model user's engagement with the iCat using a Bayesian network.

4.1 Application in MyFriend Scenario

In order to carry out a rigorous design of an affect recognition system, it is necessary that affective cues and states that the companion must be sensitive to are defined according to a specific interaction scenario. In the "MyFriend" scenario, an iCat robot plays the role of a game companion by playing chess with children. This is a naturalistic scenario that requires a companion sensitive to affective states that are both related to the game and to the social interaction with the iCat, in order to sustain long-term interactions with users.

As a collaboration between QMUL and INESC-ID, an experiment was carried out in order to investigate the role of non-verbal behaviours displayed by children while playing with the iCat (Castellano *et al.*, 2009b) and the contextual information of the game (Castellano *et al.*, 2009a) to help in discriminating among the selected affective states: the valence of the feeling experienced by the user and her engagement with the iCat.

4.1.1 User's Affective States

Given this specific interaction scenario, an affect sensitive companion should be able to capture user states that are both related to the game and the social interaction with the iCat. The valence of the feeling experienced by the user was chosen to measure the degree to which the user's affect is positive or negative (Russel, 1980). This categorisation of affect appears to be adequate for the purpose of describing the overall feeling that the user is experiencing throughout the game.

On the other hand, the user's engagement with the iCat was chosen to describe the level of social interaction established between them. We believe that the correct interpretation of the user's level of engagement with the iCat is a factor of primary importance for the establishment of a long-term interaction. 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" (Poggi, 2007). We regard engagement with the iCat as being characterised by an affective and attention component (see (Peters *et al.*, 2008) 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 it is doing something or not and whether the iCat's behaviour or the current state of the game induce a positive or a negative feeling in the user.

4.1.2 Non-verbal behaviours

Given that this is a short-range interaction scenario, where the user and the companion interact face-to-face, the non-verbal behaviours that the framework should consider may include user's facial expressions and head movements. Following this line of thought, and considering the behaviours displayed more often by the children while playing chess with the iCat (see Figure 15), the selected non-verbal behaviours were the following:

- Looking at the iCat
- Looking at the iCat after own move
- Looking at the iCat after iCat move
- Looking at the iCat during the game

- Looking at the chessboard
- Looking elsewhere
- Smiling
- Mouth fidget
- Hand on mouth
- Scratching face or head
- Blinking while looking at the iCat
- Raising eyebrows
- Approaching
- Moving away

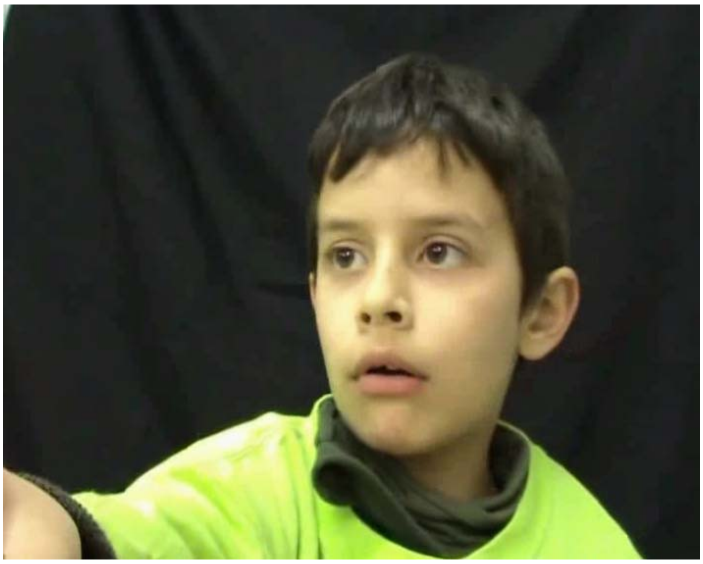


Figure 15. Example of non-verbal behaviour displayed by one of the participants while interacting with the iCat.

4.1.3 Contextual information of the game

In chess, like in most of the games, there is a lot of contextual information that one can retrieve from the game itself that may be relevant to model some of the user's states. The choice of using contextual features of the game in this particular scenario relies on the assumption that, when the user is playing with the iCat, most of the experienced affective states may be related to the events happening in the game or to the behaviour and expressions exhibited by the robot. Following we describe each one of the contextual features selected for investigation in this scenario.

Game state
It is a value that represents the condition of advantage/disadvantage of the user in the game. This value is computed by the same chess evaluation function that the iCat uses to plan its own moves. 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.

Game evolution

Game evolution is given by 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 (although without necessarily being in a condition of advantage), while a negative value means that the user's condition is getting worse with respect to the previous move.

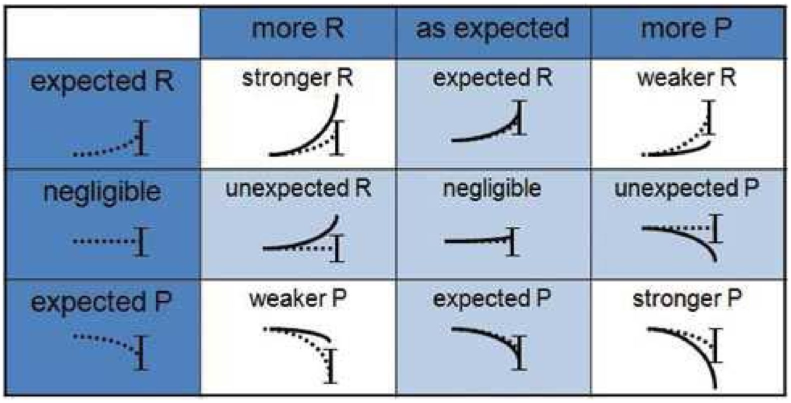
Captured pieces

This feature indicates if, in the last move played by the user and the iCat, a piece was captured either by the user or the iCat.

User sensations

User sensations are calculated through the same mechanism used by the iCat to generate its own affective reactions, but from the user's perspective, i.e., taking into account the user's game state.

The iCat’s affective reactions are determined by the *emotivector* (Martinho & Paiva, 2006), an anticipatory system that generates an affective signal resulting from the mismatch between the expected and the sensed values of the sensor to which it is coupled to. In this case, the emotivector is coupled to the values of the chess evaluation function that determines the moves played by the iCat. After each move played by the user, the chess evaluation function returns a new value, updated according to the current state of the game. The emotivector captures this value and, by using the history of evaluation values, an expected value is computed applying the moving averages prediction algorithm (Hannan, 1985). Based on the mismatch between the expected and the actual sensed value (the last value received from the evaluation function), the system generates one out of nine affective signals for that perception (Leite *et al.*, 2008), as presented in Figure 16.



	more R	as expected	more P
expected R	stronger R	expected R	weaker R
negligible	unexpected R	negligible	unexpected P
expected P	weaker P	expected P	stronger P

Figure 16. Emotivector model: the first row displays the possible outcomes when reward (R) is expected, the second row shows the possible outcomes when the state is expected to remain the same and the third row contains the possible outcomes when punishment (P) is expected. Depending on whether the sensed value is higher, within, or lower than a confidence interval, the affective signal will belong to the first column (more reward), second (as expected) or third (more punishment).

For example, after three moves in the chess game, if the iCat has already captured an opponent's piece, it might be expecting to keep the advantage in the game (i.e., expecting a reward) after the user's next move. Therefore, if the user makes a move that is even worse than the one the iCat was expecting (e.g., by putting her queen in a very dangerous position), the generated affective signal will be a “stronger reward”, which means “this state of the game is better than what I was expecting”. Now consider the same example, but from the user’s perspective: after three moves in the game the user has lost one piece, so she might expect the iCat to continue holding the advantage (i.e., the user is expecting a “punishment”). If the user plays a terrible move, she might be experiencing something closer to a “stronger punishment” sensation.

By computing the emotivector sensations from the user’s perspective, we attempt to predict the user's possible sensations throughout the game.

iCat's facial expressions

As described earlier, after each move played by the user, the iCat displays a facial expression generated by the emotivector system. We used animations from the iCat's animation library, since they have been previously submitted to tests ensuring that users perceive those emotional expressions in the iCat's embodiment (Bartneck, 2004). We have made a correspondence between those animations and the nine emotivector sensations, as depicted in Table 2. The emotional animations were not enough to map all the sensations. Thus, we adapted other animations (such as apologize or confirm) to some of the less intense sensations. The animations marked with "*" were slightly modified in order to recreate an emotion. The choices for the sensation-animation mapping were based on the meaning of the sensations. For example, the "stronger reward" sensation means that we experienced a reward much better than we were expecting and therefore the correspondent emotion is "excitement".

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

Table 2. Mapping between the nine emotivector sensations and the iCat animations

4.1.4 Experiment and data collection

An experiment was performed at a primary school where once a week children have two hours of chess lessons[1]. Five eight year old children (3 male and 2 female) took part in the experiment. Each participant played two different exercises, one with low and one with medium difficulty. By using different levels of difficulty in each exercise we expected to elicit different types of affective states and behaviours in the children. The exercises consisted of middle game chess positions chosen by a chess instructor who was familiar with each student's chess skills. The children were told to play and try to win against the iCat. The robot begins the interaction by inviting the user to play. The user is always the first to play a move. After each move played by the user the iCat asks her to play its move, as its embodiment does not allow it to do so itself. Each interaction session ends when the user completes both exercises (i.e., by winning, loosing or drawing).

For each exercise, a log file containing the contextual features was created in real-time during the interaction. After each user move, a new entry was written in the log, containing the time since the beginning of the interaction (for synchronisation purposes) and the current value of the contextual features: the game state from the user's perspective, the game evolution, whether there were pieces captured either as a consequence of the current user move or the upcoming iCat move, the user sensations, and the upcoming iCat facial expression.

[1] Two other experiments were conducted at chess clubs, but for this particular evaluation those subjects were not considered.

All interaction sessions were recorded with three video cameras: one capturing the frontal view (e.g., the face of the children), one the side view and one the iCat. The videos recorded with the frontal camera were annotated in terms of user states and contextual features by three annotators. The annotation was based on the behaviour displayed by the children and the state of the game. 72 video segments were selected starting from the 10 collected videos (two for each participant) using ANVIL, a free video annotation tool (Kipp, 2008). Each video segment had a duration of approximately 7 seconds. Before starting the annotations, the annotators agreed on the meaning of each label to describe the user's state in the videos. As for the valence of the feeling, annotators could choose one out of three options: "positive", "negative" and "cannot say". To describe the engagement with the iCat, annotators could choose among "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 annotation processes were then compared for each video segment: a label was selected to describe the state 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". From the annotation process, we randomly selected 15 video segments labelled as "positive", 15 as "negative", 15 as "engaged with the iCat" and 15 as "not engaged with the iCat". Each group of videos contains 3 samples for each participant.

After annotating the affective states, a similar process was adopted to annotate the non-verbal behaviours displayed by the children while interacting with the iCat. Based on their occurrence during the interaction with the robot, the following non-verbal behaviours were identified: *Looking at the iCat*, *Looking at the chessboard*, *Looking elsewhere*, *Smiling*, *Mouth Fidget*, *Hand on mouth*, *Scratching face or head*, *Blinking while looking at the iCat*, *Raising eyebrows*, *Approaching*, *Moving away*. Contextual information such as the phase of the game was also considered to define when children were looking at the iCat. This process generated additional non-verbal behaviours for the annotation: *Looking at the iCat after own move*, when the iCat generates an affective reaction; *Looking at the iCat after iCat move*, when the user receives feedback from the robot, such as approval or disapproval; and *Looking at the iCat during the game*, when the user is thinking and the iCat is performing idle behaviours such as blinking and looking sideways. Two coders annotated the portions of video segments in which the children exhibited these behaviours. For each video segment, each behaviour was assigned a value given by the average number of frames that that specific behaviour was displayed in the video segment.

The extracted contextual features were also added to the annotation of each video. To associate the appropriate log entry to each video, we selected the one that occurred within or right before the video segment.

4.1.5 Results and discussion

4.1.5.1 Non-verbal behaviours

Two coders annotated the videos where the non-verbal behaviours were displayed: 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.

Statistical analysis was performed to identify which behaviours allow for a discrimination of the identified affective states in this interaction scenario.

Two repeated measures t tests (N = 15) were performed for each behaviour to explore whether there is a significant difference in the occurrence of the selected behaviour between positive and negative and between engaged with the iCat and not engaged with the iCat samples. In each test the type of behaviour was considered as the dependent variable, while the valence of the feeling or the engagement with the iCat as the independent variable (two levels). A summary of the results is reported in Table 3.

Results highlight a main effect of the valence of feeling on the following non-verbal behaviours: *Looking at the iCat*, *Looking at the iCat during the game* (i.e., when the user is thinking and the iCat is performing idle behaviours such as blinking and looking sideways), *Looking at chessboard* and *Smiling*. When the feeling is positive, the children tend to look at the iCat more overall [$t(14) = 3.84$; $p < 0.001$] and during the game [$t(14) = 1.91$; $p < 0.05$], they smile more [$t(14) = 4.51$; $p < 0.001$] and look at the chessboard less [$t(14) = -5.03$; $p < 0.001$] than when their feeling is negative.

As far as the engagement with the iCat is concerned, a significant effect emerged on the following behaviours: *Looking at the iCat*, *Looking at the iCat after own move* (i.e., immediately after the user's move, when the iCat generates an affective reaction), *Looking at the chessboard*, *Smiling* and *Blinking while looking at the iCat.* In case of engagement with the iCat, children tend to look more at the iCat overall [$t(14) = 5.88$; $p < 0.001$] and after their own move [$t(14) = 4.77$; $p < 0.001$], they smile more [$t(14) =2.68$; $p < 0.01$], and they look at the chessboard less [$t(14) = -5.64$; $p < 0.001$] than when they are not engaged with the iCat. While results show that children blink more while looking at the iCat when they are engaged with it [$t(14) = 2.20$; $p < 0.05$], this result is not very significant, given that most of the times that the children look at the iCat they are engaged with it.

Non-verbal behaviour	**Valence of feeling**	**Engagement with the iCat**
Looking at the iCat	**p < 0.001**	**p < 0.001**
Looking at the iCat after own move	N.S	**p < 0.001**
Looking at the iCat after iCat move	N.S.	N.S.
Looking at the iCat during the game	**p < 0.05**	N.S.
Looking at the chessboard	**p < 0.001**	**p < 0.001**
Looking elsewhere	N.S.	N.S.
Smiling	**p < 0.001**	**p < 0.01**
Mouth fidget	N.S.	N.S.
Hand on mouth	N.S.	N.S.
Scratching face or head	N.S.	N.S.
Blinking while looking at the iCat	N.S.	**p < 0.05**
Raising eyebrows	N.S.	N.S.
Approaching	N.S.	N.S.
Moving away	N.S.	N.S.

Table 3. Summary of results of the two t-tests performed for each non-verbal behaviour (N.S. = not significant).

4.1.5.2 Contextual information of the game

Statistical analysis was also performed to study if and how the selected contextual features co-occur with the user's states identified in our interaction scenario (i.e., *feeling* and *engagement with the iCat*). In this section we present the results obtained for each one of the contextual features.

Game State

In order to investigate whether there is a significant difference between the means of the game state in correspondence with the two conditions for valence of feeling (*positive* and *negative*) and engagement with the iCat (*engaged with the iCat* and *not engaged with the iCat*), two repeated measures *t* tests (N=15) were performed for the game state (the dependent variable) with the valence of the feeling and the engagement with the iCat as the independent variables (two levels) respectively in the first and in the second test. It was predicted that higher values of the game state would be observed for *positive* and *engaged with the iCat* samples, and that the game state would better differentiate *positive* from *negative* than *engaged with the iCat* from *not engaged with the iCat*.

As far as the valence of feeling is concerned, results show that there is a significant difference between the means of the game state: when the feeling is positive the values of the game state are higher than when the feeling is negative [$t(14) = 3.61$; $p < 0.01$]. Regarding the engagement with the iCat, as expected, the values of the game state are significantly higher when the children are engaged with the iCat than when they are not [$t(14) = 2.43$; $p < 0.05$]. Figure 17 (a) and Figure 17 (b) show the error bar graphs that illustrate the confidence intervals for each sample mean. Note that game state discriminates better between *positive* and *negative* than between *engaged with the iCat* and *not engaged with the iCat* samples: this is also in line with our hypothesis, as it was expected that the game state would have greater effects on the valence of the feeling than the engagement with the iCat.

Summary of results: when the user is winning, her feeling tends to be positive and her level of engagement with the iCat is higher.

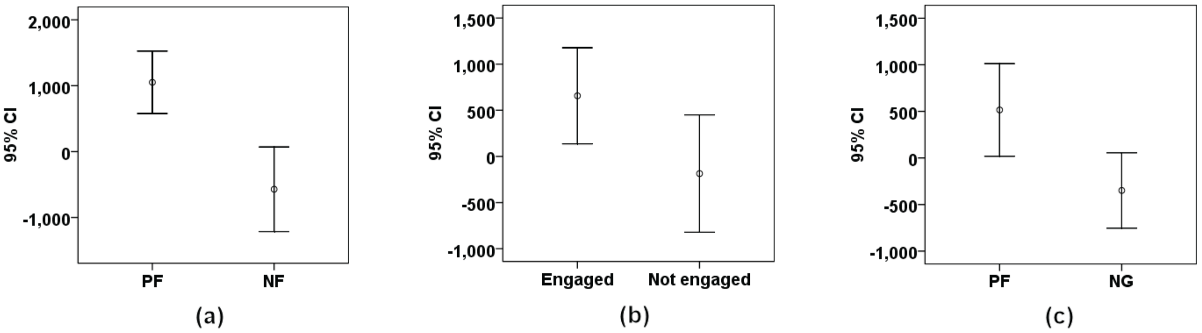


Figure 17. Error bar chart for game state and feeling (a), game state and engagement (b) and game evolution and feeling (c). PF = positive feeling; NF = negative feeling.

Game evolution

To investigate whether there is a significant difference between the means of the game evolution in correspondence with the two conditions for valence of feeling and engagement with the iCat, two repeated measures *t* tests (N = 15) were performed for the game evolution (the dependent variable) with the valence of the feeling and the engagement with the iCat as the independent variables (two levels) respectively in the first and in the second test. As for the game state, each t-test was based on fifteen samples for each level of the independent variables. It was predicted that higher values of the game evolution would be observed for the samples annotated as *positive* and *engaged with the iCat* and that the game evolution would better differentiate *positive* from *negative* than *engaged with the iCat* from *not engaged with the iCat* samples.

In terms of the valence of feeling, when the feeling is positive the values of the game evolution are significantly higher than when the feeling is negative [$t(14) = 2.62$; $p < 0.01$]. Figure 17 (c) shows the error bar graph that illustrates the confidence intervals for each sample mean. As far as the engagement with the iCat is concerned, no significant difference was found between the means of the game evolution. First of all, even though this result does not confirm our hypothesis, it was expected that the game evolution would have greater effects on the valence of the feeling than the level of engagement with the iCat. Furthermore, improvement in the game does not necessarily mean that the user is winning: the user can improve from one move to another while at the same time being in an overall condition of disadvantage. On the other hand, while it was found that both game state and game evolution significantly affect the valence of the feeling ($p < 0.01$), the computed significance was different ($p = 0.002$ for the game state and $p = 0.010$ for the game evolution).

Summary of results: when the user is improving her condition in the game, her feeling tends to be more positive than negative.

Captured pieces

In order to establish whether there is an association between the pieces captured during the game and the feeling and engagement with the iCat, two chi-square tests were carried out by considering only the video samples where the event of capturing pieces was present. Feeling (two levels, N = 10) and engagement with the iCat (two levels, N = 8) were considered as the influencing variables respectively in the first and the second test and the pieces captured during the game (two levels: *captured by the user*, *captured by the iCat*) as the dependent variable in both tests. It was expected that the captured pieces would influence the feeling and that pieces *captured by the user* would be associated with a positive feeling, while pieces *captured by the iCat* with a negative feeling.

Results show that there is no significant association between engagement with the iCat and the event of capturing pieces, but a significant association was found between the latter and the feeling [$X^2 = 4.29$, $df = 1$, $p < 0.05$]: as expected, when the feeling is positive most of the times the user has captured a piece (60%), while when it is negative it is more likely that a piece has been captured by the iCat (100%).

Summary of results: when the user captures a piece during the game her feeling is more likely to be positive, while when a piece is captured by the iCat the user's feeling is more likely to be negative.

User sensations

To study the association between user sensations and feeling and engagement with the iCat two chi-square tests were performed. Feeling and engagement with the iCat were regarded as the influencing variables (two levels, N = 30) for the first and second test respectively and the user's sensations (eight levels) as the dependent variable in both tests. It was expected that an association would be found between the valence of the feeling and specific categories of the sensations: *expected P*, *expected R*, *stronger P*, *stronger R*, *unexpected P*, *unexpected R*, *negligible* and *none.* No samples in our corpus of videos included instances of the categories *weaker P* and *weaker R*, so that the latter were not considered in the analysis.

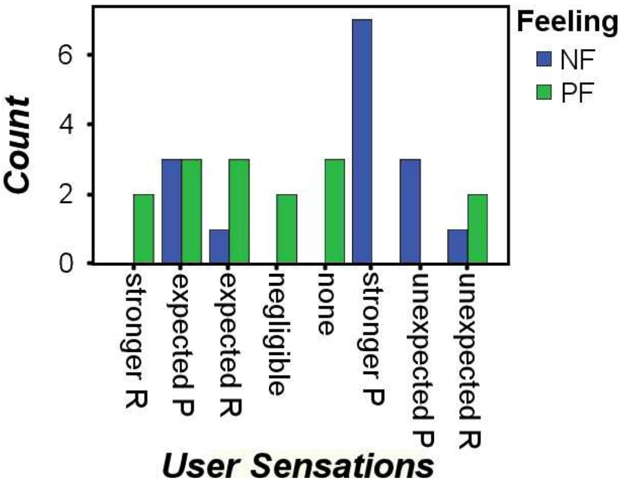


Figure 18. Clustered bar chart for user sensations and feeling (PF = positive feeling; NF = negative feeling). The graph shows that most of the times the feeling is positive rather than negative in case of expected reward (*expected R*, 3 counts vs. 1), stronger reward (*stronger R*, 2 counts vs. 0) and unexpected reward (*unexpected R*, 2 counts vs. 1), while the feeling is more likely to be negative rather than positive in correspondence with stronger punishment (*stronger P*, 7 counts vs. 0) and unexpected punishment (*unexpected P*, 3 counts vs. 0).

Results show a significant association between the valence of the feeling and the user sensations [X^2 = 18.33, *df* = 7, *p* < 0.05]. Figure 18 shows how the frequencies of the different categories of the user's sensations are linked to the feeling. The feeling is more likely to be positive rather than negative for conditions of reward (20% vs. 6.7% for *expected R*, 13.3% vs. 0% for *stronger R* and 13.3% vs. 6.7% for *unexpected R*) rather than punishment, which are associated more with a negative feeling rather than with a positive feeling (46.7% vs. 0 % for *stronger P* and 20% vs. 0% for *unexpected P*). No difference between positive and negative feeling was found for expected punishment (*expected P*, 20% vs. 20%). Furthermore, when the user expects the state of the game not to change from one move to another (*negligible*) the feeling is more likely to be positive (13.3 % vs. 0%) and the same occurs in correspondence with particular circumstances of the game where no expectation is involved, such as the very beginning of the game before the user plays her first move, when the iCat gives up and when the user wins, looses or draws (*none*, 20% vs. 0%). No significant association was found between the engagement with the iCat and the user's sensations: this suggests that the level of engagement of the user is not necessarily related to any specific condition of expectation in the game.

Summary of results: in correspondence with conditions of reward most of the times the user's feeling is positive, while in correspondence with conditions of punishment the feeling is more likely to be negative.

iCat facial expressions

Two chi-square tests were carried out to investigate whether the generation of a facial expression by the iCat would influence the feeling or the level of engagement of the user. In these tests, the presence of a facial expression generated by the iCat was used as the influencing variable (two levels, N = 30) and the feeling and the engagement with the iCat (two levels) as the dependent variables for the first and second test respectively. It was hypothesised that the generation of a facial expression by the iCat would increase the level of engagement.

Results confirm the hypothesis: when the iCat generates a facial expression it is more likely that the user is engaged with it (86.7%), while when no facial expression is displayed, most of the times the user is not engaged with the iCat (53.3%) [X^2 = 5.4, *df* = 1, *p* < 0.05]. No significant association between the presence of a facial expression and the feeling was found. Two chi-square tests were also carried out to investigate whether the generation of a specific facial expression by the iCat would influence the feeling or the level of engagement, but no significant association was found.

Summary of results: when the iCat displays a facial expression during the game, the level of engagement of the user towards the iCat increases.

4.1.6 Conclusions of this study

The main contribution of this study consists of exploiting different interaction modalities to model naturalistic user’s states that originate both from the task that the user is accomplishing (playing chess in this case) and the social interaction with the robot. Contextual information related to the game, the behaviour displayed by the iCat and the user’s non-verbal behaviours were analysed as possible indicators to be used for discriminating among the identified states.

In terms of contextual features, results highlight a key role of game state, game evolution, event of capturing pieces and user sensations (as a result of the mismatch between their expectations and what really happened in the game) to discriminate between positive and negative feeling (i.e., the user’s feeling tends to be more positive than negative when the user is winning or improving in the game, when a piece is captured by the user instead of by the iCat and in case of conditions of reward rather than punishment during the game). Game state and the display of facial expressions by the iCat proved successful to predict the level of engagement with the iCat: higher levels of engagement with the iCat were found when the user is winning in the game and when the iCat displays facial expressions.

As for the non-verbal behaviours, the study identified the behaviours displayed by the children that allow for a better discrimination of the valence of the feeling experienced by the children during the game and their level of engagement with the iCat. The results show that the behaviours that are mainly affected by the feeling and the engagement with the iCat are eye gaze and smiles. These findings will provide the foundation for the design of an affect recognition system for a game companion in the “MyFriend” scenario.

4.2 Modelling user engagement with a Bayesian approach

Based on the interaction scenario and the results presented in the previous sections, we designed a Bayesian approach to detect the user's engagement with the iCat robot (Castellano *et al.*, 2009c). 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.

Results reported in the previous sections show 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. Results also show a correlation between the level of engagement of the user with the iCat and 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:

User behaviour:

- User looking at the iCat
- User smiling

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 19 shows the identified cause-effect relationships in our interaction scenario.

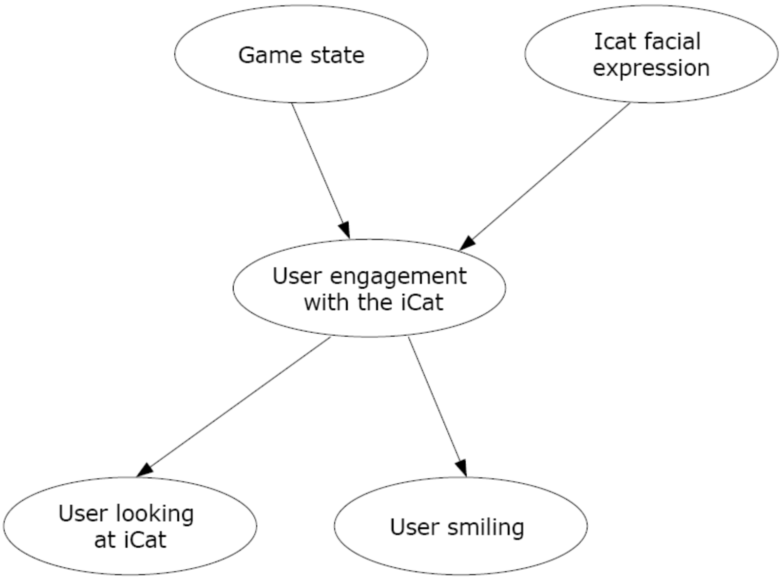


Figure 19. Cause-effect relationships between user engagement with the iCat and task and social interaction-based features in our scenario.

4.2.1 Experimental results

To obtain the experimental data used to train and evaluate our model, the data collection and procedure described in Section 4.1.4 was extended to include 8 subjects and more video samples. 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.

4.2.1.1 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.

Non-verbal behaviour

As described at the beginning of Section 4.2, eye gaze and smiles of the user were considered as indicators of the user's engagement with the iCat. Two coders annotated the portions of video segments in which these behaviours were displayed by the children based on the same procedure described in Section 4.1.4.

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**: As described in Section 4.1.3, the game state represents the condition of advantage/disadvantage of the user in the game
- **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. 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 where 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 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. 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, looses 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.

4.2.1.2 Multimodal fusion

The multimodal fusion step aggregates the user's behavioural features and the contextual information. 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.

4.2.2 Evaluation results and conclusion

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. 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 subjects. 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.

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. Table 4 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. 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.

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 4. Recognition rates and ROC area values for the different classifiers.

5 Conclusions

In this document we reported the work that has been conducted so far towards the development of a joint framework for the perception of user actions in LIREC scenarios. In terms of technology to support the perception of user's facial expressions, body movements and speech, currently we have:

- FacET, a library for the automatic extraction of facial features;
- An identification module that distinguishes users by their faces;
- A system for approach/withdraw detection;
- Two automatic speech recognition systems.

These modules have been developed using open source libraries such as OpenCV. We are using this technology as the ground for the development of competencies that perceive meaningful user activity for the LIREC scenarios. We started by focusing on two crucial actions: locating the user and affect sensitivity. Locating the user is extremely relevant in mobile robot scenarios, as it is a pre-requisite that the companion must fulfil for establishing further interactions with the user, and recognize other actions that require low proximity distances. Affect sensitivity is also very important because it supports the development of social relations between the user and the companion.

Although these two actions might seem very distinct, the ground technology that they use is similar, which highlights the importance of having a common set of basic modules (facial feature extraction, body posture detection, etc...) that can be shared by different systems responsible for perceiving distinct user actions.

However, some aspects may reduce the common applicability of the framework. First of all, there are user actions that are context dependent, and require more specific modalities than user's verbal and non-verbal behaviour, for instance the chess moves played by the user in the "My Friend" scenario. Furthermore, even among the same scenarios, some modules might be more suitable for certain embodiments than others: as a companion may take different forms (e.g. it may present itself as a robot, a mobile handheld device or a graphical character), these different forms may determine, or even restrict, the different modalities for recognizing user actions. All these aspects are being carefully analysed and will be addressed in further development of the framework.

6 References

Abad, A. and Neto, J. (2008). Incorporating Acoustical Modelling of Phone Transitions in a Hybrid ANN/HMM Speech Recognizer. In Proc. *INTERSPEECH 2008*, Brisbane, Australia.

Bartneck, C., Reichenbach, J. And Breemen, A. (2004). 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*, Ankara, Turkey.

Breazeal, C. (2003). Emotion and sociable humanoid robots. E. Hudlicka (Ed.), International Journal of Human-Computer Studies, 59(1-2), pp. 119-155.

Camurri, A., Lagerlöf, I. and Volpe, G. (2003) Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 213–225.

Castellano, G., Aylett, R., Paiva, A. and McOwan, P. W. (2008a). Affect recognition for interactive companions, Workshop on Affective Interaction in Natural Environments (AFFINE), ACM International Conference on Multimodal Interfaces (ICMI'08), Chania, Crete, Greece, 24 October.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A. and McOwan, P. (2009a). "It's All in the Game: Towards an Affect Sensitive and Context Aware Game Companion". (Accepted for publication in the *International Conference on Affective Computing & Intelligent Interaction – ACII 2009*)

Castellano, G., Leite, I., Pereira, A., Paiva, A. and McOwan, P. (2009b). Affect Recognition for Interactive Companions: Challenges and Design in Real World Scenarios. (Submitted to special issue of the *Journal on Multimodal User Interfaces*, Springer)

Castellano, G., McOwan, P. W. (2009). Analysis of affective cues in human-robot interaction: A Multi-level approach, 10th International Workshop on Image Analysis for Multimedia Interactive Services, Special Session on Affective Interaction in Natural Environments, London.

Castellano, G., Mortillaro, M., Camurri, A., Volpe, G. and Scherer, K. (2008b). Automated Analysis of Body Movement in Emotionally Expressive Piano Performances, *Music Perception*, vol. 26, no. 2, pp. 103–119.

Castellano, G., Pereira, A., Leite, I., Paiva, A. and McOwan, P. (2009c). Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features. (Submitted to the 11th *International Conference on Multimodal Interfaces – ICMI 2009*)

Chen, L. *et al*. (2005). A robust algorithm for eye detection on gray intensity face without spectacles. *Journal of Computer Science and Technology*.

Dautenhahn, K. (2007). Socially Intelligent Robots: Dimensions of Human-Robot Interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), pp. 679-704.

Hannan, E., Krishnaiah, P. and Rao, M. (1985). *Handbook of Statistics 5: Time Series in the Time Domain*. Elsevier.

Huang, Y., Chiang, C. (2006). A rule-based real-time face detector. Department of Computer Science and Information Engineering, National Dong-Hwa University, Shoufeng, Hualien.

Intel. OpenCV Library, http://www.intel.com/technology/computing/opencv/index.htm

Kapoor, A., Burleson, W. and Picard, R.W. (2007). Automatic prediction of frustration, *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736.

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

Leite, I., Pereira, A., Martinho, C. and Paiva, A. (2008). 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.

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

Meinedo, H., Caseiro, D. and Trancoso, I. (2003). AUDIMUS.media: a broadcast news speech recognition system for the european portuguese language. In Proc. *6th International Workshop on Computational Processing of the Portuguese Language - PROPOR 2003*, Faro, Portugal.

Namysl, M. (2008). Vision system in human emotions recognition. Master's Thesis, Institute of CECR WRUT, Wroclaw (in Polish).

Peters C., Asteriadis, S., Karpouzis, K. and de Sevin, E. (2008). 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.

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

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

Walker, W. *et al*. (2004). Sphinx-4: A Flexible Open Source Framework for Speech Recognition. In SMLI, SUN MICROSYSTEMS INC. Technical report TR2004-0811.

Watson: Head tracking and gesture recognition library.
http://projects.ict.usc.edu/vision/watson

Zeng, Z., Pantic, M., Roisman, G.I., and Huang, T.S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence.

7 Appendix: List of WP3 Submitted Papers

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., McOwan, P. (2009). “It’s All in the Game: Towards an Affect Sensitive and Context Aware Game Companion”.
(Accepted for publication in the International Conference on Affective Computing & Intelligent Interaction – ACII 2009)

Leite, I., Martinho, C., Pereira, A., Paiva, A. (2009). “As Time Goes by: Long-term Evaluation of Social Presence in Robotic Companions”.
(Accepted for publication at the 18th IEEE International Symposium on Robot and Human Interactive Communication – RO-MAN 2009)

Castellano, G., Leite, I., Pereira, A., Paiva, A., McOwan, P. (2009). “Affect Recognition for Interactive Companions: Challenges and Design in Real World Scenarios”.
(Submitted to special issue of the Journal on Multimodal User Interfaces, Springer)

Castellano, G., Pereira, A., Leite, I., Paiva, A., McOwan, P. (2009). “Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features”.
(Submitted to the 11th International Conference on Multimodal Interfaces – ICMI 2009)

Leite, I., Pereira, A., Martinho, C., Paiva, A., McOwan, P., Castellano, G. (2009). “Towards and Empathic Chess Companion”. Workshop on Empathic Agents, AAMAS’09, Budapest, Hungary.