



## Deliverable 3.3

***"Design and implementation vision system"***

Contract number: **FP7-215554 LIREC**

Living with Robots and intEractive Companions

Start date of the project: 1<sup>st</sup> 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

<b>Project ref. no.</b>	<b>FP7-215554</b>
<b>Project acronym</b>	LIREC
<b>Status &amp; version</b>	"D3.3"
<b>Contractual date of delivery</b>	June 2010
<b>Actual date of delivery</b>	
<b>Deliverable number</b>	D3.3
<b>Deliverable title</b>	"Design and implementation vision system"
<b>Nature</b>	Other
<b>Dissemination level</b>	PU
<b>WP contributing to the deliverable</b>	
<b>WP / Task responsible</b>	WP3
<b>Editor</b>	Ana Paiva
<b>Editor address</b>	INESC-ID / Instituto Superior Técnico - Tagus Park Av. Prof. Dr. Cavaco Silva, 2780-990 Porto Salvo, Portugal
<b>Author(s) (alphabetically)</b>	Ginevra Castellano, Sibylle Enz, Márta Gácsi, Dave Griffiths, Iolanda Leite, Peter W. McOwan, Ana Paiva, André Pereira, Caroline Spielhagen, Marek Wnuk
<b>EC Project Officer</b>	Pierre-Paul Sondag
<b>Keywords</b>	Vision-based social perception, face detection and tracking, facial features detection and tracking, expression recognition, affect recognition, human-full body movement analysis
<b>Abstract (for dissemination)</b>	This deliverable aims to provide a survey of the vision-based social perception abilities that are currently under investigation in the LIREC project. These are the result of the collaboration among different partners in LIREC and they are applied to different scenarios. We provide a description of the first prototypes of these abilities, highlighting relevant issues and open questions, and we discuss them with reference to different scenarios.

# CONTENTS

- 1 Introduction
- 2 Psychological background
  - 2.1 Affective intelligence and empathy
  - 2.2 Social motives for getting involved in a relationship
- 3 Social perception: ethological considerations
  - 3.1 Visual communication
  - 3.2 Individual recognition
- 4 Social perception abilities in the LIREC scenarios
- 5 Face detection and tracking
  - 5.1 Colour-based face tracking
  - 5.2 Face tracking based on particle filters
  - 5.3 Approach / withdraw detection using face detection and tracking
- 6 User recognition
- 7 Facial features detection and tracking
  - 7.1 The FacET library
    - 7.1.1 Library concept
    - 7.1.2 Face detection and segmentation
    - 7.1.3 Facial features coding
  - 7.2 Tracking facial features with particle filters
- 8 Affect sensitivity in the “My Friend” scenario
  - 8.1 User’s affective states
  - 8.2 System overview
  - 8.3 Relevant contextual information and non-verbal behaviours
  - 8.4 Initial framework to model user engagement
  - 8.5 The Inter-ACT corpus
    - 8.5.1 Annotation
  - 8.6 A prototype system for smile detection
- 9 Automatic analysis of full-body movement: motion direction detection
  - 9.1 Application in the LIREC scenarios
- 10 Conclusions

# 1 Introduction

A vital requirement for artificial companions is the ability to perceive and interpret social, affective expressions and states of humans, so as to be able to engage in and behave appropriately during sustained social interactions (Breazeal, 2003). A fundamental component in these abilities is the automatic analysis and interpretation of social signals and human affective behaviour (Vinciarelli et al. 2009; Zeng et al., 2009) from sensory input.

One of the objectives of WP3 is to design and evaluate a social, affective perception framework for artificial companions. The key idea is to endow artificial companions with the ability to perceive application-dependent users' states and expressions so that they can use this information to plan how to start, maintain or improve the interaction with the users. We refer to these abilities as *affect sensitivity* and, in general, *social perception* (Castellano et al., 2010), as they concern the way social affective cues displayed by the user can be used to infer higher level information such as affective and mental states.

The scenarios investigated in the LIREC project are quite numerous and diverse and they involve different types of robots interacting in different ways with human users. This has a relevant impact on the choice of user states and expressions that a companion should be sensitive to and on the implementation of automatic methods for their analysis and interpretation. First of all, in LIREC we are considering a multi-level approach, that is, we are focusing on different cues and expressions depending on the physical distance between user and robot: human-companion interactions in LIREC are classified as short-range interaction (e.g., "My Friend" scenario), medium- or long-range interaction (e.g., Spirit of the Building and Robot House scenarios) (Castellano & McOwan 2009). The task the companion and the user are involved in is also important to determine which user states and expressions a companion should be sensitive to, as well as the behaviour displayed by the companion. In LIREC we refer to all this as contextual information and we adopt the strategy of (1) designing social perception and affect recognition abilities based on the role played by context and (2) using contextual information in addition to users' behaviour to improve automatic affect recognition performances.

This deliverable aims to provide a survey of the vision-based social perception abilities that are currently under investigation in the LIREC project. These are the result of the collaboration among different partners in LIREC and they are applied to different scenarios. We provide a description of the first prototypes of these abilities, highlighting relevant issues and open questions, and we discuss them with reference to different scenarios. An example of contextualised design of social perception and affect recognition framework is also reported for the "My Friend" scenario. Note that many of these abilities are still work-in-progress and represent first prototypes that are currently under test in the LIREC scenarios with the objective to be improved and account for a larger spectrum of variables related to social perception in human-companion interaction.

This document is organised as follows: Section 2 provides an overview of the psychological background behind the design of social perception abilities for artificial companions, while Section 3 includes some ethological considerations related to social perception in human-companion interaction; Section 4 presents an overview of the social perception abilities and their application in the LIREC scenarios; Section 5, 6 and 7 describes vision-based abilities applicable to face-to-face interaction and Section 8 presents a case study of affect sensitivity in the "My Friend" scenario; Section 9 addresses social perception abilities related to full-body movement analysis and their application in the LIREC scenarios; Section 10 provides a summary of the open questions and challenges in social perception for artificial companions and describes some of the ethical issues related to social perception that may arise in the LIREC scenarios.

## 2 Psychological background

### 2.1 Affective intelligence and empathy

How do people interact with others in social contexts? In long-term social interactions, empathy was shown to be related to communication styles, conflict resolution and relationship satisfaction (Franzoi et al., 1985; Long & Andrews, 1990; Davis & Kraus, 1991): in successful long-term social interactions, people try to be empathic, i.e. to understand the thoughts and feelings of others and take this understanding into consideration when they interact with them. Empathy as a psychological concept embraces two perspectives (Davis, 1996; Holz-Ebeling & Steinmetz, 1995): cognitive and affective empathy. Cognitive empathy can be defined as understanding a target person's feelings and thoughts in a given situation, whereas affective empathy refers to emotional reactions in the empathiser due to the perceived feelings and thoughts of a target person.

The empathic process itself is based on social perception of the target person, i.e. the perception of cues that stem from the target person's expression (mimic, gesture, posture, paraverbal cues) or from the situation the target person acts in (Bischof-Köhler, 1994). Thus, social perception is one core antecedent of empathic processes. On the other hand, empathic processes influence the reactions towards an interaction partner and thus link social perception with actions in social contexts.

Empathy in companions helps them to engage realistically and believably in social interaction, since it allows them to dynamically understand the internal processes in the user, e.g. emotions and expectations, and to react accordingly.

### 2.2 Social motives for getting involved in a relationship

Why do people interact with others? According to Fiske (2004) social motives help us to survive in the environment. According to her "we are motivated to get along with other people because it is adaptive to do so" (Fiske, 2004, p.14). Fiske addresses five social motives (belonging, understanding, controlling, self-enhancing and trusting) that are orientated toward making people fit better in social groups.

The motivation for getting involved in a relationship in the first place is largely affected by the way people perceive each other, e.g. the attention for and the processing of information about others. This so called social perception (Greenberg & Baron, 2000) starts with the first impression of a person and accompanies almost all further social interactions.

Two motives that are in particular related to social perception are "understanding" and "trust": for example, "**understanding**" motivates us to make sense of our surroundings and to predict what will happen in case of uncertainties. Since "sharing the same frames of reference facilitates interaction with other people" (Fiske, 2004, p. 87), individuals prefer to develop meanings that are shared with other people. The "**trusting**" motive involves confidence or faith that some other, upon whom we depend, will not act in a way that causes us painful consequences (Boon & Holmes, 1991). In principle people trust other people and expect them to be basically benign. People are biased to see the best in others and are basically motivated to restore their sense that the world is trustworthy. According to Yamagishi and Yamagishi (1994) trust is a form of social intelligence that facilitates attachment and interdependence in close relationships.

From a psychological point of view companions should address some of the above mentioned social core motives (Fiske, 2004) in order to promote the development of human-companion long-term relationships. The motives "understanding" and "trusting" are very much concerned with *attraction*, as a desire for a voluntary relationship (Huston & Levinger, 1978). Individuals are attracted by those they feel they can understand easily and by which

fulfil their expectations, features that in technical contexts are subsumed – among others – under the term “usability”.

### **3 Social perception: ethological considerations**

Humans and dogs are able to communicate in particular situations without specific prior training of the dog. Dogs have a large set of signals, for example for affective communication (to express their inner states), and also can perceive and comprehend several human-specific signals such as pointing gestures (Lakatos et al. 2009) or verbal commands.

#### **3.1 Visual communication**

There are many experimental evidences that one of the most typical behavioural elements of dog-human interactions is the seeking for and maintaining eye contact (Miklósi et al. 2003). This trait probably evolved during domestication to serve the purpose of attention getting and communication with humans.

The ability to recognise the other's attention is especially important in the visual modality of communication when the orientation of the receiver is crucial. Thus in case of visual signals the sender should actively modify his/her own behaviour to become the focus of the other's attention by either producing attention-getting signals or moving into the visual field of the receiver. The ability of dogs to use behavioural and/or facial cues to detect of human attention has been proved in several studies. Gácsi et al. (2004) investigated the ability of dogs to recognize human attention in different experimental situations: (a) facing versus not facing human; (b) visible versus non-visible human eyes. Results showed that the efficiency of dogs to discriminate between *attentive* and *inattentive* humans depended on the context of the test (game or task), but they definitely could rely on the orientation of the body and the orientation of the head. There were also indications that dogs were sensitive to the visibility of the eyes because they showed increased hesitant behaviour when approaching a blindfolded owner, and they also preferred to beg from the person with visible eyes. Virányi et al. (2004) assessed the dogs' responsiveness to their owner's tape-recorded verbal commands. Results showed that dogs were more ready to follow the command if the person attended them during instruction compared to situations when she faced a human partner or was out of sight of the dog. Importantly, dogs showed intermediate performance when the person was orienting towards the wall during the replayed verbal commands. This suggests that dogs are able to differentiate the focus of human attention.




#### **3.2 Individual recognition**

There are multiple factors that a domestic animal could use to recognize its human handler including face recognition, speech patterns, olfactory signals, and movement style. In contrast to cats, dogs are able to discriminate their owner from another human based solely upon face recognition (Lomber & Cornwell, 2005). In another study Adachi and colleagues (2009) showed that in dogs the sound of an owner's voice conjures up a mental image of the owner's face, and this leads to confusion when another face appears instead

### **4 Social perception abilities in the LIREC scenarios**

As previously mentioned in the Introduction Section, in LIREC we adopt a multi-level and context-based approach in the design of social perception abilities.

Figure 1 shows an overview of the social perception abilities currently under investigation and development in the LIREC scenarios. The next sections will provide a description of the vision-based social perception abilities and examples of their application in the LIREC scenarios, with the exception of the use of face detection for user proxemics in the Spirit of the Building and Robot House scenario, which are very closely linked to WP6 and will be reported in the next WP6 deliverable ([add ref](#)), and the user activity recognition ability, which is not vision-based.

<p style="text-align: center;"><b>My Friend</b></p> 	<p style="text-align: center;"><b>Spirit of the Building</b></p> 	<p style="text-align: center;"><b>Robot House</b></p> 
<ul style="list-style-type: none"> <li>- Face detection and tracking</li> <li>- User recognition</li> <li>- Facial features detection and tracking</li> <li>- Expression recognition</li> <li>- Affect recognition</li> <li>- Motion direction detection</li> </ul>	<ul style="list-style-type: none"> <li>- Face detection (e.g., user proxemics)</li> <li>- Motion direction detection</li> </ul>	<ul style="list-style-type: none"> <li>- Face detection (e.g., user proxemics)</li> <li>- User activity recognition (GEO system)</li> </ul>

**Figure 1: Social perception abilities currently under investigation in the LIREC scenarios.**

## 5 Face detection and tracking

The face detection capability in the LIREC scenarios is based on the OpenCV method, which consists of a cascade of boosted classifiers using Haar-like features (Bradski & Kaehler, 2008). The final classifier consists of many simple classifiers that are built using boosting techniques. The classifier, which is pre-trained with positive samples (images of faces) and negative samples (arbitrary images) of the same size, can be applied to a region of interest in an input image and classify it as “Face”, if the region is likely to contain a face, or “Not face” otherwise. A search window is moved across the image and every location is checked using the classifier. The classifier can be resized in order to be able to find objects

of interest at different sizes (see Bradski & Kaehler, 2008 for more details on the OpenCV Haar-based classifier).

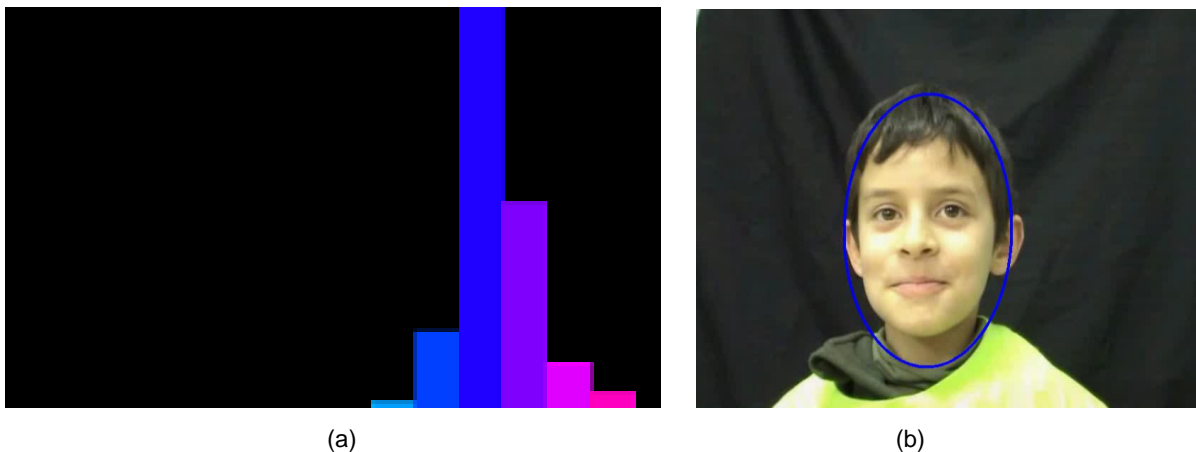
This method is quite robust to the presence of lateral and rotated faces if a classifier trained with lateral face images is also used in addition to the one trained with frontal faces, although this adds computational complexity. A drawback of this method is the generation of false positives in complex backgrounds, which requires constraints to be added.

When a face is detected in the scene, a tracking of the face bounding box can be performed.

### 5.1 Colour-based face tracking

QMUL developed a module for face tracking based on the Camshift algorithm (Bradski & Kaehler, 2008), which tracks a combination of colours. When a face is detected in the scene, the tracking is automatically initialised using the face bounding box returned by the Haar classifier. In short, the Camshift algorithm creates a colour histogram to represent the face (see Figure 2a), calculates a face probability for each pixel of the input image, shifts the location of the face region at each video frame and adjusts the size and angle of the face rectangle each time it shifts it. Figure 2b shows an example of output of face tracking based on the Camshift algorithm.

Colour-based tracking has the advantages of being fast and lightweight and of being able of dealing with lateral and rotated views of the face, but, since it is based on the tracking of a combination of colours, is not very robust in complex backgrounds including skin-like coloured objects, and hence it requires some constraints to be adopted in the scenario. Moreover, problems may arise if hands are placed in front of the face and, since the neck is also of the same colour of the face, it is usually segmented with the face.



**Figure 2: Example of colour histogram used by the Camshift algorithm to represent a face (a); example of output of the Camshift algorithm (b).**

### 5.2 Face tracking based on particle filters

QMUL and FoAM jointly developed a software module for face tracking using an approach based on estimators and, specifically, on a category of estimators called particle filters. The idea behind the use of estimators is to estimate motion in a way that makes the most out of the measurements provided by the sensors (Bradski & Kaehler, 2008). By making use of several measurements it is possible to extract motion information that does not arise from noise. An important characteristic of estimators is the need for a model of the motion to be analysed. Given this information, estimators work in two main phases: a prediction phase, where information learned in the past is used to further refine the model, and a correction

phase, in which a measurement is made and that measurement is further reconciled with the model's prediction.

Particle filters are estimators that allow for a variable of interest to be tracked as it evolves over time and, in comparison with other estimators such as the Kalman filter, present the advantage of being able to represent non-gaussian and multimodal distributions.

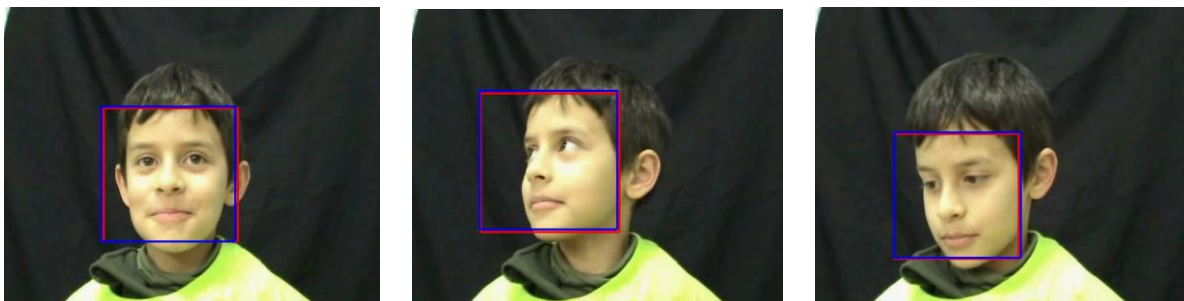
Particle filters are characterised by a sample-based representation of the pdf, with multiple copies (or particles) of the variable of interest. Each particle has a weight, which is related with the quality of the particle. The algorithm is recursive and consists of a prediction and an update phase (Rekleitis, 2004):

**Prediction phase:** each particle is modified according to the existing model and random noise is added

**Update phase:** each particle's weight is re-evaluated based on the latest available sensory information

At times the particles with infinitesimally small weights are eliminated, a process called *resampling*. The estimate of the variable of interest can be obtained using different methods, such as *weighted mean*, *best particle*, or *robust mean*. The weighted mean has the disadvantage of failing with multimodal distributions; the best particle introduces a discretization error; the robust mean computes the weighted mean in a small window around the best particle and represents the best method, although it is computationally expensive (Rekleitis, 2004).

Tracking of the face bounding box requires the x and y coordinates of the face bounding box extremes. These values represent our variable of interest or particle. Figure 3 shows an example of results of tracking based on particle filters.



**Figure 3: Example of output of tracking based on particle filters. The red bounding box represents the face detected by the Haar classifier, while the blue bounding box represents the face tracked by the particle filter.**

### 5.3 Approach / withdraw detection using face detection and tracking

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 4 and Figure 5).

This system runs on Windows and Linux. If used in the “My Friend” scenario, it could be useful for the detection of “approach” and “withdraw” head gestures of children playing with the iCat robot.

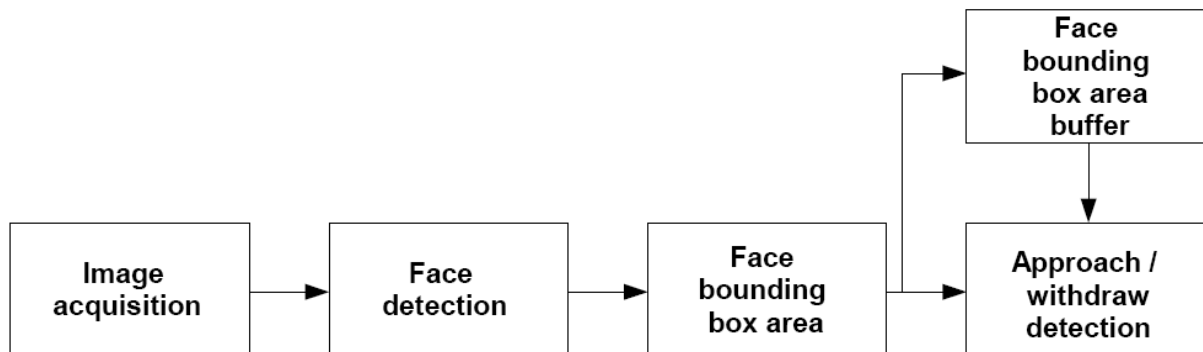


Figure 4. Main components of the system for approach / withdraw detection.



Figure 5. From left to right, user is staying still, approaching the camera and withdrawing.

## 6 User recognition

FoAM extended its user identification competency (see Deliverable D3.2 for more details on the first prototype) with a new recognition algorithm based on eigenface (Turk & Pentland, 1991) subspace modelling. The YARP interface and usage is identical to the previous sum of squared difference image algorithm, while internally the user's face is now parameterised and compared against other user's face parameters created with a face space generated by processing the Spacek<sup>1</sup> face database (comprising 395 individuals) using PCA. This approach was benchmarked against the previous SSD algorithm in faces with varying lighting conditions using the Yale Face Database B<sup>2</sup> and found to be more robust for user classification with respect to lighting conditions. Figure 6 shows that while lighting changes

<sup>1</sup> <http://cswww.essex.ac.uk/mv/allfaces/index.html>

<sup>2</sup> <http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html>

have an effect on the parameters generated by the face space, overlap is minimal, and the identity is still classifiable. Some additional work was also carried out to prototype generating eigenface parameters to differentiate and classify user expressions using the same approach.

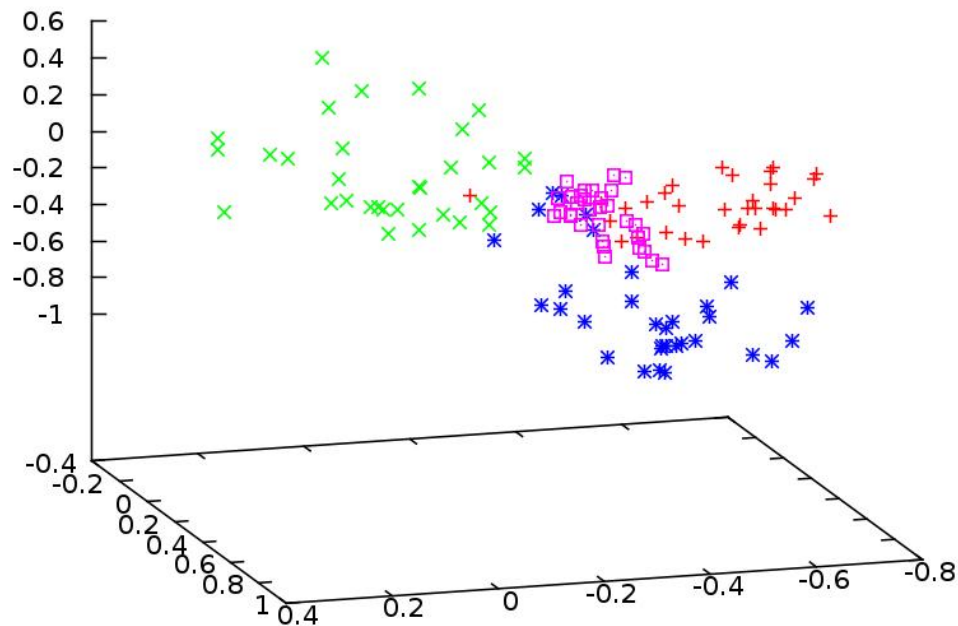


Figure 6: An example of plotting face images against 3 dimensions of the face appearance model. The points represent images of 4 individuals in many different lighting conditions.

## 7 Facial features detection and tracking

Facial features detection is based on the FacET library, developed by WRUT, while facial features tracking is done using an approach based on particle filters, jointly developed by QMUL and FoAM.

### 7.1 The FacET library

FacET (Facial Expression Tracker) is a library of image processing procedures, designed for detecting and parameterising face components (e.g., eyes, eyebrows, lips, forehead wrinkles). The parameters are derived from the face image by measuring certain lengths, areas and angles (e.g. the distance between the eyelids, the area of visible teeth, declination of the eyebrows). The extracted feature values form a vector, describing a facial expression. This vector is suitable for variety of pattern recognition methods (e.g. k-nearest neighbours, naive Bayes, neural networks, fuzzy classifiers). The extracted face features can be useful for classification of human emotions based on facial expression (Namysl, 2009) and facial action coding system (FACS) (Ekman & Friesen, 1978).

The current version of FacET is based on the application created for the Master thesis: "Vision system in human emotions recognition" (in Polish) by Marcin Namysl<sup>3</sup>. It is free software in the sense of the GNU General Public License version 3.

### 7.1.1 Library concept

The main public function of the library is face features detection:

```
void detectFeat(IplImage *src, IplImage *dst);
```

It fills the facesList - list of facepar\_t structures, describing selected features for all the faces detected in the source image.

The structure:

```
std::list<facepar_t> facesList;
```

describes the detected face features in the following order:

```
typedef struct {
    double roix, // face ROI upper left corner X coordinate [pixels]
           roiy, // face ROI upper left corner Y coordinate [pixels]
           angle, // face declination angle (not verified, for future use)
           LEbBnd, // left eyebrow bend angle (top)
           LEbDcl, // left eyebrow declination angle (side)
           LEyOpn, // distance between the right eyelids (relative to the eyeball
subregion)
           LEbHgt, // distance between left pupil and eyebrow top (relative to the
eye subregion)
           REbBnd, // right eyebrow bend angle (top)
           REbDcl, // right eyebrow declination angle (side)
           REyOpn, // distance between the right eyelids (relative to the eyeball
subregion)
           REbHgt, // distance between right pupil and eyebrow top (relative to the
eye subregion)
           LiAspt, // aspect ratio of the lips bounding box (percents)
           LLiCnr, // Y position of the left corner of the lips (relative to the lips
bounding box)
           RLiCnr, // Y position of the right corner of the lips (relative to the
lips bounding box)
           Wrnkls, // number of horizontal wrinkles in the center of the forehead
           Nstrls, // nostrils baseline width (relative to the face width)
           TeethA; // area of the visible teeth (relative to the lips bounding box)
} facepar_t;
```

Several parameters can be controlled in the application:

- face\_detection\_scale (set by setFaceScale()),
- eyebrow\_proportions (set by setEyebrowsRatio()),
- lips\_proportions (set by setLipsRatio()).

The preset values are used by Haar classifier and the histogram-based threshold calculation procedure, used by the eyebrows and the lips extraction. Haar cascades for the face and its regions are loaded with read\*Cascade().

---

<sup>3</sup>Supervised by Marek Wnuk, Wrocław University of Technology, Institute of Computer Engineering, Control and Robotics

### 7.1.2 Face detection and segmentation

Three main tasks are performed:

- detect/locate a face (find face ROI),
- locate face subregions (eyes, mouth, nose, forehead)
- extract and measure selected face features (members of `facepar_t` structure).

Face detection is based on Haar classifier and makes use of available Haar cascades for faces: `haarcascade_frontalface_alt.xml`<sup>4</sup>. It is implemented using the OpenCV library (Bradski & Kaehler). The `cvHaarDetectObjects()` function returns a list of faces detected in the input image.

For each face in the list, the subregions (eyes, forehead, nose, mouth) are detected and parameterised.

#### Eyes

The subregions of eyes are found with intensity and gradient profiles (Chen et al., 2005) (Figure 7)<sup>5</sup>. The maximum of the horizontal profile of the gradient image defines the central line of the eyes. The maxima of the vertical profile define left and right eye region boundaries. The vertical central line is defined by the maximum of the vertical profile of the intensity image.

This method returns two separate rectangles for both eyes and is much faster than Haar classifier (Namysl, 2008; see Figure 8).

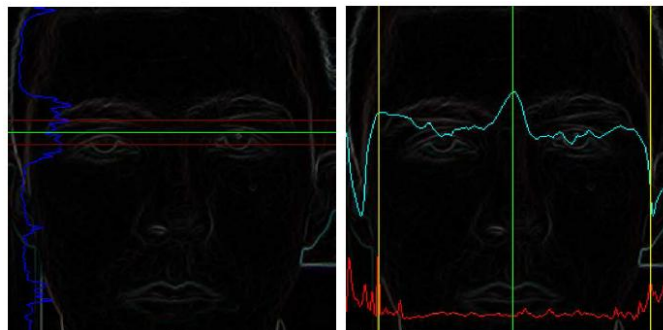


Figure 7: Finding eyes subregions with the gradient profiles.

---

<sup>4</sup> For this and the following Haar classifiers see: Reimondo A.. Haar cascades, <http://alereimondo.no-ip.org/OpenCV/34>

<sup>5</sup>All the figures courtesy the author of (Namysl, 2008)

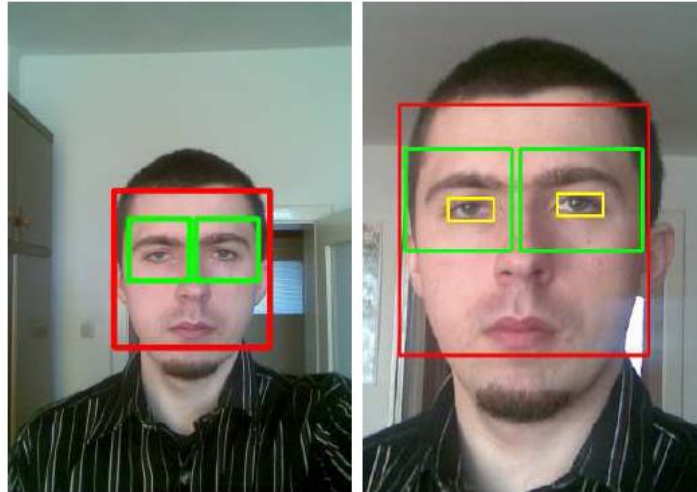


Figure 8: Eyes subregions (left) and bounding boxes of the eyeballs (right).

### Eyeballs

Assuming that the subregions of the eyes are detected, the eyeballs are located inside (Figure 8).

Two methods are selectable with `bool eyeballsHaar:`

- Haar cascade for eyeball bounding box (`thing_shan_eye_cascade.xml` cascade).
- template matching (`FastMatchTemplate()`)<sup>6</sup>

### Eyebrows

The eyebrows detection is performed in the upper part of the eye region (above the eyeball bounding box). It uses hysteresis thresholding (Namysl, 2008) and morphological filtering (region boundary propagation) (Namysl, 2008) in order to extract single dark blob representing the eyebrow silhouette. The leftmost, the topmost, and the rightmost pixels of this silhouette form a two-segment broken line representing the eyebrow shape (Figure 9).



Figure 9: Eyebrow shape representation.

<sup>6</sup> Georgiou T., Fast Match Template, <http://opencv.willowgarage.com/wiki/FastMatchTemplate>

## Nose

Two methods can be selected with `bool noseHaar:`

- Haar cascade for nose region `Nariz.xml`. It is a time consuming method.
- Setting the rectangle between eyes and mouth subregions, width equal to mouth subregion width. It is much faster, but depends on detection of eyes and mouth subregions.

## Mouth

Two methods can be selected with `bool mouthHaar:`

- Haar cascade for mouth region `Mouth.xml`. It is relatively robust, but rather slow.
- gradient projections (Namysl, 2008). It is much faster.

## Forehead

The forehead subregion is defined as a rectangle over the top of the eyebrows, with (experimentally selected) left, right and top margins with respect to the face ROI. Its detection obviously depends on successful detection of at least one eyebrow (Figure 10). Figure 10 also shows the detected face subregions.

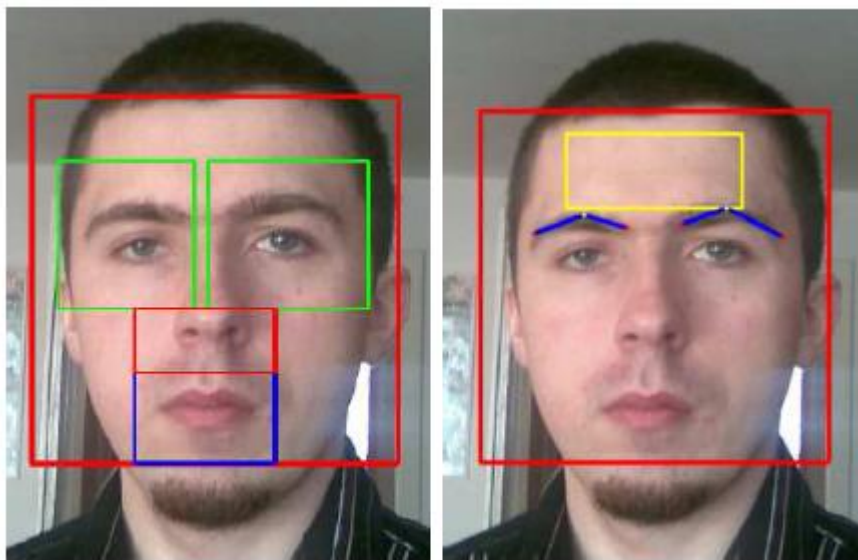


Figure 10: Primary subregions and the forehead area.

### 7.1.3 Facial features coding

The `facepar_t` structure members are calculated as geometric parameters of the extracted face features.

#### Eyebrow shape

[LR]EbBnd - Left/Right Eyebrow Bend - the angle "A" at the top of the eyebrow line (Figure 11)

[LR]EbDcl - Left/Right Eyebrow Declination - the less of the declination angles ("B", "C") of the two-segment eyebrow model (Figure 11).

[LR]EbHgt - Left/Right Eyebrow Height - vertical distance "d" between the centre of the eyeball bounding box and the eyebrow top (Figure 11). It is normalised to the eye subregion height.

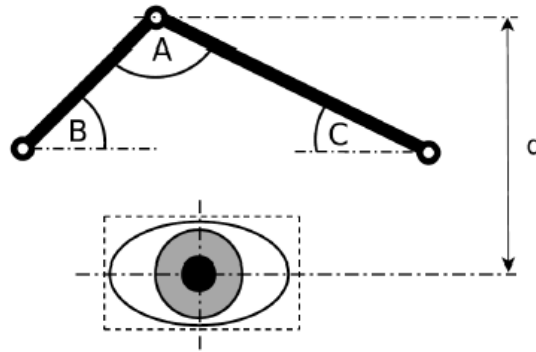


Figure 11: Eyebrow features coding.

## Eye opening

[LR]EyOpn - Left/Right Eye Opening - vertical distance between the eyelids. It is measured as the vertical size of the dark area found inside the eyeball bounding box (Figure 12) It is normalised to the eyeball bounding box height.



Figure 12: Eye opening measurement.

## Lips shape

The lips bounding box is found with the CCM (Chromatic Curve Map) (Eveno et al., 2001), which intensifies the contrast between the skin and the lips (Figure 13 b). Hysteresis thresholding around the mean region intensity, morphological filtering and gradient are used for the lips contour detection (Figure 13 c).

LiAspt - Lips Aspect ratio - aspect ratio (width/height) of the lips bounding box (Figure 13 d),

[LR]LiCnr - Left/Right Lips Corner position - vertical position of the leftmost/rightmost pixel of the lips silhouette within the lips bounding box, normalised to the bounding box height (Figure 13 d).

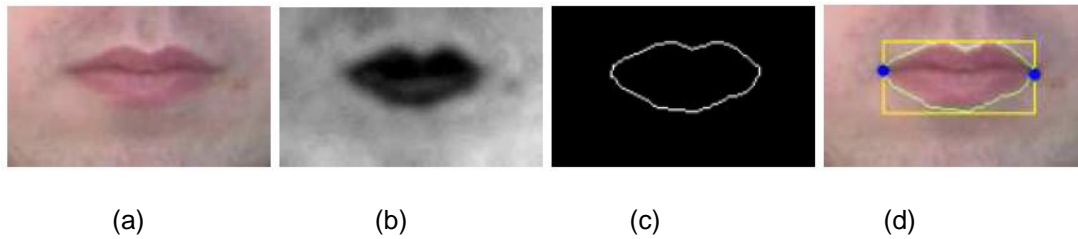


Figure 13: Lips detection and coding

### Teeth visibility

TeethA - Teeth Area - the ratio of visible teeth area to the lips bounding box area. The teeth are detected as the white region inside the lips area. Hysteresis thresholding around the mean region intensity and morphological filtering (boundary propagation and removal) are used for the detection (Figure 14).



Figure 14: Teeth area detection

### Nostrils width

Nstrls - the width of the Nostrils base line. The line is found with horizontal and vertical projections of the gradient image of the nose subregion (Figure 15). Its length is normalised to the face ROI width.

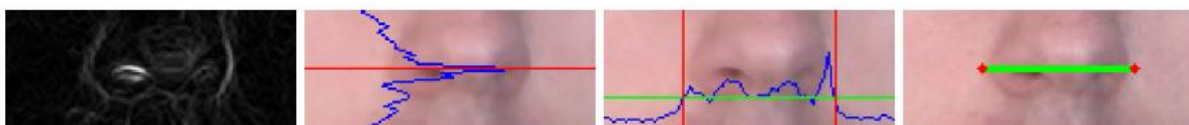


Figure 15: Nostrils baseline definition

## Forehead wrinkles

`Wrnkls` - number of horizontal wrinkles in the forehead area. The wrinkles are counted as straight lines found by the Hough transform in the vertical gradient image of the forehead subregion obtained by the Sobel convolution kernel (Figure 16).

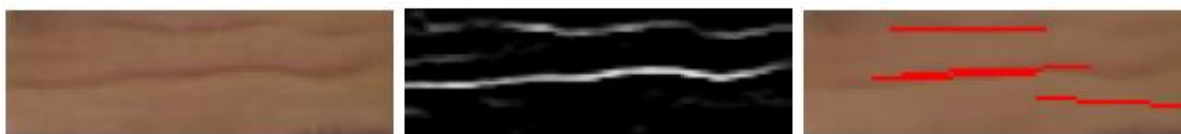


Figure 16: Forehead wrinkles detection

The results of face features detection is presented in Figure 17.

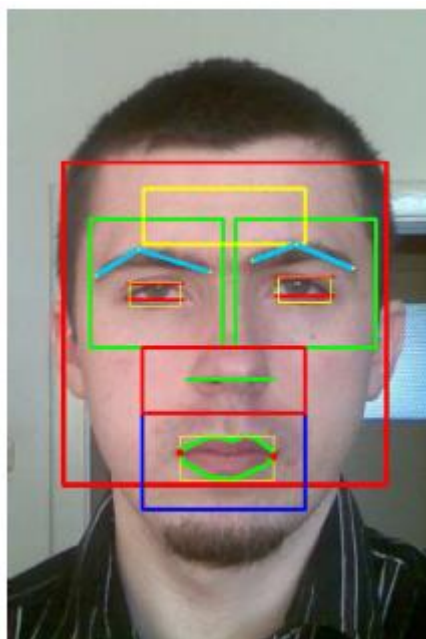
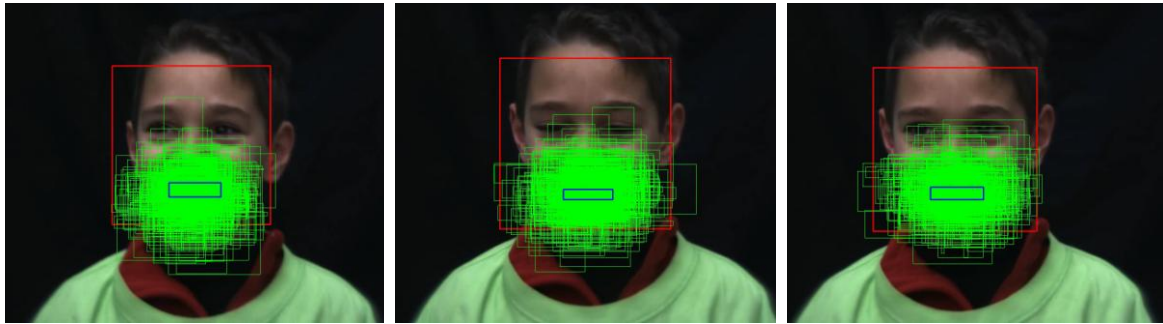


Figure 17: Presentation of the extracted face features

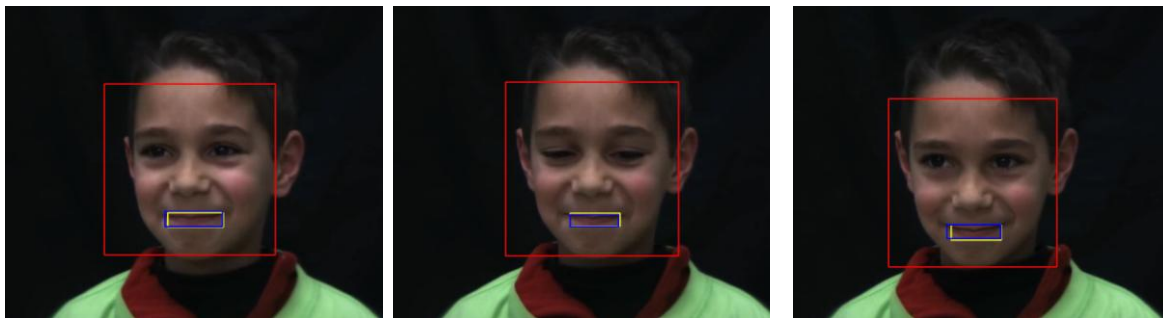
## 7.2 Tracking facial features with particle filters

Facial features tracking is performed using an approach based on particle filters (for more details see Section 5.2). Here the variable of interest (particle) is represented by the x and y coordinates of the facial features' bounding boxes. Figures 18 and 19 show different outputs of the tracking based on particle filter.

A module for the bounding boxes' check verifies that the current position and size of the bounding boxes are not too dissimilar from the previous ones.



**Figure 18. Multiple copies (particles) of the mouth's bounding box (in green) and final estimate (the blue bounding box).**



**Figure 19. Examples of output of mouth tracking based on particle filters. The yellow bounding box is the result of the detection based on the FacET library, while the blue bounding box represents the mouth bounding box tracked by the particle filter.**

## 8 Affect sensitivity in the “My Friend” scenario

One of the most relevant prerequisites of companionship is the ability to sustain long-term interactions. To do so, it is important that robot companions are endowed with a module for affect recognition that analyses both multimodal behavioural cues of the user and information about the context in which the interaction takes place.

In the “My Friend” scenario, the iCat robot acts as a game companion, helping children to improve their chess skills. While playing with the iCat, children receive feedback on their moves through the robot's facial expressions that are generated by an affective system influenced by the state of the game. The iCat's affective system is self-oriented, which means that when the user makes a good move, the iCat exhibits sad facial expressions, and when the user makes a bad move, it expresses positive reactions. However, this may change with the introduction of an affect sensitivity competence; for example, the robot may inhibit some of its happy expressions when it detects that the user's affective state is negative, or propose a new game if the user is not engaged. Therefore, by including an affect sensitivity competence in this scenario, we expect that children perceive the iCat as more caring and empathic, and consequently that this will influence in a positive way the possible long-term social relationship established between them.

In the following sections we present QMUL and INESC-ID joint work aiming at designing and evaluating an affect sensitivity competence in the “My Friend” scenario.

## 8.1 User's affective states

In the "My Friend" scenario the companion must be sensitive to user states that are both related to the game and the social interaction with the iCat. We identified the following user states as the most important to detect to improve the interaction with the companion:

**Valence of feeling:** 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 (Russell, 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.

**Interest towards the iCat:** the user's interest towards the iCat was identified to describe in what measure the user pays attention to the iCat over time (Peters et al. 2010). This, together with contextual information, could be used to detect the user's level of engagement with the iCat.

**Engagement with the iCat:** the user's engagement with the iCat was chosen to describe the level of social interaction established between them. We follow the description by Poggi (Poggi, 2007), who defined engagement 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".

## 8.2 System overview

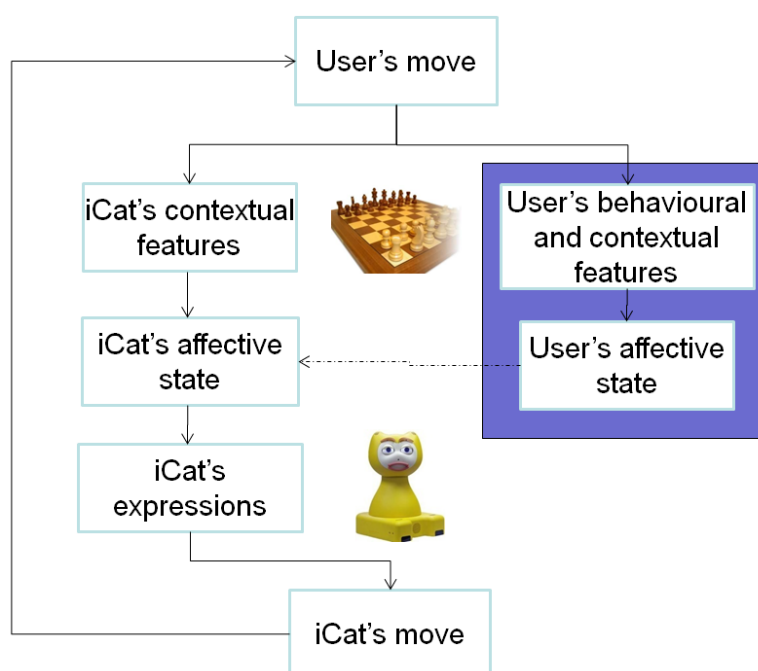
The iCat's affective state is determined by the *emotivector* system (Martinho & Paiva, 2006), an anticipatory system that generates an affective signal resulting from the mismatch between an expected and a sensed value. In this case, the emotivector is coupled to the values of a chess evaluation function that evaluates the quality of the moves played by the user, and determines the iCat's next move. After every move played by the user, the chess evaluation function returns a new value updated according to the current state of the game. The emotivector system captures this value and, by using the history of evaluation values, an expected value is computed (applying the moving averages prediction algorithm). 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). The nine possible outcomes of the emotivector system, and consequent facial expressions of the iCat, are described in Table 1.

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

**Table 1. Outcomes of the emotivector system, and the corresponding affective facial expressions displayed by the iCat.**

For instance, 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".

Figure 20 presents an overview of what happens internally during the users' interaction with the companion. When the user plays a new move, the iCat uses the contextual information of the game from its "own" perspective to update its affective state (particularly, the state of the game given by the chess heuristic function). With the introduction of an affect sensitivity competence, the iCat's affective state will be influenced not only by the contextual information from the iCat's perspective, but also from the user's affective state, determined by the user's behaviour and the contextual features of the game from the user's perspective (right part of the diagram of Figure 20).



**Figure 20. Overview of the system.**

The contextual features of the game selected for investigation in this scenario are the following:

**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.

From the contextual features that can be extracted in real-time during the game, 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.

**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, she might experience something closer to a "stronger punishment" sensation.

### 8.3 Relevant contextual information and non-verbal behaviours

In Phase 1 we performed experiments to investigate what user non-verbal behaviours and contextual information are effective in discriminating among the selected affective states in the "My Friend" scenario.

As far as the contextual information is concerned, experiments highlighted the following results (Castellano et al., 2009a):

- Game state is higher when the feeling is positive ( $p < 0.01$ ) and the user is engaged with the iCat ( $p < 0.05$ )
- Game evolution is higher when the feeling is positive ( $p < 0.01$ )
- When the user captures a piece, their feeling tends to be more positive than negative ( $p < 0.05$ )
- There is a significant association between the valence of feeling and the user sensations ( $p < 0.05$ )
- When the iCat displays a facial expression, the user's level of engagement towards it increases ( $p < 0.05$ )

Experiments on user non-verbal behaviours emerging in the "My Friend" scenario identified a subset of discriminative behaviours (Castellano et al., 2010):

- When the feeling is positive, the users look at the iCat more overall ( $p < 0.001$ ), look at the chessboard less ( $p < 0.001$ ) and smile more ( $p < 0.001$ )
- When the users are engaged with the iCat they look at the iCat more overall ( $p < 0.001$ ), they look at the chessboard less ( $p < 0.001$ ) and they smile more ( $p < 0.01$ )

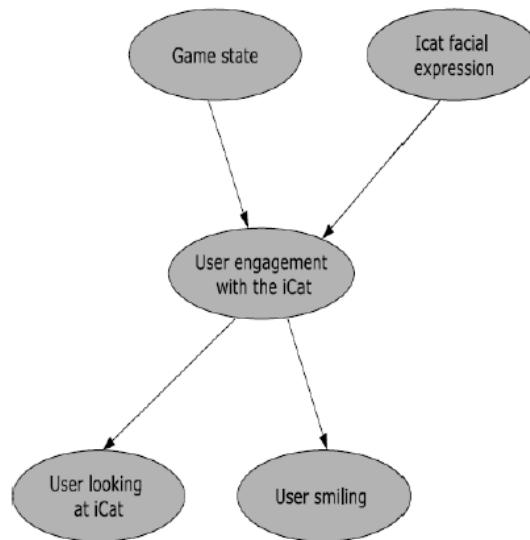
These results represent the first step of the design of an affect recognition system for a game companion in the "My Friend" scenario.

### 8.4 Initial framework to model user engagement

Based on the results presented in the previous sections, an initial framework for automatic affect recognition based on a Bayesian network was proposed to model the user's engagement with the iCat robot (Castellano *et al.*, 2009b). The 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. In our scenario user engagement is modelled using a number of task and social interaction-based features: *user looking at the iCat*, *user smiling*, *game state* and *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 21).



**Figure 21. Cause-effect relationships between user engagement with the iCat and task and social interaction-based features in our scenario.**

Training and testing the system with a person-independent approach using annotated non-verbal behaviour data and automatically extracted contextual information showed that an approach based on multimodal fusion of task and social interaction-based features outperforms those based solely on non-verbal behaviour or contextual information (see Table 2). Results also 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. For more information on the proposed framework and the experimental results see Deliverable D3.2.

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

**Table 2. Recognition rates and ROC area values for the different classifiers.**

## 8.5 The Inter-ACT corpus

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 (Zeng et al., 2010). 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 (Afzal & Robinson, 2009; Castellano et al., 2010). 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 in and the behaviour generated by the robot itself.

Following a preliminary data segmentation of our collection of videos with children interacting with an iCat robot in the "My Friend scenario" performed in Phase 1, we designed the Inter-ACT (INTERacting with Robots - Affect Context Task) corpus, which 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 (Castellano et al., submitted).

The Inter-ACT corpus consists of 156 six-second "thin slices" of the interaction between children and an iCat robot that plays chess. Each slice of the interaction is described by multimodal data: a frontal video capturing the face and the upper body of the children (captured by two cameras: (1) 15 fps, 1024X768 spatial resolution; (2) 25 fps, 720X576 spatial resolution), a lateral video (25 fps, 720X576 spatial resolution) capturing their lateral posture and full-body movements, a video capturing the iCat (standard 25 fps webcam), and a series of synchronised contextual features that describe the events of the game and the behaviour displayed by the robot. Figure 22 shows some examples of frames from the frontal and lateral view.



**Figure 22. Examples of frames from the frontal and the lateral view in the Inter-ACT corpus.**

### **8.5.1 Annotation**

Pre-segmentation was performed starting from the videos including the frontal views of the children so as to include coherent samples of behaviour: the corpus includes samples displaying full expressions, no expressions and blends of expressions. The Inter-ACT corpus is provided with affective labels that describe each "thin-slice" of the interaction.

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. Specifically, the annotators were asked to label each video in terms of the valence of feeling and the user's interest towards the iCat.

As far as the valence of the feeling is concerned, they were required to choose among the following options: positive or negative (first step of the annotation); positive, negative or neutral (second step of the annotation). In terms of interest towards the iCat, the annotators could choose among the following labels: high interest or low interest (first step of the annotation); high interest, low interest or medium interest (second step of the annotation). 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 at the different levels, with an average Fleiss' kappa value of 0.29 ( $\sigma = 0.08$ ).

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. Future studies will compare the initial annotation results with those of an annotation performed by expert coders.

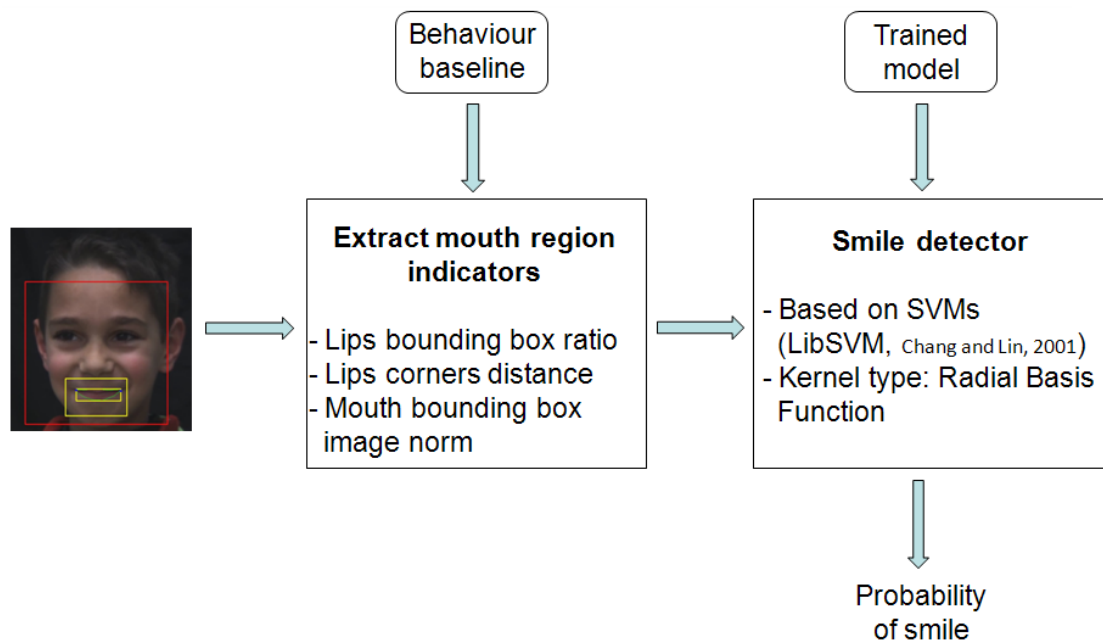
## **8.6 A prototype system for smile detection**

Results from the experiments performed in Phase 1 showed that smiles are quite important to discriminate among the selected user states in the "My Friend" scenario (see Deliverable D3.2; Castellano et al., 2010). QMUL built a prototype system for automatic smile detection as part of the affect sensitivity competency in the "My Friend" scenario.

The prototype is based on Support Vector Machines (SVMs) and it was trained using 512 samples of indicators of the mouth regions extracted using an approach based on the FacET library from naturalistic video data of children playing chess with the iCat.

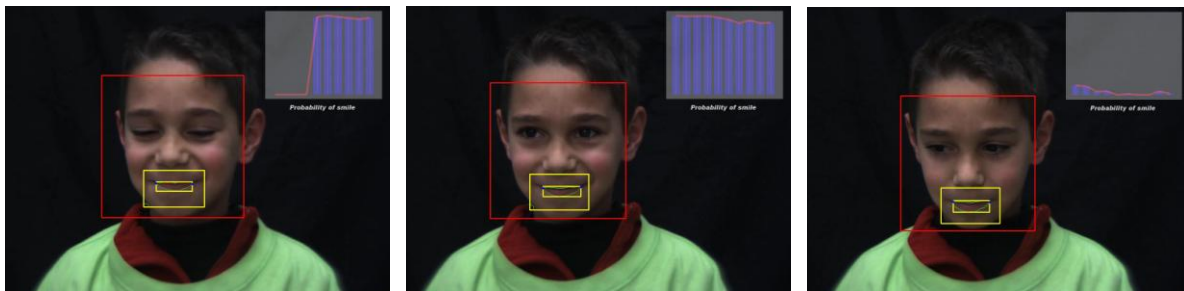
The detector extracts the following mouth region's indicators: (1) lips bounding box ratio, (2) lips corners' distance and (3) mouth bounding box image norm.

The final indicators used for training the prototype system are based on the comparison of the above indicators and a behaviour baseline computed at a local and global level: the difference between the value of the above indicators at the current frame and a local and global baseline is performed. The idea behind the behaviour baseline is to abstract from individual differences, which allows for a person-independent expression detector to be obtained. Figure 23 shows an overview of how the prototype system for automatic smile detection works.



**Figure 23. Overview of the modules composing the smile detection prototype system.**

The detector was trained with 512 samples (165 smiles samples, 347 neutral samples) using a Radial Basis Function kernel type for the SVMs. A grid search using cross-validation for parameter selection was performed. Results of a “leave-one-subject-out” cross validation with five subjects showed a recognition accuracy of 92.77 %. The system provides as output a value of the probability of smile for each frame. Figure 24 shows an example of output of the smile detector prototype system.



**Figure 24. Output of the smile detection prototype.**

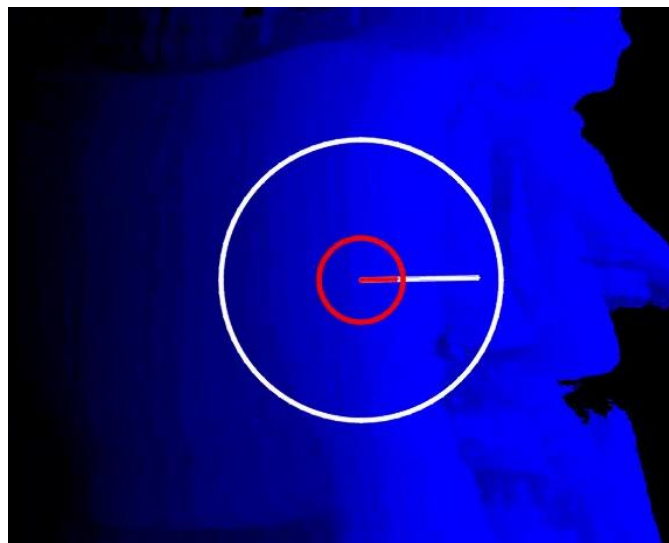
This first smile detection prototype system was trained with samples of children, but seems to work with adults as well. Robustness needs to be improved (e.g., quick movements, head tilts, lighting, and usual hard-to-solve computer vision issues), as well as real-time performances.

The smile detection prototype system is currently in use in the “My Friend” scenario and it has been successfully integrated in the iCat system.

## 9 Automatic analysis of human full-body movement: motion direction detection

Automatic analysis of global indicators, such as human full-body movement's characteristics, can provide information about the user's affect and intentions (Castellano, 2008). This information can be very useful for a companion in order to improve the interaction with the user.

QMUL developed a software module to automatically detect the overall motion direction of a user. The motion direction detection ability is based on the motion templates routines provided by OpenCV (Bradski & Kaehler, 2008). Using motion templates requires the automatic extraction of the user's silhouette. Given the user's silhouette, a motion template is built using a motion history image (mhi) and an indication of the overall motion can be derived by computing the gradient of the mhi (see Figure 25). The motion direction detection ability is currently under test in the LIREC scenarios.



**Figure 25: A measure of the overall motion direction detection of the user computed using motion templates.**

### 9.1 Application in the LIREC scenarios

A socially intelligent companion must be able to assess the appropriateness of an interaction initiation condition with the user. In a mobile interaction scenario where robots and users are free to move in the environment, such as the "Spirit of the Building" scenario, information about the position, movement and expressive behaviour of the users is of key importance for the robot to evaluate the user's willingness to interact with the robot and to plan whether interaction initiation with the user is appropriate.

In the "Spirit of the Building" scenario our robot companion is able to detect the presence of motion within a specific spatial area and compute the overall motion direction of a person walking in an orthogonal direction to a camera. This can be achieved by placing a camera on the robot or at key locations in the lab.

Planning interaction initiation with a user requires the robot companion to have the ability to evaluate when this is appropriate. In an office environment scenario, the first thing that the robot needs to evaluate is whether a user has entered or left the room.

The motion direction detection ability is useful to detect whether a user has entered the room. A camera placed in an orthogonal direction to the door captures entries and exits of users from the room and a computation of the overall motion direction of the user allows the

robot to infer whether the user has entered or left the room. If a user is present in the room, the overall motion direction of the user can be used to evaluate whether the latter is walking towards the robot or entering a specific area under which the robot can respond. Motion direction computation at different spatial locations and temporal instants can then be used to help detect whether the user is willing to start an interaction with the robot. This ability could also be used for application in the “Robot House” scenario.

The motion direction detection ability is also under investigation in the “My Friend” scenario. In this scenario it could be used to detect affective postural changes of children playing chess with the iCat (Figure 26). Videos containing a lateral view of the interaction between children and the iCat are currently under test to assess whether postural changes can be associated with different affective states experienced by the users.

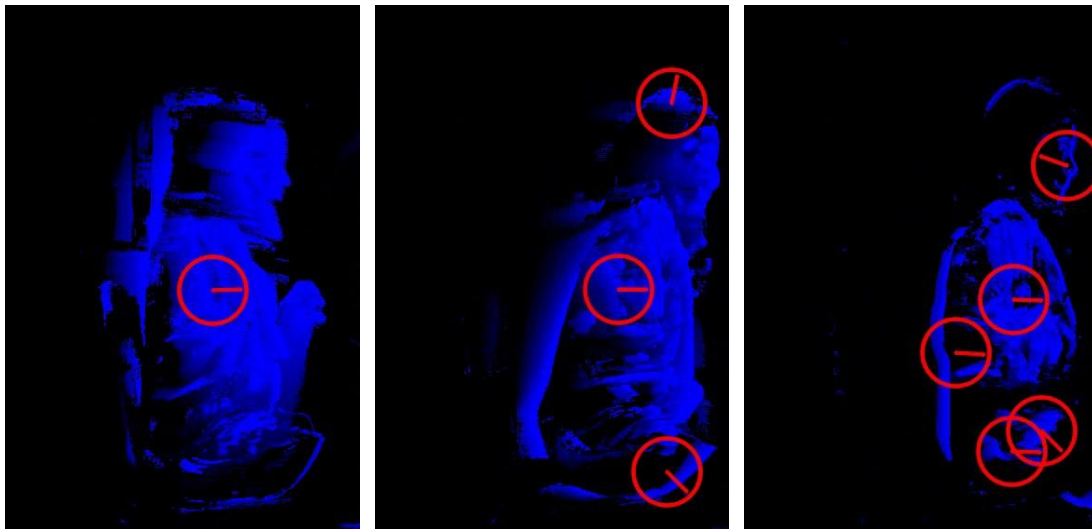


Figure 26: Example of motion direction detection data to analyse postural changes.

## 9 Conclusions

In this document we provided an overview of the vision-based social perception abilities that are currently under investigation in the LIREC scenarios. Social perception is discussed in light of psychological and ethological requirements and considerations with respect to its importance in human-companion interaction.

First prototypes of social perception abilities, including face and facial features detection and tracking, user recognition, expression recognition, affect sensitivity and motion direction detection, have been presented and their application in the LIREC scenarios described.

Note that many of these abilities are still work-in-progress and are currently under test in the LIREC scenarios with the objective to be improved. Open questions in the design of social perception abilities include the need to build systems that work in naturalistic scenarios, that is, systems that are robust in real-world conditions. Real-time performances are also required in order for the interaction between user and robots to be smooth. These issues will be addressed in further developments of these abilities. Future work will also investigate further the role of context in the social perception abilities and the modelling of the user's affective states over time.

Finally, we would like to note that our design of social perception abilities is taking into consideration possible ethical issues. So far, we have been dealing with privacy issues with the design of the Inter-ACT corpus, due to the presence of children. Protection of sensible

data is our primary concern. Further ethical issues may arise when social perception abilities will be used for closing the “affective loop” with the companions. Questions such as “*What does the companion have the right to know about the user’s state? Who owns the companion’s memory? What kind of cognitive and affective behaviour can it express? When can a companion persuade the user to engage in an interaction or be involved in a task?*” are likely to arise as a consequence of further developments of the modelling of the interaction between our companions and human users. We believe that addressing issues of this kind is of key importance when researching interaction between humans and artificial companions.

## References

- Adachi I, Kuwahata H, & Fujita K (2007). Dogs recall their owner's face upon hearing the owner's voice. *Animal Cognition* 10: 17–21.
- Afzal, S., & Robinson, P. (2009). 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 1–7, 2009.
- Bischof-Köhler, D. (1994). Selbstobjektivierung und fremdbezogene Emotionen. Identifikation des eigenen Spiegelbildes, Empathie und prosoziales Verhalten. *Zeitschrift für Psychologie*, 202, 349-377.
- Boon, S. D., & Holmes, & J. G. (1991), The dynamics of interpersonal trust: resolving uncertainty in face of risk. In R. A. Hinde, & J. Groebel (eds.), *Cooperation and Prosocial Behavior* (pp.190-211). Cambridge: University Press.
- Bradski, G. & Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly.
- Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2), pp. 119-55.
- Castellano, G. (2008). Movement Expressivity Analysis in Affective Computers: From Recognition to Expression of Emotion. Ph.D. Thesis, Department of Communication, Computer and System Sciences, University of Genoa, Italy, February 2008.
- Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A, & McOwan, P. W. (submitted). Inter-ACT: An Affective and Contextually Rich Multimodal Video Corpus for Studying Interaction with Robots.
- Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A, & McOwan, P. W. (2010). Affect Recognition for Interactive Companions: Challenges and Design in Real World Scenarios. *Journal on Multimodal User Interfaces*, 3(1), 89-98, Springer, DOI 10.1007/s12193-009-0033-5.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2009a). It's All in the Game: Towards an Affect Sensitive and Context Aware Game Companion. *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII'09)*, Amsterdam, The Netherlands. IEEE Press.

Castellano, G., & McOwan, P. W. (2009). Analysis of Affective Cues in Human-Robot Interaction: A Multi-Level Approach. *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services* (pp. 258-261), London, UK. IEEE Press.

Castellano, G., Pereira, A., Leite, I., Paiva, A., & McOwan, P. W. (2009b). Detecting User Engagement with a Robot Companion Using Task and Social Interaction-Based Features. *Proceedings of International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI'09)* (pp. 119-126), Cambridge, MA. ACM, New York, NY.

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

Davis, M. H. (1996). *Empathy: A social psychological approach*. Boulder, CO: Westview Press.

Davis, M. H., & Kraus, L. A. (1991). Dispositional empathy and social relationships. In W. H. Jones & D. Perlman (Eds.), *Advances in personal relationships* (Vol. 3). London: Jessica Kingsley.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, Calif.: Consulting Psychologists Press.

Eveno N., et al.. A new color transformation for lips segmentation, Lab. des Images et Signaux, Inst. Nat. Polytechnique de Grenoble, 2001.

Fiske, S. T. (2004). *Social Beings. Core Motives in Social Psychology*. New York: Wiley.

Franzoi, S. L., Davis, M. H., & Young, R. D. (1985). The effects of private selfconsciousness and perspective-taking on satisfaction in close relationships. *Journal of Personality and Social Psychology*, 48, 1584- 1594.

Gácsi, M, Miklósi, Á, Varga, O, Topál, J, & Csányi, V (2004). Are readers of our face readers of our minds? Dogs (*Canis familiaris*) show situation-dependent recognition of human's attention. *Animal Cognition* 7, 144-153.

Greenberg, J., & Baron, R. A. (2000). *Behavior in organizations* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Holz-Ebeling, F., & Steinmetz, M. (1995). Wie brauchbar sind die vorliegenden Fragebogen zur Messung von Empathie? Kritische Analyse unter Berücksichtigung der Iteminhalte. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 16, 11-32.

Huston, T., & Levinger, G. (1978). Interpersonal Attraction and Relationships. *Annual Reviews*, 29, 115-56.

- Lakatos, G, Soproni, K, Dóka, A, & Miklósi Á (2009). A comparative approach to dogs' (Canis familiaris) and human infants' understanding of various forms of pointing gestures. *Animal Cognition* 12: 621-631.
- Leite, I., Pereira, A., Martinho, C., & Paiva, A. (2008). Are emotional robots more fun to play with? In *Proceedings of 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*. pp. 77-82.
- Lomber, S.G., & Cornwell, P. (2005) Dogs, but not cats, can readily recognize the face of their handler [Abstract]. *Journal of Vision*, 5(8):49, 49a, <http://journalofvision.org/5/8/49/>, doi:10.1167/5.8.49.
- Long, E. C. J., & Andrews, D. W. (1990). Perspective taking as a predictor of marital adjustment. *Journal of Personality and Social Psychology*, 59, 126-131.
- Martinho, C., & Paiva, A. (2006). Using anticipation to create believable behaviour. In *American Association for Artificial Intelligence Technical Conference*, pp 1–6, Boston.
- Miklósi, Á, Kubinyi, E, Topál, J, Gácsi, M, Virányi, Zs, & Csányi, V (2003) A simple reason for a big difference: wolves do not look back at humans but dogs do. *Current Biology*, 13, 763-766.
- Namysl M., Vision system in human emotions recognition (in Polish), Master thesis, Wrocław University of Technology, 2008.
- Peters, C., Asteriadis, S., & Karpouzis, K (2010). Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces*, 3(1-2):119–130.
- Poggi, I. (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, Berlin.
- Rekleitis, I. (2004). A Particle Filter Tutorial for Mobile Robot Localization. *Technical Report TR-CIM-04-02*, Centre for Intelligent Machines, McGill University, Montreal, Quebec, Canada.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.
- Turk, M., & Pentland, A. (1991). "Eigenfaces for recognition". *Journal of Cognitive Neuroscience* 3 (1): 71–86. doi:10.1162/jocn.1991.3.1.71.
- Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing Journal*, 27(12), pp. 1743-1759.
- Virányi, Zs, Topál, J, Gácsi, M, Miklósi, Á, & Csányi, V. (2004). Dogs can recognize the focus of attention in humans. *Behavioural Processes*, 66, 161-172.
- Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, 18, 129-166.
- Zeng, Z., Pantic, M., Roisman, G. I. & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), pp. 39-58.