Title of proposal

Visual Perception System for a Social Robot (VISOR)

Coordinator

Francisco Sandoval Hernández

Dpto. Tecnología Electrónica, ETSI Telecomunicación, Universidad de Málaga

Campus de Teatinos, s/n 29071-Málaga (Spain)

Phone: +34 952 13 1352

FAX: +34 952 13 1447

sandoval@dte.uma.es

Type of project: **Topical Research Studies (TRS)**

Amount of funding requested: 70.000,00 €

List of partners: Dept. of Electronic Technology, University of Málaga (Spain)

Institute of Systems and Robotics – Coimbra (Portugal)

The emerging field of *Human-Robot Interaction* (HRI) represents an interdisciplinary effort that addresses the need to integrate social informatics, human factors, cognitive science and usability concepts into the design and development of robotic technology. The aim is the development of social robots. A social robot can be defined as an embodied agent that is part of a heterogeneous society of robots or humans (Dautenhahn and Billard, 1999). As the physical capabilities of robot improve, the reality of using them in everyday environments such as offices, factories, homes and hospitals is quickly becoming more feasible. In these environments, the socially interactive robot is able to recognize each other and engage in social interactions. It must be capable of communicate and interact with humans and other social robots, understand and even relate to humans, in a personal way. Besides, it possesses history (perceive and interpret the world in terms of their own experience), and it explicitly communicates with and learns from each other. Therefore, social learning and imitation, gesture and natural language communication, emotion and recognition of interaction partners are all fundamental factors.

To interact with humans, social robots must simultaneously perceive a great variety of natural social cues from visual and auditory channels, and to deliver social signals. This social interaction imposes that socially interactive robots must address three important issues (Breazeal, 2002):

- ✓ *Human-oriented perception*. In order to sense and interpret the same phenomena that humans observe, social robots must perceive the environment as humans do. Besides, socially interactive robots must proficiently interpret human activity and behaviour.
- ✓ *Intelligible social cues*. In order to ease the human-robot interaction, social robots must deliver social signals. These signals include facial expression, body and pointer gesturing or vocalization. They also permit that the robot provides feedback of its internal state.
- ✓ *Natural human-robot interaction*. Socially interactive robots must establish appropriate social expectations and it must follow social convention and norms.

On the other hand, it is also necessary that the robot can operate at human interaction rates. This implies that the robot must exhibit real-time performance.

The aim of this research study is to generate in-depth knowledge of the visual human-oriented perception system for a social robot. As we have commented, this is a fundamental key in the development of a socially interactive robot. If most human-oriented perception is based on passive sensing (artificial vision and auditory), the vision system is the responsible of solve the problems of identifying faces, measuring head and hands poses, capturing human motion, recognizing gestures and reading facial expressions to emulate human social perception. This information permit that the robot be able to identify who the human is, what the human is doing, how the human is doing it and even to imitate the human motion. Thus, the robot could treat the human as an individual, understand his/her surface behaviour, and potentially infer something about his/her internal states (e.g., the intent or the emotive state). On the other hand, these human-related tasks must be run in parallel with object-related ones, which permit the robot to recognize objects extracted from the environment. Our research study will investigate how the requirements for accomplishing a visual task determine the optimal architecture of a vision system. The issues that will be considered are: human and object representation, selection of low-level and high-level vision modules and control schemata. Traditionally, computer vision research has emphasized investigating each vision module as a general and isolated item. Such research efforts often generate unrealistic solutions from ill-defined assumptions. In contrast, we will consider a vision system as a whole and emphasize the research on interactions among modules as well as the research on each vision module.

An anticipate outcome of this study is that social robot visual perception cannot be viewed as a passive reconstructive activity, but as one intimately related to action. In the reconstructive paradigm, the vision task is to reconstruct physical scene parameters from image input, to segment the image into meaningful parts, and to describe the visual input in such a way that higher-level modules can act on the description to accomplish general tasks. Therefore, it requires a huge amount of computational capacity to manage the visual data acquisition and processing. Although substantial progress has been done in reconstructive vision during the last decades, it appears to be nearing its limits without reaching its goal. On the other hand, animate vision is based on two principles:

- ✓ *Active vision*. Vision must operate continuously and it must furnish results within a fixed delay. Rather than obtain a maximum of information from any one image, the camera is an active sensor giving signals that provide only limited information about the scene.
- ✓ *Task-oriented techniques*. Since visual sensing is performed with limited resources, visual strategies must to be planned so that only necessary information will be obtained. The generation of the appropriate visual strategy entails knowing what information to extract, where to get it, and how to get it. This is facilitated by the knowledge of the task.

Thus, the visual system interacts deliberately with the environment by controlling the gaze and moving the focus of attention. In animate vision, the data acquisition and the information extraction processes are closely related and they depend on the current task. Therefore, the perception process becomes an active mechanism that extracts the most relevant information from the huge amount of input data depending on the application. This selection or pre-attention mechanism allows to efficiently exploiting the available computational resources either by dedicating all of them to a specific perceptual task or by sharing them among a small set of tasks. Therefore, other central outcome of this research study is the development of a general-purpose attention mechanism based on the *feature integration theory* (Treisman and Gelade, 1980), which will be capable of handling dynamic environments, and detecting both human faces and hands and objects of interest in a fast way. The aim is to study current proposals (Backer and Mertsching, 2003) and to develop a task-based attention mechanism. Besides, it must be noted that attention capacity also helps the robot to interact socially in a natural manner. Thus, attention is, together with expression, one of the capabilities that can achieve that a human will consider that a robot is acting in a rational manner (Fong *et al*, 2003). Finally, it must be remarked that, to benefit communication and social learning, it is fundamental that both robot and human find the same sorts of perceptual features interesting.

Other very important task that must solve the visual perception system is the ability to capture the human motion. It will be especially interesting to track the human hands. In addition to facial expressions, non-verbal communication is often conveyed through gestures and body movement. Human hands are highly deformable articulated objects with many degrees of freedom and can, through different postures and motions, be used to express information. The human motion capture problem has been accurately solved using marker-based systems that usually require the human to wear especial gear and move within a constrained capture space. In this research study, we will analyse the vision-based approaches to human motion capture. We will especially focus our study in fast solutions, based on hierarchical tracking of regions (hands, head, or torso) and edges (human silhouette). One of the aims is to track a number of well-defined hand postures that represent a limited set of commands that humans can give to the robot.

This research project is aligned with the objectives of the programs Beyond Robotics Proactive Initiative (COGNIRON project), e-inclusion and Technology-enhanced Learning, inside the priority IST (Information Society Technologies of the VI Frame Program of the European Union). The process of dissemination of results will be carried out in conferences and journals, both of acknowledge prestigious in the robotic and vision research areas.

References

(Breazeal, 2002) C. Breazeal, Designing sociable robots, MIT Press, Cambridge-MA, 2002

(Dautenhahn and Billard, 1999) K. Dautenhahn and A. Billard, Bringing up robots or –the psychology of socially intelligent robots: From theory to implementation, in: Proc. of the Autonomous Agents, 1999

(Fong *et al*, 2003) T. Fong, I. Nourbakhsh and K. Dautenhahn, A survey of socially interactive robots, Robotics and Autonomous Systems, 42, pp. 143-166, 2003

(Treisman and Gelade, 1980) A.M. Treisman and G. Gelade, A feature integration theory of attention, Cognitive Psychology, 12(1), pp. 97-136, 1980