UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

Departamento de Informática y Sistemas

TESIS DOCTORAL

Contribuciones al análisis y desarrollo de robots sociables

(An Engineering Approach to Sociable Robots)

Óscar Déniz Suárez

Las Palmas de Gran Canaria, Abril 2006

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

Departamento de Informática y Sistemas

Tesis titulada *Contribuciones al análisis y desarrollo de robots sociables,* que presenta D. Óscar Déniz Suárez, dentro del programa de doctorado *Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería,* realizada bajo la dirección del Catedrático de Universidad D. Fco. Mario Hernández Tejera y la codirección del Doctor D. José Javier Lorenzo Navarro.

Las Palmas de Gran Canaria, Abril de 2006

El director

Fdo. Fco. Mario Hernández Tejera

El codirector

Fdo. José Javier Lorenzo Navarro

El doctorando

Fdo. Óscar Déniz Suárez

To my family and Gloria

Acknowledgments

The work described in this document would not have been possible without the disinterested help of many people. First of all, I want to acknowledge the continuous support of my tutor Mario Hernández. Besides being very knowledgeable about the topics covered in my work, Mario is one of the most cheerful guys I've ever met. My second tutor Javier Lorenzo also supported me on many occasions, impressing me with his incredibly fast thought.

Many people contributed sporadically: Jorge Cabrera, Daniel Hernández, Antonio Carlos Domínguez, Cayetano Guerra, Josep Isern, Antonio Falcón, Juan Méndez, Irael Pérez and David Hernández, all from my research group. Modesto Castrillón was also a good companion, from back when I started working on face recognition. Now I know he is the person I should resort to if I need to know something about facial analysis or travelling around the world. Modesto kindly provided me with the ENCARA face detection module. David let me use its ZagaZ module for action selection. Also, the French mechanical engineers of IFMA who visited our laboratory, Yann, Patrick, Celine, Vincent and Olivier, were decisive in the mechanical design of CASIMIRO.

Other people did not contribute so specifically, but they did in other subtle ways that I can't articulate: Yeray, Sanyu, José Luis, Claudio, Carlos Moisés, Oscar, Fayna, Atamán, Jaime, Carmelo, Cristóbal, Patricia, Antonio and Marilola. Much to my surprise, instant messaging turns out to be a good concept when you spend so many hours working in front of a computer screen. I also want to thank my pupils and fellow teachers at the School and Faculty of Computer Science, especially Miguel Angel Pérez, Alexis Quesada, Carmelo Rubén García, José Antonio Muñoz, Francisca Quintana and Santiago Candela, who were very understanding and helpful with respect to my teaching duties.

I want to thank my parents and my sisters María Jesús (Susi) and Lorena. Hopefully, now that I have finished this work Susi and Lorena will have more access to our home Internet connection.

I am grateful for the funding granted by the entities that supported our research group:

Gobierno de Canarias, Ministerio de Ciencia y Tecnología and Universidad de Las Palmas de Gran Canaria. I am personally indebted to Universidad de Las Palmas de Gran Canaria, and specially to the research and academic staff vice-chancellors, who have supported me since I got a research grant back in 2000.

Finally, I want to thank the person who's holding this document right now for showing interest in my work. I hope to be back soon. It is my hope that CASIMIRO's abilities will be expanded in the future. I've always thought that the most important thing is not the destination, but the journey itself. *In itinere sumus*.

Contents

List of Figures

List of Tables

Abstract

A relatively new area for robotics research is the design of robots that can engage with humans in socially interactive situations. These robots have expressive power (i.e. they all have an expressive face, voice, etc.) as well as abilities to locate, pay attention to, and address people. In humans, these abilities fall within the ambit of what has been called "social intelligence". For this class of robots, the dominant design approach has been that of following models taken from human sciences like developmental psychology, ethology and even neurophysiology.

We show that the reproduction of social intelligence, as opposed to other types of human abilities, may lead to fragile performance, in the sense of having very different performances between training/testing and future (unseen) conditions. This limitation stems from the fact that the abilities of the social spectrum, which appear earlier in life, are mainly unconscious to us. This is in contrast with other human tasks that we carry out using conscious effort, and for which we can easily conceive algorithms. Thus, a coherent explanation is also given for the truism that says that anything that is easy for us is hard for robots and vice versa.

For some types of robots like manipulators one can extract a set of equations (or algorithms, representations,...) that are known to be valid for solving the task. Once that these equations are stored in the control computer the manipulator will always move to desired points. Sociable robots, however, will require a much more inductive development effort. That is, the designer tests implementations in a set of cases and hopes that the performance will be equally good for unseen (future) cases. Inductive processes crucially depend on *a priori* knowledge: if there is little available one can have good performance in test cases but poor performance in unseen cases (overfitting).

In Machine Learning, complexity penalization is often used as a principled means to avoid overfitting. Thus, we propose to develop sociable robots starting from simple algorithms and representations. Implementations should evolve mainly through extensive testing in the robot niche (the particular environment and restrictions imposed on the robot tasks,

physical body, etc.). Such approach places more emphasis in the engineering decisions taken throughout the robot development process, which depend very much on the niche.

This work describes the ideas and techniques involved in the design and development of CASIMIRO, a robot with a set of basic interaction abilities. The robot has been built following the mentioned approach. In particular, the main difficulties lay in parsimoniously exploiting the characteristics of the robot niche in order to obtain better performances.

Publications

This thesis is based on research that was previously reported in the following publications:

- *"Face Recognition Using Independent Component Analysis and Support Vector Machines"*, O. Déniz, M. Castrillón, M. Hernández. IX Spanish Symposium on Pattern Recognition and Image Analysis, Castellón (Spain), May 2001.

- *"Face Recognition Using Independent Component Analysis and Support Vector Machines"*, O. Déniz, M. Castrillón, M. Hernández. Procs. of the Third International Conference on Audio- and Video-Based Person Authentication. Lecture Notes in Computer Science 2091, pp. 59-64. Halmstad, Sweden, June 2001.

- *"Estudio Experimental sobre la Combinación Temporal de Resultados en el Reconocimiento de Caras con Secuencias de Vídeo"*, O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández. IX Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA-2001, Vol. I pp. 273-282, Gijón, November, 2001.

- *"El Método IRDB: Aprendizaje Incremental para el Reconocimiento de Caras"*, O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández. IX Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA-2001, Vol. I pp. 273-282, Gijón, November, 2001.

- *"Aprendizaje Incremental para la Identificación Facial a Partir de Secuencias de Vídeo"*, O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández. Revista Española de Visión por Computador. vol. 6, March, 2002.

- *"Modelado de expresiones para una cara robótica"*, O. Déniz, A. Falcón. Revista Buran, vol. 10, nº 18, April 2002.

- *"An Incremental Learning Algorithm for Face Recognition"*, O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández. Lecture Notes in Computer Science, vol. 2359, pp. 1-9,

Springer-Verlag, PostECCV'02 Workshop on Biometric Authentication, Copenhagen, Denmark, June 2002.

- *"CASIMIRO: A Robot Head for Human-Computer Interaction"*, O. Déniz, M. Castrillón, J. Lorenzo, C. Guerra, D. Hernández, M. Hernández. 11th IEEE International Workshop on Robot and Human Interactive Communication, pp. 319-324, Berlin, Germany, 25-27 September, 2002.

- "A Computational Mechanism for Habituation in Perceptual User Interfaces", O. Déniz, J. Lorenzo and M. Hernández. Int. Conference on Computational Intelligence for Modelling, Vienna, Austria, February 12-14, 2003.

- *"Building a Sound Localization System for a Robot Head"*, O. Déniz, J. Cabrera and M. Hernández. Revista Iberoamericana de Inteligencia Artificial, 18 - Winter 2003.

- *"Multimodal Attention System for an Interactive Robot"*, O. Déniz, M. Castrillón, J. Lorenzo, M. Hernández and J. Méndez. Lectures Notes in Computer Science, vol. 2652. 1st Iberian Conference on Pattern Recognition and Image Analysis. 4-6 June 2003, Pto. Andratx, (Mallorca, Spain).

- *"Face Recognition using Independent Component Analysis and Support Vector Machines"*, O. Déniz, M. Castrillón and M. Hernández. Pattern Recognition Letters, vol 24, issue 13, pp. 2153-2157, September 2003.

- *"BDIE: A BDI like architecture with emotional capabilities"*, D. J. Hernández, O. Déniz, J. Lorenzo and M. Hernández. AAAI Spring Symposium, 22-24 March 2004.

- *"A simple habituation mechanism for perceptual user interfaces"*, O. Déniz, J. Lorenzo and M. Hernández. Revista Iberoamericana de Inteligencia Artificial, vol. VIII, 33, pp. 7-16, 2004.

- *"Useful Computer Vision Techniques for Human-Robot Interaction"*, O. Déniz, A. Falcón, J. Méndez, M. Castrillón, ICIAR 2004, International Conference on Image Analysis and Recognition, September 2004, Porto, Portugal.

- *"Expressive Robotic Face for Interaction"*, O. Déniz, L. Antón-Canalís, M. Castrillón, M. Hernández. VII Workshop de Agentes Físicos (WAF'06), April 2006, Las Palmas de Gran Canaria, Spain.

Chapter 1

Introduction

"Man is by nature a social animal"

Aristotle, *Politics*.

In recent years there has been a surge of interest in a topic called social robotics. As used here, social robotics does not refer to groups of robots that cooperate and interact with each other. For a group of robots, communication is relatively simple from a technological point of view, they can use whatever complex binary protocol to "socialize" with their partners. Back in the 40's [Holland, 1997] robotic tortoises already interacted in a "social" manner using headlamps attached to the robots and by means of phototaxis. This field of Artificial Intelligence evolved and produced concepts like "swarm" robots, "ant-like" robots, self-organization, etc.

For us, the adjective social refers to humans. In principle, the implications of this are much wider than in the case of groups of robots. Although the distinction may be much fuzzier, social robotics aims at building robots that do not seem to have a specific task (like playing chess with a human) but to simply interact with people. Socializing with humans is definitely much harder, not least because robots and humans do not share a common language nor perceive the world (and each other) in the same way.

Many researchers working on this topic use other names like human-robot interaction, perceptual interfaces or multimodal interfaces. However, as pointed out in [Fong *et al.*, 2003] we have to distinguish between conventional human-robot interaction (such as that used in teleoperation scenarios or in friendly user interfaces) and socially interactive robots. In the latter, the common underlying assumption is that humans prefer to interact with robots in the same way that they interact with other people.

Traditional robots can carry out tasks that are well beyond human capabilities, with greater precision, no risk, and a large number of times. Then why so much interest in social robots?

Admittedly robots are built with the aim of imitating or reproducing human intelligence. "Before" social robotics, most robots excelled at certain tasks, although they were incapable of doing other "simple" things, not even in a partial way. Almost in an unconscious fashion, researchers realized that those robots were extremely practical, but were not considered more intelligent (at least by the general public). In a sense, the more practical and precise the robot, the less human it was considered.

The fact is that humans have a wide range of abilities. Comparatively, robots tend to have far fewer abilities, although with high proficiency levels, in some cases even higher than in humans (see Figure 1.1). Such performance unbalance may be key to building social robots. It suggests that two necessary conditions for achieving a more "human" robot may be:

1. to replicate a large number of human abilities, and
2. that these abilities have similar development levels.

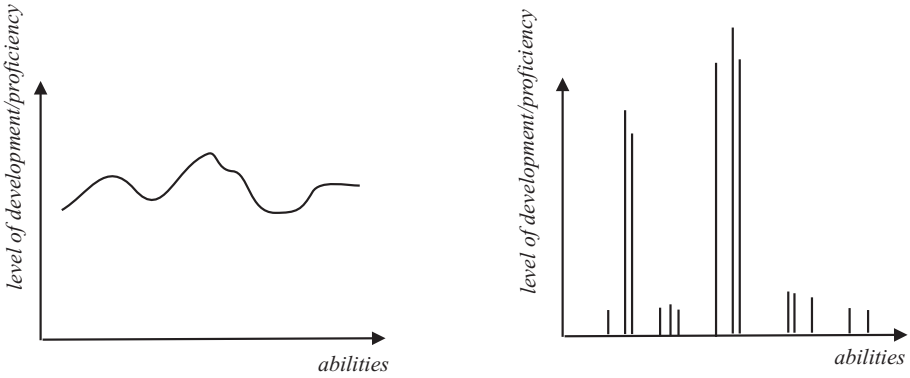


Figure 1.1: Abilities vs. level of development. On the left: human performance. On the right: typical robot performance.

Human intelligence does not seem to be restricted to certain abilities. According to Gardner's theory [Gardner, 1983], now widely accepted, human intelligence is not a unitary capacity that can be measured by IQ tests. He proposes eight classes of intelligence (see Table 1.1): linguistic, musical, logical-mathematical, spatial, bodily-kinesthetic, interpersonal, intrapersonal and naturalist. Intelligence has two meanings. First, it is a species-specific

characteristic (humans can exhibit those eight intelligences). Second, it is also an individual-specific characteristic (each individual shows his particular blend of intelligences). Gardner claims that the eight intelligences rarely operate independently. Rather, the intelligences are used concurrently and typically complement each other as individuals develop skills or solve problems. For example, a dancer can excel in his art only if he has 1) strong musical intelligence to understand the rhythm and variations of the music, 2) interpersonal intelligence to understand how he can inspire or emotionally move his audience through his movements, as well as 3) bodily-kinesthetic intelligence to provide him with the agility and coordination to complete the movements successfully. The theory states that all eight intelligences are needed to productively function in society.

Intelligence	*Operations*
Linguistic	syntax, phonology, semantics, pragmatics
Musical	pitch, rhythm, timbre
Logical-mathematical	number, categorization, relations
Spatial	accurate mental visualization, mental transformation of images
Bodily-kinesthetic	control of one's own body, control in handling objects
Interpersonal	awareness of others' feelings, emotions, goals, motivations
Intrapersonal	awareness of one's own feelings, emotions, goals, motivations
Naturalist	recognition and classification of objects in the environment

Table 1.1: The eight human intelligences proposed by Gardner.

Gardner's proposal of multiple intelligences was prompted by several signs, like isolation by brain damage, in which certain forms of intelligence are impaired while others remain relatively intact. Damage to the prefrontal lobes of the cerebral cortex can impair personal and social intelligence, while other abilities remain normal [Damasio, 1994]. Gardner mentions exceptional cases like autistic savants, and he notes that both the Down syndrome and the Alzheimer's disease are cognitive impairments that do not severely hinder a person's ability to get along with other people. By contrast, Pick's disease seriously diminishes a person's interaction abilities, though some intellectual capacities are not affected [Kihlstrom and Cantor, 2000].

Not only theory favours the fact that there are different types of intelligence. In a series of empirical studies, [Sternberg *et al.*, 1981] found that people generally identify intelligence with three core abilities: problem-solving, verbal and social competence. Hudson's theory of convergent and divergent thinking is also related with the idea emphasized above [Hudson, 1967]. Convergent thoughts are those that lead to the solution of a concrete problem, while divergent thought are those that are prompted by stimuli. Convergent thoughts are related to mathematics and sciences. Divergent thoughts are related to creativity, arts and

humanities. Every individual has a little amount of each component, and some individuals show preference for (or are more inclined to) a particular component. Problems appear when one of the components is not well developed.

All these findings suggest that social intelligence aspects must be addressed if we are to build robots that imitate human intelligence. What is more, social intelligence could be even more important than other capacities. There is evidence that in primates social intelligence is one important prerrequisite for the evolution of non-social, domain-independent intelligence [Dautenhahn, 1995]. As an example, the highly elaborated ability of symbolization is suggested to be a social act of agreeing [Bullock, 1983]. Some authors contend that social intelligence is also necessary for the development of generic intelligence in humans, see [Lindblom and Ziemke, 2003]. This 'Social Situatedness Hypothesis' (also known as Machiavellian Intelligence Hypothesis) emphasizes the importance of designing and building robots that can interact with other agents. In this respect, the work of Brooks already showed the importance and benefits of this approach [Brooks, 1991].

This document describes the experiences, ideas and techniques involved in building CASIMIRO [1], a robot with basic social abilities. CASIMIRO is a robotic head that includes facial features and neck movements. The robot was not designed for performing a certain precise task. If any, its task would be to interact with humans (note the vagueness of that). At some point in the development process the option of restricting the robot's behaviour to a game or entertaining task was considered, although that was always discarded. For some members of our group (including me) this produced at times a feeling that the robot would end up doing anything. When I have that sort of *horror vacui* feeling, I think in babies. They do not seem to be doing anything, although they already have some physical and mental capabilities that are unattainable for a machine.

1.1 Previous Work

In order to provide the reader with a more practical viewpoint on social robotics, in this section a brief description of the most influential social robots built is given. Not all of such robots appear here. Being an emergent field, their number seem to increase on a monthly basis. This section is intentionally short, as other robots will be referenced and described in the rest of the document.

The following three major application areas can now be distinguished for interaction

[1] The name is an Spanish acronym of "expressive face and basic visual processing for an interactive robot"

robots:

- "Robot as a persuasive machine" [Fong *et al.*, 2003]: the robot is used to change the behaviour, attitudes or feelings of people. Also, any social robot that is used as a (bidirectional) user interface may also fall into this category, as would robots that are social for entertainment or demonstration purposes.
- Robots that are designed and used to test theories of human social development or communication. This and the first type overlap and complement each other. Many social robots are inspired by theories of social development, and that effort is obviously in the direction of making the robot more persuasive, as observed by humans.
- Robots that must interact/collaborate with people in order to accomplish the assigned tasks. Here the main task is not that of interacting or socializing with people. Collaboration/interaction is just used to get the main task done.

An example of the first type of robots can be found in the AURORA project, which uses robots in autism therapy [Dautenhahn and Werry, 2001]. Potential groups of application are children and the elderly. Another such robot is the seal robot Paro, which has been tested at nursing homes and with autistic and handicapped children, and has been recognized by the Guinness Book of Records as the "world's most therapeutic robot" [Shibata and Tanie, 2001].

Social robots that fall into the first category explained above find very fitting working scenarios in places like museums and exhibition centres. These guide robots may also fall into the third category (this depends on the importance of the robot itself as an attraction as compared to the importance of the museum tour task). Minerva [Schulte *et al.*, 1999] is an interactive tour-guide mobile robot that displays four basic expressions, namely happy, neutral, sad and angry. A state machine is used to transition between those emotional states, in that specific order. The robot is happy when it can move freely and sad when people block its way. Perhaps one of the most attractive aspects of Minerva is its capability to learn how to attract people. It performs series of actions (facial expression, gaze direction and sound) and then evaluates them based on a reinforcement signal. The signal is positive when there is an increase in closeness and density of people around the robot. This shows the importance of adapting to the user's behaviour and, particularly, to maintain a memory that allows the robot to change its own behaviour.

Kismet [Breazeal, 2002] has undoubtedly been the most influential social robot appeared. The most important robot that CASIMIRO relates to is Kismet, and it was taken from the beginning as a model and inspiration (CASIMIRO's external appearance is in fact

very similar to that of Kismet, albeit this was not achieved intentionally). It is an animal-like robotic head with facial expressions. Developed in the context of the Social Machines Project at MIT, it can engage people in natural and expressive face-to-face interaction. Kismet was conceived as a baby robot, its abilities were designed to produce caregiver-infant exchanges that would eventually make it more dexterous. An overview of Kismet is available at [MIT AI lab, Humanoid Robotics Group, 2003].

Kismet is equipped with visual, auditory and proprioceptive inputs. The vision system includes four colour CCD cameras. Two of them have a larger field of view, and the other two are foveal cameras that allow higher-resolution processing. Kismet's gaze is controlled by three degrees of freedom in the eyes and another three in the neck. An attentional system based on basic visual features like skin tone, colour and motion allows it to direct its attention to relevant stimuli and gaze toward them. The auditory system processes signals gathered by a wireless microphone worn by the caregiver. The auditory system can recognize the affective intent of the speaker, i.e. it can recognize praise, prohibition, attention, and comfort.

Kismet's face has 15 DOF that allows it to display facial expressions like fear, accepting, tired, unhappy, disgust, surprise, anger, stern and content. The vocalization system is based on a DECtalk v4.5 speech synthesizer in which parameters are adjusted to convey personality (Kismet babbles like a young child) and emotional state.

Another MIT robot, Cog [Adams *et al.*, 2000, Brooks *et al.*, 1999, Scassellati, 2000], has a trunk, head and arms, for a total of 22 degrees of freedom. The head has 4 DOF in the neck and 3 DOF in the eyes. Its capabilities include human-like eye movements, head and neck orientation (in the direction of a target), face and eye detection, imitation of head nods, basic visual feature detectors (colour, motion and skin tone), an attentional system that combines them, sound localization, reflex arm withdrawal, shared attention (see below), reaching to visual targets and oscillatory arm movements. Recently, the work of Arsenio [Arsenio, 2004] has endowed the robot with significant high-level capabilities such as object, face and scene recognition, acoustic segmentation and recognition, activity recognition, etc.

Both Kismet and Cog were developed with the aim of exploiting scaffolding. Scaffolding is a teaching strategy introduced by Constructivist psychologist Vygotsky's sociocultural theory. Parents modulate and facilitate learning by creating an appropriate environment for the infant. Aspects like the use of simple phrases and achievable goals can facilitate learner's development. In the case of a robot, an instructor would guide interactions so as to foster novel abilities. The instructor may for example mark the critical aspects of the task or concept to learn, reduce the degrees of freedom, show the robot the effects of its actions with

respect to the task to learn, etc.

Infanoid [Kozima, 2002, Kozima and Yano, 2001] is a robot that can create and maintain shared attention with humans (it is an upper-torso humanoid, with a total of 24 degrees of freedom). Infanoid was inspired by the lack of attention sharing observed in autism. Attention sharing, the activity of paying attention to someone else's attentional target (Figure 1.2), plays an indispensable role in mindreading (see Chapter 2) and learning, and it has been shown to be important for human-robot communication [Yamato *et al.*, 2004]. The robot first detects a human face and saccades to it. Then, the robot detects the eyes and extract gaze direction. It then starts looking for a target with a clear boundary in that direction. Recently, however, it has been shown that shared attention actually requires more than gaze following or simultaneous looking [Kaplan and Hafner, 2004].

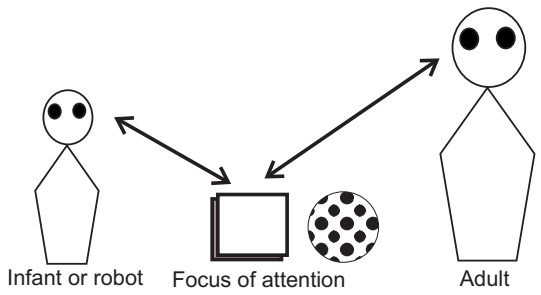


Figure 1.2: Shared attention.

Babybot [Metta *et al.*, 2000*a*, Metta *et al.*, 2000*b*] is a baby humanoid built at LIRA-Lab. The latest version at the time of writing has 18 degrees of freedom distributed along the head, arm, torso, and hand. The head was custom designed at the lab. The Babybot's sensory system is composed of a pair of cameras with space-variant resolution, two microphones each mounted inside an external ear, a set of three gyroscopes mimicking the human vestibular system, positional encoders at each joint, a torque/force sensor at the wrist and tactile sensors at the fingertips and the palm. The scientific goal behind Babybot encompasses the developmental approach to building social robots, also called *Epigenetics* (see [Lungarella *et al.*, 2004] for a survey). The developmental approach emphasizes the use of minimum representations and algorithms so that the robot can learn and develop by itself its own abilities (and adapt to the environment), especially with the help of human teachers.

Studies on human emotions have also been a major source of inspiration. The ability to show emotions is crucial for social interaction. Feelix (FEEL, Interact, eXpress) is a LEGO robot that displays different facial emotional expressions in response to tactile stimulation [Cañamero and Fredslund, 2001, Cañamero and Fredslund, 2000]. Using two eyebrows and two lips it can display six basic emotions: neutral, anger, sadness, fear, happiness

and surprise. Feelix implements two models of emotional interaction and expression inspired by psychological theories about emotions in humans. Tactile sensations translate directly into emotions. This low perceptual bandwidth does not require attentional mechanisms, which would be necessary with more complex sensors such as vision.

Advanced perception techniques are obviously necessary for successful interaction. SIG [Okuno *et al.*, 2002] is a humanoid robot designed as a test bed of integration of perceptual information to control a large number of degrees of freedom. The visual perceptual subsystem is in charge of a stereo vision system, each camera having 3 DOF. The most remarkable aspect of SIG is its sound localization ability. It uses a total of four microphones. The body is covered with an isolating material. Two of the microphones are placed inside and the other two outside. This allows the robot to cancel the noise of its own motors (see Section 5.2).

The ability to talk to multiple users in the same session is addressed in ROBITA [Matsusaka *et al.*, 1999]. This poses many problems, such as recognition of who is speaking and to whom he is speaking. To solve them, the robot implements face recognition, face direction recognition, sound localization, speech recognition and gestural output. It has 2 CCD cameras on his head, 2 DOF for each of them and 2 DOF in the neck. The robot uses a total of 9 computers.

The availability of hardware elements has led to a profusion of social robots. A big laboratory and group of researchers is no longer necessary to complete a complex robot. There is a clear -and positive- tendency to fall into the I-want-to-build-one-too, do-it-yourself fever. Interestingly, I-want-to-build-one-too systems, despite being simple in hardware and techniques, are comparable to more complex systems. That is, excessive or advanced hardware does not seem to have a crucial effect on the overall "quality" of the robot. Two outstanding robots built with low-cost components are ARYAN and SEGURITRON, see Chapter 4.

1.2 Analysis

The study of the available literature about social robots allows to extract a number fundamental ideas. First, there seems to be a clear tendency to use psychology, ethology and infant social development studies as the main inspiration source. In fact, the development of almost all of the robots built for social interaction revolve around one or more human models. This is the case of Kismet, for example, the design of which is inspired by infant-caregiver relationships. Other robots have used concepts like scaffolding, autism, imitation, shared

attention, mindreading, empathy or emotions. In some cases, only human models have been implemented, to the detriment of other aspects like the engineering point of view (called "functional design" in [Fong *et al.*, 2003]) or the question of the validity of known human models.

Not only high-level psychological and cognitive models are being used. Neurophysiological knowledge is also being frequently used for building social robots or modules for them. The work of Arsenio, mentioned above, is particularly significant in this respect. His PhD dissertation is in fact organized according to functional structures of the human brain, the so-called *brainmap*. Something similar appears in the recent work of Beltrán-González [Beltrán-González, 2005], whose work on prediction in perceptual processes for artificial systems revolves around the function of the human cerebellum.

To what extent are these models, concepts and theories valid for building social robots? Are they just useful metaphors? or will, in effect, the reproduction of the models achieve similar performances as in humans? These questions are now important, as the interdisciplinary character of the problem is becoming evident. In particular, two aspects seem specially relevant: 1) the validity of human models, concepts and theories for a robot and 2) the level of detail at which those models should be reproduced.

Another common feature of this type of robots is the extensive integration of different techniques and hardware. The robot designers normally choose state-of-the-art techniques for solving certain tasks and use them in the robot. Is that always the appropriate approach? Can we expect those advanced techniques, tested in other scenarios, to work well in our particular robot?

On the other hand, a careful analysis of the related work may lead to the question of whether these and other robots that try to accomplish social tasks may have a robust behaviour. For industrial robots, robustness can be easily measured. That is, we can be as sure as we want that the robot will keep on working as in the test conditions. This warranty stems from the fact that the processes and algorithms that the robots implement for solving the tasks are known to be valid. As an example, consider a robot manipulator, for which precise kinematic equations have been previously obtained. Once these equations are obtained and stored in the control computer only a few tests will be necessary for us to have the guarantee that the manipulator will move to indicated points.

The case seems to be different for social robots. In face recognition for example (the social ability par excellence) the number of papers describing working implementations is significantly low compared with the total. Face recognition techniques are extremely sensitive to illumination [Adini *et al.*, 1997, Georghiades *et al.*, 1998, Gross *et al.*, 2002], hair

[Liao *et al.*, 1997], eyeglasses, expression, pose [Beymer, 1993, Gross *et al.*, 2004], image resolution [Wang *et al.*, 2004], aging, etc. Pose experiments, for example, show that performance is stable when the angle between a frontal image and a probe is less than 25 degrees and that performance dramatically falls off when the angle is greater than 40 degrees [Blackburn *et al.*, 2001]. As for the other factors, acceptable performances can be achieved only under certain circumstances, see Table 1.2.

Category	*False reject rate*
Same day, same illumination	0.4 %
Same day, different illumination	9 %
Different days	11 %
Different days over 1.5 years apart	43 %

Table 1.2: Face verification error (for a fixed false alarm rate of 2%) when test conditions differ from training conditions (taken from [Martin *et al.*, 2000]).

Also, speech recognition performance decreases catastrophically during natural spontaneous interaction. Factors like speaking style, hyperarticulation (speaking in a more careful and clarified manner) and emotional state of the speaker significantly degrade word recognition rates [Oviatt, 2000]. Above all, environmental noise is considered to be the worst obstacle [Wenger, 2003]. The mouth-microphone distance is in this respect crucial. The typical achievable recognition rate for large-vocabulary speaker-independent speech recognition is about 80%-90% for clear environment, but can be as low as 50% for scenarios like cellular phone with background noise.

Sound localization is very hard to achieve in noisy environments. Sound signals tend to interfere with each other and are transformed by collisions with furniture and people [Good and Gilkey, 1996, Shinn-Cunningham, 2003]. On the other hand, the physical construction of the robot itself can have a negative effect on sound localization (in humans, it has been demonstrated that hearing protection affects our sound localization abilities.

In summary, there is the impression (especially among researchers) that performance would degrade up to unacceptable levels if conditions were different from those used to test the implementations. In test scenarios, performance is acceptable. However, it would seem that there is little guarantee that it remains at the same levels for future, unseen conditions and samples. How can we explain this negative impression? Note that it does not appear for other types of robots, say industrial manipulators, where the robot performance is somehow "under control". This leads us to the important question: is building a social robot in any sense different than building other kinds of robots?

Chapter 2

The Case of Social Robotics

"We see what we know"

A. Snyder, T. Bossomaier, J. Mitchell, in *Concept formation: 'Object' attributes dynamically inhibited from conscious awareness*, 2004.

Social robotics is having an enormous influence on the part of disciplines like cognitive sciences, neurobiology and ethology. Scassellati contends that developmental psychology and robotics can even complement each other (see [Scassellati, 2000]). On the one hand, robotics will rely on human development for inspiration and practical theories. On the other hand, human development can profit from the evaluation and experimentation opportunities that robotics offers. In this sense, the most well known relationship is perhaps that between social robots and autism.

Autism is a developmental disorder characterized by impaired social and communicative development, and restricted interests and activities [Frith, 1989]. Many theories try to identify the etiology of autism. From a cognitive point of view, the Theory of Mind (TOM) hypothesis has dominated research since the mid-1980's. This theory was developed from findings of Baron-Cohen, Leslie and others that indicated that autism was related to a lack or impairment in the models that we have of the others as people with their own sets of beliefs, desires, etc (i.e. *mindreading*). 'False belief' tests are normally used to identify this deficit. The Sally-Anne test is representative [Baron-Cohen, 1995]. In a room there is a box, a basket and a puppet, Anne (see Figure 2.1). Sally, another puppet, enters the room and puts its ball into the basket. Sally leaves the room, and then Anne puts Sally's ball into the box. Then Sally comes back. The child under test, who has been watching the process, is then asked where is Sally going to look for its ball. Normal children can correctly answer 'in the

basket', though autistic children's answer is 'in the box'. However, it is not clear whether the lack of mindreading abilities is a cause or an effect of another more basic impairment. Some autistic symptoms (like stereotypic behaviour, restriction of interests, oversensitivity to physical stimuli and notably the phenomenon of autistic savants (see below) are not well accounted for by the TOM hypothesis [Hendriks-Jansen, 1997].

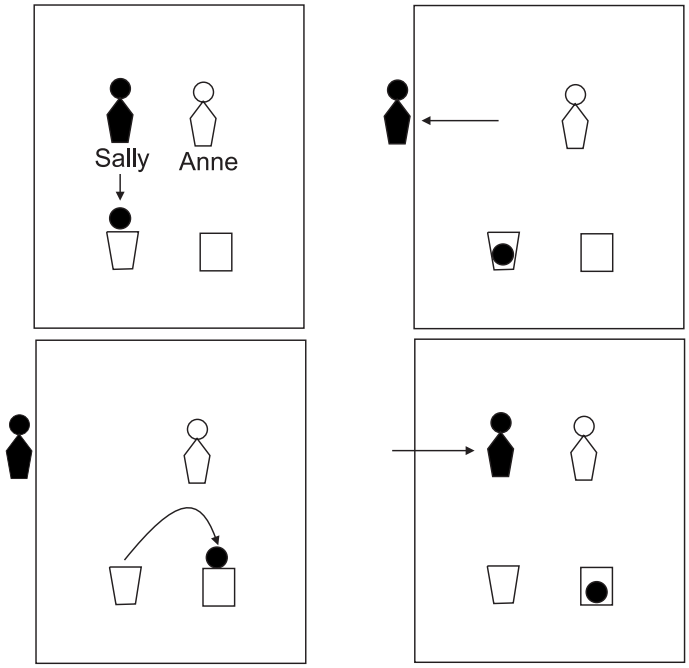


Figure 2.1: the "Sally-Anne Test" of understanding false belief.

It is important to note that the relationship between autism and social robotics has been twofold. On the one hand, social robots have been used as a therapeutic tool in autism, as in the AURORA project. On the other hand, autism has been an inspiration for building social robots by attending to the lack of abilities that autistic people have -i.e. by considering autism as an analogy of non-social robots (this is the approach taken by Scassellati [Scassellati, 2001]).

2.1 The Savant Syndrome

Interestingly, autism is not the most precise analogy for non-social robots. If we were to identify a condition that could be associated to the robots that we build, the most precise one should be that of autistic savants. This idea has been already suggested informally by some authors [Blumberg, 1997]. The autistic savant is an individual with autism who have

extraordinary skills not exhibited by most people [Treffert, 1989]. The most common forms involve mathematical calculations, memory feats, artistic and musical abilities. Autistic savants behave very much like computers (this should not be taken as a negative trait). From [Treffert, 1989]:

> "The savant develops into a well-trained 'robot' with little ability to change with changing conditions; in the face of change or challenge, the savant holds to obsessive, stereotyped, concrete responses."

One of the best known cases of autistic savants is that of the twin calculators, John and Michael, who were born with the same physical and neurological deficiencies. Despite being autistic, they could perform certain arithmetic calculations at incredible speeds, see [Sacks, 1985].

Autistic savants are very rare, the estimated prevalence of savant abilities in autism is 10%. Their skills are very concrete, restricted to rather limited domains, as evidenced in the most widely known account of an autistic savant: the *Rain Man* movie (there are other two, *Being there* with Peter Sellers, and possibly *Forrest Gump* with Tom Hanks).

There are many theories that seek to explain the strange phenomenon of autistic savants. Many researchers find the case extremely interesting because of the fact that abnormalities can help discover what is "normal" in human brains (*"Everything we know about health we have learned from the study of disease"*, [Treffert, 1989]). According to a recent theory proposed by Snyder and Mitchell [Snyder and Mitchell, 1999, Snyder and Thomas, 1997*a*], autistic savants have not developed in their brains the part that gives meaning and interpretation to the things that we perceive. Autistic savants perceive "everything", they have an eye for detail, they perceive images and sounds with extreme precision. However, they have difficulty in attributing high-level meaning to such detail. They struggle to isolate the important part of the data, the part that conforms a representation of high-level concepts. That inability can render them socially incompetent.

Healthy people can isolate useful data, and in fact we do that in order to survive, for we must make useful decisions quickly. We are concept-driven. Only with very hard training can healthy people manage detailed information and use it for example for drawing (our concept-driven thought can be seen as an advantage but also as a serious obstacle [Snyder and Thomas, 1997*b*]). According to Snyder and Mitchell, healthy people can also perceive with extreme detail, though as that ability is unconscious, they become more conscious of the final product, the concept ("we see what we know"), see Figure 2.2. On the

contrary, autistic savants would be able to access raw information [1].

Figure 2.2: Children can not identify the couple sharing sexual intimacy (right) because they lack prior memory associated with such scenario. They can see nine dolphins (left).

Representative of this theory is the case of the autistic savant child Nadia [Selfe, 1977]. At about three and a half years, Nadia could draw natural images with great detail, without training. She could draw a horse with amazing detail and perspective. Normal children aged four drew the horse with much less fidelity, usually side-on and with rough ellipses representing the main body parts. Normal children drew according to their mental schema, Nadia drew what she saw. Another astonishing example of this theory is represented in Temple Grandin [Grandin, 1995, Grandin and Scariano, 1986], an autistic but highly functional person, who has an extraordinary visual intelligence. She herself points out that she stores information in his head as if it were a CD-ROM disk. Her thought patterns work with specific examples, while normal people move from the general concept to specific examples. Often, she struggles to comprehend things that people consider evident.

The reason that leads us to associate autistic savants with computers/robots sheds light as to what makes a robot be observed as such. Typically, we tend to think in terms of what could make a machine more human-like. Here, the approach is to think in terms of what

[1] Recently, a deficit in the activity of "mirror neurons" (MN) has been linked to autism (see [Williams *et al.*, 2001] for example). MNs are active when monkeys perform certain tasks, but they also fire when the monkeys watch someone else perform the same specific task. A similar mechanism has been indirectly observed in healthy humans, but not at the same level in autistic people. Note that, essentially, the phenomenon involves two different perceptions that are "unified" at a neurological level. Such unification may be indicative of the formation or use of the concept ("hand-waving", for example). An explanation based on a difficulty in working with concepts at a conscious level (instead of the detailed perceptions) would also account for this phenomenon as a particular case.

makes a human more machine-like [2]. This is not just unbalanced performance (i.e. abilities that are present in humans but not in the robot). The point is that autistic savants can easily accomplish tasks that require much effort for us (calculus, calendars, etc.) whereas they struggle to accomplish tasks that are trivial and almost automatic for us (like social tasks) [3].

We try to build robots that look and behave like (healthy) humans. Then why do non-autistic people build machines that turn out to be "autistic"? On the other hand, in Artificial Intelligence it is a well-known fact that certain tasks that are trivial for humans are hard for computers/robots and vice versa. But why is it so? The following sections propose an explanation for these two important questions.

2.2 Unconscious Mental Processes

From the discussion in the previous section now it seems clear that if we try to somehow model human intelligence and behaviour we will do it from an unavoidable concept-driven point of view. Some details will remain hidden to our awareness, for it seems that, in healthy adults, an enormous amount of processing is done unconsciously.

Nowadays, the existence of unconscious processes in our brain seems to be beyond doubt [Wilson and Keil, 1999]. Freud's work already acknowledged that unconscious ideas and processes are critical in explaining the behaviour of people in all circumstances. Helmholtz [H. von Helmholtz, 1924], studying vision, pointed out that even basic aspects of perception require deep processing by the nervous system. He argued that the brain constructs perceptions by a process of unconscious inference, reasoning without awareness.

Blindsight [Weiskrantz, 1998] refers to a set of residual visual functions, like the ability to locate a source of light, in cortically ("physically") blind patients. The cortical blindness is a result of the visual cortex's destruction, caused by tumours or other causes. The fact is that patients have those abilities, though they consistently claim not to see the stimuli. Moreover, the abilities are present even in unconscious (comatose) patients.

Face recognition, an ability which obviously falls within the ambit of social intelligence, is another example of unconscious processing. We do not know what features in a face tell us that the face belongs to individual X. Yet, we carry out that process fast and

[2] An interesting aspect to explore is the "Confederate Effect": in the annual Loebner's Contest (a version of the Turing test in which human judges try to distinguish chatbots from humans in a text conversation) some human interlocutors have been considered machine-like [Shah and Henry, 2005].

[3] Note that this confirms the two conditions for building a more "human" robot, already suggested in the previous chapter. That is, a) replication of a large number of human abilities and, b) that these abilities have similar performance levels.

robustly every day. Prosopagnosia is another rare neurological disorder characterized by a specific inability to recognize faces [De Renzi, 1997]. As is the case in blindsight, patients with prosopagnosia show an ability to perceive and process the visual input of which they claim not to be aware. They see faces, though they can not recognize them. These patients are further evidence for the existence of unconscious processing.

In linguistics something similar has also been observed. There is evidence that speakers unconsciously assign a structure in constituents to sequences of words. Besides, the rules that constitute knowledge of the language, i.e. the rules that enable us to produce grammatically correct sentences, are unconscious [Chomsky, 1980]. On the other hand, studying knowledge representation, Bartlett was led to propose the concept of schemata: much of human knowledge consists of unconscious mental structures that capture the generic aspects of the world [Wilson and Keil, 1999].

Many other examples of conscious/unconscious dissociation have been encountered. Researchers have demonstrated that this dissociation is present in perception, artificial grammar learning, sequence learning, etc., see [Cleeremans, 2001] for descriptions of experiments. These findings suggest that the presence of unconscious influences on behaviour is pervasive.

Some authors argue that newborns, unlike adults, perceive the world with every detail, and only with their growing do concepts appear [Snyder *et al.*, 2004]. Infants have eidetic imagery, which tend to disappear as they grow. Experiments show that infants at 6 months can discriminate between monkey faces as well as between human faces, but not after 9 months. They also possess "absolute pitch" (the ability to identify or produce any given musical tone without the aid of any reference tone). Absolute pitch is very rare in adults, although all musical savants and many individuals with autism possess it. The same authors argue that metaconcepts are also developed. Details (or the concepts that constitute the metaconcepts) tend to fade into our unconscious.

In the context of learning, the "conscious competence model" explains the process and stages of learning a new skill (or behaviour, ability, etc.) (see [Kirkpatrick, 1971]). We move through conscious/unconscious levels of learning:

1. Unconscious incompetence: we are incompetent and ignorant of it.
2. Conscious incompetence: we are incompetent but we can recognize our incompetence.
3. Conscious competence: learnings that develop more and more skill and understanding.
4. Unconscious competence: the skill becomes so practised that it enters the unconscious

Property	*Conscious*	*Unconscious*
Engaged by	Novelty, emergencies, danger	Repetition, expected events, safety
Used in	New circumstances	Routine situations
Capacity	Tiny	Huge
Controls	Volition	Habits
Persists for	Tenths of seconds	Decades (lifelong)

Table 2.1: Properties of the cognitive conscious and the cognitive unconscious (taken from [Raskin, 2000]).

parts of the brain. Common examples are driving, sports activities, typing, manual dexterity tasks, listening and communicating. It becomes possible for certain skills to be performed while doing something else, for example, knitting while reading a book. The person might now be able to teach others in the skill concerned, although after some time of being unconsciously competent the person might actually have difficulty in explaining exactly how they do it (the skill has become largely instinctual).

Why are some mental processes unconscious? Unconscious mental processes are fast, allowing us to do things like riding a bicycle without having to think about how to control each of our movements (see Tables 2.1 and 2.2). Some authors contend that it is practice and habituation what makes details go unconscious [Baars, 1988, Mandler, 1984]. Only novel, informative inputs trigger conscious processing (or "mental attention", which is a scarce resource). When events are easily predictable, they are no longer conscious [Berlyne, 1960, Sokolov, 1963]. This actually constitutes an attractive theory of learning: we learn when we habituate to certain details or stimuli that appear repeatedly. As we habituate we have less and less conscious consideration for the stimuli (i.e. we put less mental attention to those redundancies). Being consciousness a scarce -limited- resource, this process would perfectly account for the progress of the individual that learning implies. This mechanism is very similar to the "chunking" technique used in the SOAR candidate unified theory of cognition, see [Newell, 1990]. Chunking allows an agent to identify and store the mental processes that consistently lead to a goal. Later, the agent does not engage in mental processing for that task, but it retrieves the stored pattern automatically. Chunking is meant to produce the "power of law of practice" that characterizes the improvements in human performance during practice.

If practice and habituation is what makes details go unconscious, then social abilities should be relatively more unconscious, as they appear first in life and are always present in the acquisition of other abilities (recall that in Chapter 1 we saw that the Social Situ-

Conscious	*Unconscious*
Explicit cognition	Implicit cognition
Immediate memory	Long-term memory
Novel, informative, and significant events	Routine, predictable, uninformative events
Attended information	Unattended information
Focal contents	Fringe contents (e.g. familiarity)
Declarative memory (facts)	Procedural memory (skills)
Supraliminal stimulation	Subliminal stimulation
Effortful tasks	Spontaneous/automatic tasks
Remembering (recall)	Knowing (recognition)
Available memories	Unavailable memories
Strategic control	Automatic control
Grammatical strings	Implicit underlying grammars
Rehearsed items in Working Memory	Unrehearsed items
Explicit inferences	Automatic inferences
Episodic memory (autobiographical)	Semantic memory (conceptual knowledge)
Normal vision	Blindsight (cortical blindness)

Table 2.2: Some widely used euphemisms for conscious and unconscious phenomena (taken from [B.J. Baars, 2004]).

atedness Hypothesis even proposes that generic intelligence evolves only after social intelligence has developed). In fact, unconscious processes are behind all or part of what we call social abilities, like face and language recognition (as opposed to other mental processes like for example solving a differential equation, which require conscious effort) [Bargh and Williams, 2006]. From [Wheatley and Wegner, 2001]:

> "Much of our behavior in social life is unconsciously automatic. There is evidence that people can respond automatically and unthinkingly to facial expressions, body gestures, hints about a person's sex, ethnicity, or sexual orientation, information about someone's hostility or cooperativeness, and a variety of other social stimuli. People also have unconscious automatic responses to things they like and dislike, from foods or books to ideas and social groups. ... Many of the automatic behaviors we do every day are things of which we are perfectly aware at the outset. We know we are getting in the car and heading off to work, for instance, or we know we are beginning to take a shower. Yet because we have done the act so often driving to work every day, showering every darn year, whether we need it or not we no longer need to think about the act after we have consciously launched it. These behaviors are often acquired skills, actions that become automatic only after significant repetition.

> When we begin to learn an action, such as driving, we think of the action at a very detailed level. We think 'engage clutch, move gear shift down into second, lift left foot off the clutch and right foot onto gas.' Skill acquisition starts off as labored, conscious learning and after consistent, frequent practice becomes more automatic and unconscious. Once the action is well learned, the behavior becomes automatic in the sense that it does not require constant conscious monitoring."

At this point it seems clear that there must be a difference in the process involved in building precision manipulators or calculus machines and that involved in building a robot that can socialize with humans as we do. Note that many other problems in robotics and perception also suffer from our lack of conscious access to the mechanisms involved. However, this effect is comparatively more accentuated in social robotics for, as explained before, social skills appear earlier in life.

2.3 The Reason why Social Robots may not be Robust

When building a social robot, for the most advanced abilities (face detection, face recognition, speech recognition,...) researchers normally resort to machine learning techniques and, particularly, to supervised learning. Machine learning studies computer algorithms that improve automatically through experience [Mitchell, 1997]. In supervised learning, given a training set (examples), algorithms can be built for providing outputs for novel samples (i.e. different from those in the training set). This capacity to learn from experience results in a system that can offer increased efficiency and effectiveness. Most importantly, it is the most appropriate option for cases in which we can not think of a concise relationship between inputs and outputs but there are many labelled examples available from which something can be learned.

Learning algorithms use the available training samples to guide a search for a solution in a hypothesis space H_i. A robust solution has to predict, i.e. produce the correct output for future, unseen samples. Training samples alone are not sufficient for this, for there is a large number of possible hypotheses that fit the training set. This is the essential idea behind Wolpert's *No Free Lunch* theorem (see [Wolpert, 1996]), which states that, on the criterion of prediction performance, there are no reasons to prefer the hypotheses selected by one learning algorithm over those of another.

The perfect fit to a training set does not guarantee low error for future, unseen sam-

ples. What about the available knowledge about the solution? A priori knowledge about the solution to the task allows to define the hypothesis space. We may know, for example, that the solution is a polynomial function. Better still, we may know that the solution is actually a quadratic. The more knowledge we have of the form of the solution the smaller the search space in which our learner will look for a candidate solution:

$$H_0 \supset H_1 \supset H_2 \supset \ldots \supset H^* \tag{2.1}$$

H^* being the objective hypothesis (the solution).

Virtually all "practical" learners employ some sort of complexity penalization technique [Scheffer, 1999], including the method of sieves and Bayesian, Minimum Description Length, Vapnik-Chervonenkis dimension and validation methods [Nowak, 2004]. The basic idea of complexity penalization is to minimize the sum of the error in the training set and a complexity measure of the hypothesis space. The quantity that we are ultimately interested in is the *risk*:

$$R(\alpha) = \int \frac{1}{2}|y - f(\mathbf{x}, \alpha)| dP(\mathbf{x}, y) \tag{2.2}$$

α being the parameters of the learner (to be set). The risk is a measure of discrepancy between the values $f(\mathbf{x}, \alpha)$ produced by our learner and the objective values y produced by H^*.

Theoretical work by Vapnik and others has allowed to obtain upper bounds on this risk [Vapnik, 1995]. With probability $1 - \eta$ the following bound holds:

$$R(\alpha) \leq e_{emp} + \sqrt{\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}} \tag{2.3}$$

where e_{emp} is the empirical risk (the error on the training set), and h a non-negative integer called VC dimension. The VC dimension of a set of functions H_i, a measure of its complexity, is the maximum number of training points that can be shattered[4] by H_i. Note that we can in principle obtain an e_{emp} as low as we want, though we would have to do it by making our hypothesis space more complex (i.e. by having a large number of degrees de freedom). Alternatively, we can consider only simple hypotheses, but then we would obtain large e_{emp}.

[4]for a 2-class recognition problem and a set of l training points, if the set can be labelled in all possible 2^l ways, and for each labelling a member of the set H_i can be found which correctly assigns those labels, then we say that the set of points is *shattered* by that set of functions.

From the definition of VC dimension the hypothesis spaces considered above satisfy:

$$h(H_0) \geq h(H_1) \geq \ldots \geq h(H^*) \tag{2.4}$$

The second term of the right-hand side of Equation (2.3) is monotonically increasing with h (see [Burges, 1998]). Thus, if we call:

$$VC(H_i) = \sqrt{\frac{h(H_i)(\log(2l/h(H_i)) + 1) - \log(\eta/4)}{l}} \tag{2.5}$$

the following holds:

$$VC(H_0) \geq VC(H_1) \geq \ldots \geq VC(H^*) \tag{2.6}$$

In Equation (2.3), e_{emp} is the error obtained in the training set for an individual hypothesis. Let $e_{empMax}(H_i)$ be the maximum training error that can be obtained in H_i (i.e. the maximum training error of all the individual hypotheses in H_i). Then:

$$R(\alpha) \leq e_{emp} + VC(H_i) \leq e_{empMax}(H_i) + VC(H_i) \tag{2.7}$$

Let us assume that the error in the training set for the objective hypothesis H^* is 0. From (2.1):

$$Card(H_0) > Card(H_1) > Card(H_2) \ldots > 1 \tag{2.8}$$

From (2.8) the following holds:

$$e_{empMax}(H_0) \geq e_{empMax}(H_1) \geq \ldots \geq 0 \tag{2.9}$$

From (2.7), (2.6) and (2.9) we can conclude that the risk decreases as we increase the amount of knowledge. With reduced knowledge good performances cannot be guaranteed.

Therefore, in our case in which available knowledge is little, we cannot guarantee good performance for samples different than those tested. Thus, if we obtain low error in a test set we could be actually overfitting to that test test! [5]. The involved inductive process

[5] In the discussion we have not mentioned the effect of the samples used. Obviously, if the test samples are "representative" then the error measured is applicable to the future samples that the system will encounter. As an example, let us suppose that the domain is made up of a very frequent set of samples (set A) plus a very infrequent set of samples (set B). Thus $P(\mathbf{x})$ is much larger for values of $\mathbf{x}$ inside A. Then if we measure low error

is valid in predicting future cases only as long as we are capturing knowledge about the solution, whether intentionally or not.

The conclusion is: in any attempt to reproduce a social ability in a certain domain there is a serious risk of overfitting to the test cases. This means that in practice we will have a situation like that depicted in Figure 2.3.

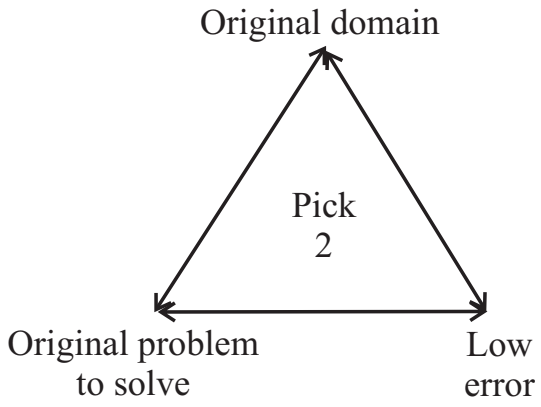


Figure 2.3: Situation when there is little knowledge about the form of the solution. We can be in one of three possibilities: 1) have low error in a domain different than the original, 2) work in the original domain but with large error, or 3) have low error in the original domain but in that case solving a problem different than the original.

The reasoning path taken in this chapter has been the following. First, instead of autism, the condition of autistic savants was seen as more analogous to the social robots that we build, based on the unbalanced intelligences and capacities that are observed. Studies on the origin of that condition show that there are unconscious mental processes that provide healthy adults with the final useful products of reasoning: concepts. Finally, unconscious mental processes, useful as they are, are formally demonstrated to hinder our capacity to reproduce some of our own social capacities. Based on these ideas, namely the unavoidable lack of robustness (specific to tasks that fall into the social range of human abilities) and the fact that we are concept-driven, a way to approach the design of social robots is outlined in the next chapter.

Note that the ideas developed in this chapter constitute a coherent explanation of the well-known fact that certain tasks that are trivial for us are hard for computers/robots and

in the set A we can obviously expect low error in the future. However, such reasoning is essentially equivalent to saying that the more (labelled) samples we know of the domain the more reliable the error figure measured on those samples. But knowing many samples of the domain is actually equivalent to knowing something about the solution function, only that it would be knowledge by enumeration, instead of by comprehension. Having little knowledge (by comprehension) of the solution is just equivalent to knowing only a very small part of the possible samples in the domain.

vice versa. We have seen that, in humans, practice and habituation makes processing details go unconscious. The lack of conscious knowledge of the form of our most "practised" processing algorithms leads to fragile performance in the implementations. On the contrary, for tasks that we carry out with conscious effort we can devise more definite algorithms, which in turn may lead to implementations that may outperform humans in terms of robustness, speed and precision.

Chapter 3

Approach and Architecture

"A scientist discovers that which exists. An engineer creates that which never was."

Theodore von Kármán.

In Chapter 2 we showed that our own consciousness may act as a barrier for obtaining knowledge about the abilities to implement. Given this fact, what could be the most appropriate way to design and build robots with social capacities? After the theoretical discussion in previous chapters here we aim at proposing a number of practical guidelines that should be present throughout the robot design process.

3.1 Analysis *vs* Synthesis

As in many other scientific and technological environments there are two basic approaches that we may follow in our problem: *analytic* and *synthetic*. In the analytic approach we observe, propose models and experiment with them in order to confirm if the model fits the data and predicts new situations. This can be done either from an empirical perspective (carrying out experiments whereby new knowledge can be extracted) or from a deductive perspective. As it was shown in Section 1.1, the analytic approach has been taken by many researchers in social robotics. It entails trying to take advantage of insights taken from disciplines like psychology, ethology, neurophysiology, etc. in the hope of gaining more knowledge and implementing plausible models. The analytic approach is particularly useful for of two reasons:

- We are obtaining knowledge of exactly that what we want to reproduce, which in a sense provides some guarantee that the effort is worthwhile. That knowledge is, to say the least, a good starting point.

- The study of human intelligence and capacities has produced and is still producing invaluable results. Advances in neuroscience have been particularly significant in the last fifteen years, in part thanks to the aid of new exploratory techniques like magnetic resonance imaging.

By contrast, the synthetic approach aims at building artificial systems, see Figure 3.1. The goal is not so much to know how the object is, but how *it should be* to fulfil a series of requirements. This can be done to pursue three goals [Pfeifer and Scheier, 1999]:

- To model biological systems

- To explore principles of intelligence

- To develop applications

As Figure 3.1 shows, analysis and synthesis complement one another. Every synthesis is built upon the results of a preceding analysis, and every analysis requires a subsequent synthesis in order to verify and correct its results.

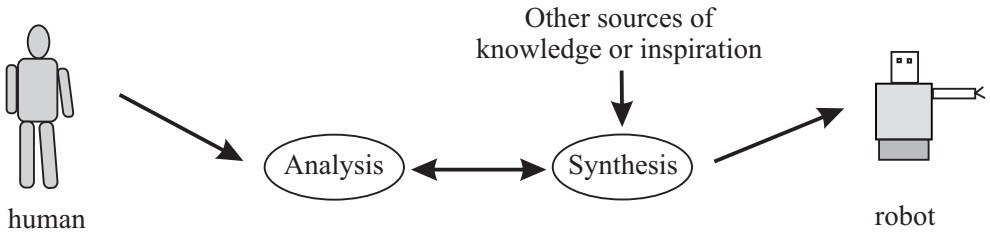


Figure 3.1: Analysis and Synthesis.

In Chapter 1 it was observed that for social robots there may be less guarantee of good performance than in other types of robots. The fact is that, for some types of machines like robotic manipulators, one can extract a set of equations (or algorithms, representations,...) that are known to be valid for solving the task. Such equations would have been obtained after analytical effort, mainly related to kinematics. Once that these equations are stored in the control computer the manipulator will always move to desired points and therefore there is a sort of deductive process involved.

On the contrary, for social tasks, the robot design and development process is more of an inductive process. In inductive processes one cannot have the guarantee of good performance for untested cases. This problem is more accentuated as less knowledge of the causal relation involved is available (a proof of this was given in the previous chapter in the context of machine learning, which is an inductive process).

Thus, the social robot design process can be seen as analogous to machine learning or inductive modelling in general, only that it is the designer (and not the machine) who looks for an appropriate hypothesis (basically one that works well in the cases that he/she tests). In any case the objective of such inductive process is to have good performance for unseen cases.

In machine learning the training set is used to seek an hypothesis. Thus, the cases used for testing the working hypothesis in the robot development process would be the counterpart of the training set in machine learning. If we partition the domain with the typical nomenclature used in both disciplines:

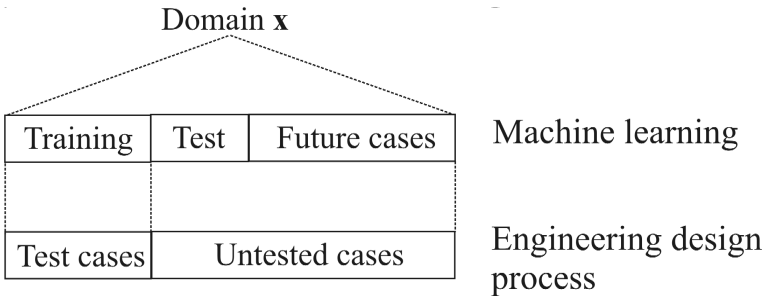


Figure 3.2: Domain partitions.

In machine learning, we have seen that when there is little available knowledge about the solution, the hypotheses to consider may be arbitrarily complex (by contrast, when we have much knowledge the hypotheses that we have to consider are simpler). Complex hypotheses may imply overfitting. In such cases good performance may still be obtained, but only for few specific cases of the domain. This problem would also exist in the robot design process: we would be measuring low error for the tested cases but the error would be large in future, unseen cases.

In machine learning there is a principled way, already mentioned in the previous chapter, to obtain hypotheses that do not overfit: complexity penalization. A well-known complexity penalization technique is Structural Risk Minimization (SRM) [Burges, 1998]. SRM is a procedure that considers hypotheses ranging from simple to complex. For each hypothesis the error in the training set is measured. The best hypothesis is that which mini-

mizes the sum of a measure of its complexity and its error in the training set. In other words, it is the simplest hypothesis that gives an acceptable error figure in the training set.

The same idea could be applied in the broader context of the robot engineering design process. Given the fact that for our problem we have little knowledge about the solutions (this was seen in Chapter 2), a principled way to search for hypotheses would be to start from simple implementations and proceed to more complex ones, each implementation being thoroughly tested. The best implementation should be the simplest one that achieves an acceptable error rate. Although applying such machine learning procedure to the whole robot design process may seem strange, more reasons will be given below to show that it is certainly appropriate.

3.2 Niches

Even though we are trying to emulate the human ability, the robot will always perform in a much restricted environment. In ecology there is an appropriate term for that and we will borrow it here: *niche* [1]. The niche is the range of environmental conditions (physical and biological) under which an organism can exist. For a robot, the niche would include aspects like room sizes, illumination, wall (background) colours, batteries, expected human presence, etc. That is, it is the range of conditions under which the robot is expected to work and perform well.

Ecology distinguish two types of niche. The *fundamental* niche is the total range of environmental conditions that are suitable for existence. The *realized* niche is the range of conditions in which an organism actually performs at a given moment. The realized niche is generally narrower than the fundamental niche, mainly because of the effects of competition and predation from other species, see Figure 3.3 [Hutchinson, 1958].

Humans themselves have their own fundamental and realized niches. For a given social task, for example, humans have a fundamental niche (the range of conditions under which the task is performed well). Obviously, if a human performs well in its fundamental niche then it also performs well in any realized niche inside. What we are trying to achieve in our context is to reproduce the task in the same fundamental niche. We would like, for example, to achieve face detection under the same range of conditions that humans experience (i.e. variations in illumination, distances to people, etc.).

[1] in our discussion we will often resort to terms and notions taken from ecology. Many parallelisms and analogies can be established between robotics and this discipline, and so they turn out to be quite useful for explaining purposes

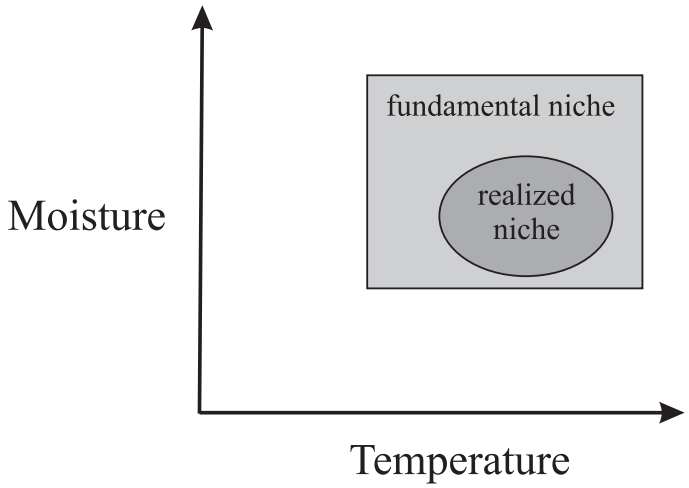


Figure 3.3: Example of fundamental and realized niches. An organism can live under a potential range of moisture and temperature conditions (i.e. they are required for survival). The realized niche is the range of conditions that the organism actually utilizes in its habitat.

The robot, however, is always expected to work in a much more restricted environment, its realized niche. Some particular cases that we ourselves encounter in our daily lives will never appear in the niche the robot is to work in. Obviously, throughout the robot development process we must be interested in minimizing error in its realized niche.

If we try to reproduce the ability as in the fundamental niche, with simple implementations we may obtain improvements in the realized niche. However, due to our limited knowledge, if we go too far and use too much of the available knowledge or intuitions we could be actually worsening performance in the realized niche. That, in fact, is something relatively frequent. Consider again the problem of face detection. If we reflect on how we detect faces, we would say that by detecting the presence of certain features simultaneously, such as skin colour, eyes, mouth, etc. Sophisticated detection systems exist that use these and other features. However, it is our experience that only skin colour blobs with certain width-height ratios are sufficient to detect faces in many practical cases. Using a more elaborated or intuitive algorithm without extensive testing and tuning *in situ* usually leads to many typical faces going undetected (i.e. false negatives).

Note that the search for implementations that work well for the robot's realized niche is not an easy one. We still have the same lack of conscious knowledge about the possible solution to the task. However, the more narrow the realized niche the more relative descriptive value have the available cases used for testing and tuning the system. In the limit, if the realized niche was just a single case, knowing the output value for that case would be obviously sufficient.

Note also that, as long as the available knowledge is still too little, the more of it we utilize the better the performance in parts of the fundamental niche, but also the more "patchy" the performance inside the realized niche. This is due to the fact that in our case there may be overfitting to the samples used for testing (see Section 2.3). Such patchy performance is obviously undesirable.

Ideally, we would like to measure low test error inside the realized niche and at the same time have some guarantee that this measured error applies to the whole niche. The more knowledge we use the more likely it is to measure low test error inside the niche, but there is still little guarantee that such error applies to the whole of it. Thus, large errors may be experienced in untested cases. This could actually mean that the global error experienced in the niche may be very large. Using only basic knowledge that situation will be less frequent. This suggests that basic knowledge will generally be applicable to most of niches, whereas the use of more of the available knowledge could lead to major improvements only in few specific niches. This, in turn, suggests that it would be advisable to make the system as simple as possible.

Note that this is not the same as saying that simpler techniques should perform better than more complex ones. The point is that complex implementations should be carefully adapted to the realized niche. The best way to achieve this is to start from simple implementations. Using an analogy, it would be like a custom suit. The tailor starts by using a standard pattern of sizes (which fits everybody relatively well) and then, through successive fittings with the client, adjusts different parameters to produce a better final product. In the end the suit will fit the client very well, although it will probably not be adequate for other clients.

It is known that, on an evolutionary scale, organisms tend to adapt to its realized niche (see for example [Kassen and Bell, 1998]). If we concede that robot designers may at times take a role similar to evolution, we will have to achieve this too. Therefore, for the kind of robots that we pursue, we see that the role of the robot niche is crucial. The robot designer has to make a significant effort in discovering the opportunities in the robot environment that allow to obtain useful implementations for the desired abilities. This may involve resorting to assumptions and algorithms that may seem even unintuitive. In other words, it may involve discarding intuitive ideas that *seem* to be useful. Obviously, if such opportunities are not found, then the designer will have to define the minimum set of restrictions to the problem that allows to obtain useful implementations.

For the reasons given above, in this work we have adopted a parsimonious and opportunistic synthetic approach. That is, our interest is in developing an application with whatever means. The main source of insight will still be the findings of other disciplines like psychol-

ogy or ethology. However, we will use that knowledge only as long as it allows to obtain practical results. As we have seen, using too much knowledge can be counter-productive. Therefore, in building the robot, we start by considering simple techniques, simple features, simple algorithms, even though they may seem unintuitive. Then, if needed, more elaborated versions are tested. As soon as the results are satisfactory in the working scenario the technique is implemented as final.

Evolution has allowed some species to have organs perfectly adapted to their environment. Actually, there has been no adaptation; only those individuals who were adapted survived. What evolution has made through natural selection, the robot builder has to achieve through a teleological, and also iterative, design perspective. In such approach low-level design aspects become more important. Issues like the robot niche and how the design is adapted to it in order to achieve a desired function are now fundamental, to the point of defining the value of the robot.

3.3 Design and Niche Spaces

Design and *niche* spaces are two concepts introduced by Sloman [Sloman, 1995] that also fit very well in our discussion, see Figure 3.4. The design space is the space of all the possible designs for the robot. Obviously, design spaces are huge and generally very complex, for they may include aspects like part sizes, types of motors, process communications, architectures, etc. The niche space implicitly defines the requirements.

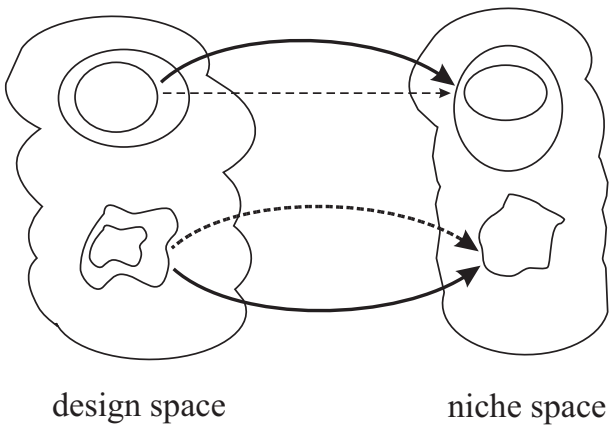


Figure 3.4: Design and niche spaces.

Sloman sees AI as the study of design and niche spaces and their interrelations. One design, for example, may fit a niche in different ways. It could be simply good or bad, or

we could say that it is good with respect to battery charge, or with respect to cost. This is indicated in Figure 3.4 with different styles of arrows. Thus, a design can be considered good for more than one niche. Besides, both spaces can be seen at different abstraction levels (the lowest level of the design space, for example, could be the implementation).

The opportunistic synthetic approach would consist of turning the robot building process into a design space exploration. It would also be an exploration in the niche space, as long as some requirements may be discarded along the process for being too difficult to fulfil. In that case, of course, the less stringent the new requirements the better. Note that we are not referring here to the exploration being made by many authors, through hundreds of papers and built robots. The scale of the exploration we refer to is that of a specific robot, from the early design stages to its final implementation details.

In such exploration we should look for a tight coupling between the robot design and implementation and its niche. The search for such coupling must start from simple versions, which are not associated to a particular niche (or, in other words, that are relatively valid for a large number of niches). Then, the exploration should develop and test new versions that may use knowledge accumulated through analysis but that, above all, exploit the characteristics of the specific niche in which the robot is to work.

In order to take into account both analysis and synthesis, we could use in the discussion both a robot design space and a human design space, see Figure 3.5. The latter would consist of whatever space one could imagine to represent the knowledge acquired about ourselves through analysis (neurophysiological level, cognitive level, etc.). Our human design for accomplishing a certain task (whatever it may be) is such that we achieve good performance under a broad range of circumstances. It is so good a design that we of course can achieve good performance on a narrower range of circumstances, like that of the specific niche of the robot. The design for the robot, however, has to be coupled (in a sense, overfitted) to its niche. This will give good results in that niche, although the solution may seem at times unintuitive or contrary to the knowledge acquired through analysis.

We want to emphasize that this tight coupling is necessary because, in our context, our knowledge about the solution to the desired task remains relatively poor. Should our knowledge be richer, we could provide solutions that do not rely so much on specific features of the niche. In a sense, we have no option but to overfit the system to its niche, which may involve discarding intuitive approaches, algorithms and representations. Much of the designer effort will revolve around this overfitting.

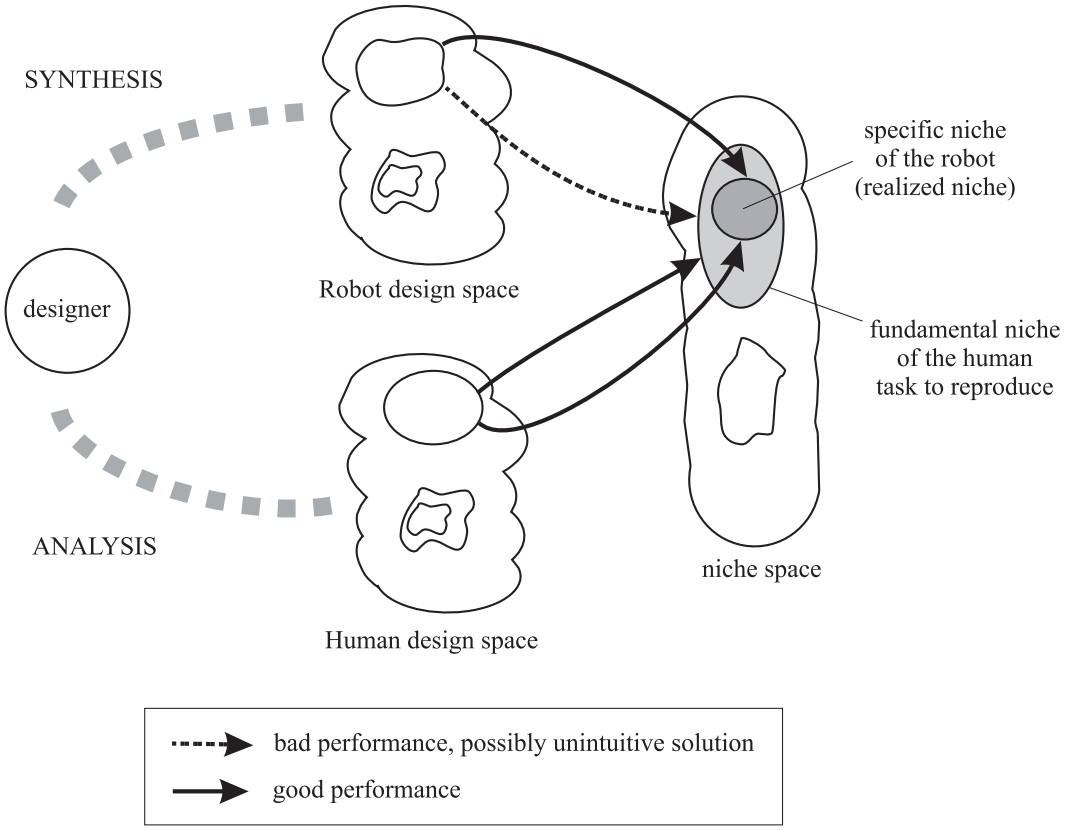


Figure 3.5: Tight coupling to the specific niche of the robot.

3.4 Design Principles

In the previous sections we have argued for a kind of opportunistic synthetic approach. The characteristics of the social robotics problem, described in the first chapters, led us to propose that methodology. It can be summarized as shown in Figure 3.6.

Such approach obviously involves the use of many heuristic ideas, especially those obtained after extensive experience in the hands-on design and testing of social robots. To us, that experience, which includes many engineering insights, is one of the most valuable aspects of the research literature on the topic. In fact, the writing of this document has been heavily influenced by that idea.

Besides the guidelines of Figure 3.6, we consider particularly appropriate for our context the set of principles introduced by Pfeifer and Scheier for designing autonomous agents [Pfeifer and Scheier, 1999]. Those authors, who also advocate for a synthetic approach, have extensive experience in building robots that try to reproduce interesting human capacities. From that experience that they have accumulated, they extracted a set of useful principles,

- Make an effort in discovering **opportunities** for improving performance in the niche of the robot. This can lead to unintuitive implementations (or, in other words, can call for originality).
- Proceed **from simple to complex** algorithms and representations. The final implementation should be **as simple as possible**.
- Perform **extensive testing** of the algorithms and representations **in the niche** of the robot. Adjustments to the (or selection of new) algorithms and representations should be guided by the the results of these tests.
- Treat available **human knowledge very cautiously**, and always following the two previous guidelines. Basic knowledge will almost always be applicable to most niches. Detailed knowledge may be appropriate only for few specific niches.

Figure 3.6: Summary of the opportunistic synthetic approach.

which we summarize in Table 3.1. These principles, in turn, overlap in many aspects with the design principles established by other authors like Brooks [Brooks and Stein, 1994] and Maes [Maes, 1989].

Principle 2 indicates that the robot should be:

- autonomous: it does not require human intervention while it is working (independence of control).
- self-sufficient: it can perform multiple tasks, can exhibit multiple behaviours in the real world over extended periods of time; that is, they do not incur an irrecoverable deficit in any of their resources.
- embodied: it has a physical body embedded in a real environment.
- situated: it has sensors that allow it to acquire information about the environment.

CASIMIRO fulfils to different degrees these four requirements. In particular, its physical body is one of the most important values, for it has a clearly visible impact on visitors. This implies that the whole process should be a hardware-software codesign.

The fourth requirement is in turn related to the principle of sensory-motor coordination (4), which argues that interesting behaviour requires coupling actions to sensory input. This, in turn, involves extensive sensory processing (that, following principle 6, comes from

Principle	*Name*	*Summary*
1	The three-constituents principle	Designing autonomous agents always involves three constituents: 1) definition of ecological niche, 2) definition of desired behaviours and tasks, and 3) design of the agent.
2	The complete-agent principle	The agents of interest are the complete agents, i.e., agents that are autonomous, self-sufficient, embodied, and situated.
3	The principle of parallel, loosely couple processes	Intelligence is emergent from an agent-environment interaction based on a large number of parallel, loosely coupled processes that run asynchronously and are connected to the agent's sensory-motor apparatus.
4	The principle of sensory-motor coordination	All intelligent behaviour is to be conceived as a sensory-motor coordination that serves to structure the sensory input.
5	The principle of cheap designs	Designs must be parsimonious and exploit the physics and constraints of the ecological niche.
6	The redundancy principle	Sensory systems must be designed based on different sensory channels with potential information overlap.
7	The principle of ecological balance	The complexity of the agent has to match the complexity of the task environment.
8	The value principle	The agent has to be equipped with a value system and with mechanisms for self-supervised learning employing principles of self-organization.

Table 3.1: Design principles of autonomous agents proposed by Pfeifer and Scheier.

many channels, both visual and auditive, some of them redundant). The paradigm of active vision [Aloimonos *et al.*, 1987, Ballard, 1991], where movement is considered to be an integral aspect of the perceptual process, is thus very useful here.

In the previous section reasons were given for the importance of carefully adapting the design to the robot niche. The principle of cheap design (5) embodies this aspiration. Although the main justification for the principle seems to be the application of the Occam's razor paradigm, we have already shown that in our context such approach is more than justified and it has been, in fact, one of the most important guidelines in the building of our robot CASIMIRO. The principle of ecological balance (7) is closely related with the principle of cheap design.

The modules that constitute the robot's software are a manifestation of the principle of parallel, loosely coupled processes (3). In this case, it is the simultaneous, asynchronous performance of a number of processes (perception of people, face detection, sound perception, ...) that give the robot its current external appearance and behaviour. There is still a "central" module that decides which actions the robot executes, though many of these processes accomplish high-level tasks on their own. As an example, the audio-visual attention system, itself implemented in a number of modules, is able to modulate the robot's behaviour in a way directly observable by the individuals interacting with the robot.

While it is not used for learning, the value principle (8) is also present in the robot. In fact, it is very important that this robot, whose main function is to interact with individuals, has a way to evaluate the current status of the interaction session. In CASIMIRO the value system is represented mainly in the emotional module, the effect of which is directly visible to the observer. This system allows the robot to have a means of judging what it is good and what it is not in the interaction. Thus it can be considered the basic capacity to develop further abilities.

3.5 Architecture

In robotics, an architecture is the high-level description of the main robot components and their interactions. The choice of an architecture already defines some of the most significant properties of the system. After introducing the design approach and basic principles, in this section we start to take design decisions in earnest. This part of the document will also serve the reader to get an overall impression of the robot software suite, with details left for subsequent chapters.

Social robot architectures generally make use of anthropomorphical models taken from human sciences. Independent of the validity or explanatory power of those models, they are always a good starting point to divide the (often complex) design process into manageable parts. In CASIMIRO we have taken advantage of that aspect and thus we have made extensive use of abstractions like emotions, memory (as an analogy of human memory), habituation, etc.

In this respect, it is important to consider the ideas developed in the previous chapter which showed that healthy people mainly work with high-level concepts. "Robotic" traits are present in individuals who, for one reason or another, are unable to form and work with such high-level concepts or abstractions. Thus, an additional objective would be to endow the robot with enough expertise to recognize and use high-level concepts, generalizations,

interpretations. Note that by high-level concepts we mean those concepts normally used by humans. The abstraction is relative to the concepts that humans have.

Figure 3.7 shows the high-level block diagram of CASIMIRO's architecture.

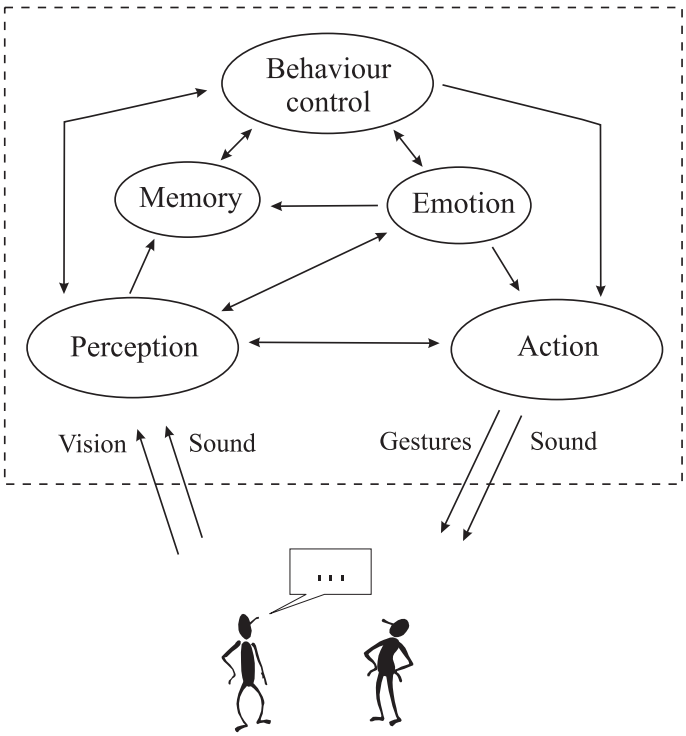


Figure 3.7: Robot architecture.

Note that, much as a "standard suit", the different elements of the architecture are very common in robotic architectures, especially the triad (perception, action , behaviour control). The following paragraphs describe and justify the presence of these and other elements.

Perception

As mentioned in Section 3.4, one of the principles that should guide the robot design is that of sensory-motor coordination. The principle emphasises the use of advanced perceptual and motor capabilities. If we think in terms of high-level concepts, it is straightforward to see that useful high-level concepts can only be identified if the appropriate sensors are used. Poor sensory data can not lead to rich high-level concepts. Therefore, the robot has been given the capacity to process data from two rich perception channels: visual and auditive.

Figure 3.8 shows a block diagram of the perception abilities in CASIMIRO. An important module is devoted to the detection of individuals in the interaction area. This is

obviously a fundamental capacity for the robot, which is expected to engage in interactions with people around it. Individuals may move around the room, and so they are also tracked during the interaction session. Both capacities will be described in Sections 5.1 and 5.4.

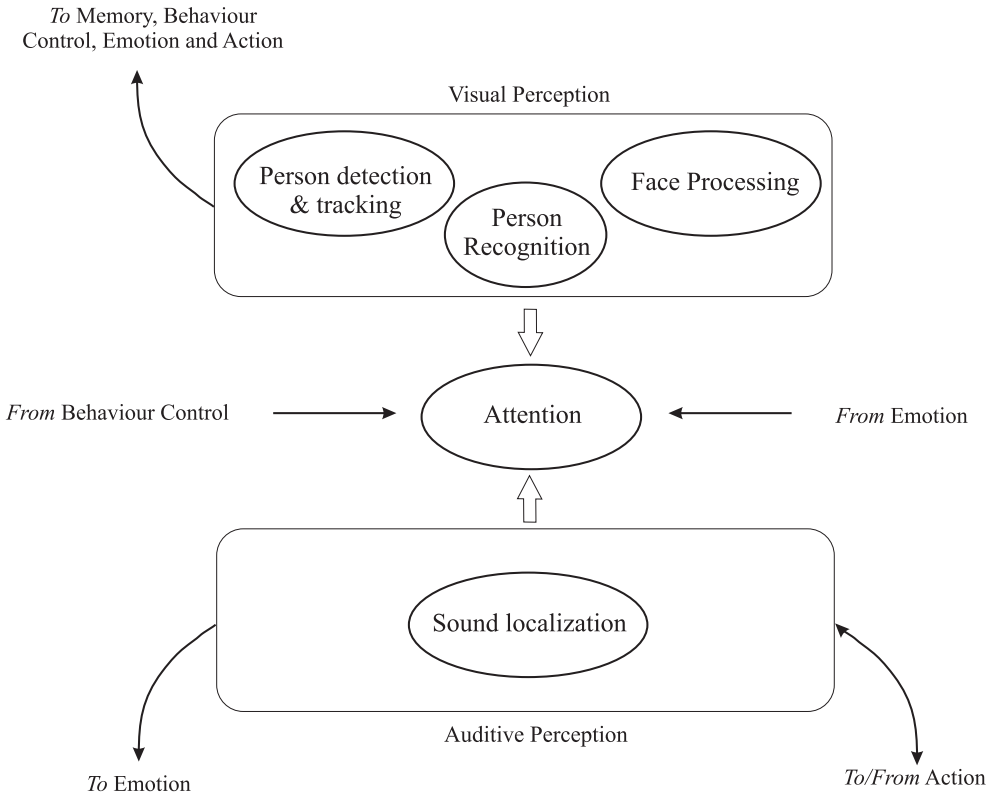


Figure 3.8: Perception.

Face processing is itself divided into two functional modules. One of them simply detects faces in the images taken by the robot "eyes". The other module is devoted to processing the face zone of the input images to detect head nods and shakes. Head nods and shakes are interpreted as yes/no responses to questions made by the robot. Although minimal, this is a valuable capacity, for it allows the robot to have a direct way to know if the interaction is going well. Moreover, it is a capacity that can be fulfilled in practice with good performance. This design decision was actually taken after weighing different niches for the robot. The ideal niche would be that of a normal conversation. Speech recognition, however, did not fulfil our performance requirements in practice and so, following the approach introduced in Chapter 3, we opted for detecting affirmative/negative face gestures. Head nod/shake detection will be covered in detail in Section 5.5.

Note that person and face detection are in a sense redundant perceptions. However, given the fact that people are the most important environment object for the robot, this is

actually desirable. In fact, we are following the principle of redundancy, already mentioned in Section 3.4. The principle emphasizes the introduction of sensory overlap to obtain more robustness in the perception processes. Sound localization and person detection are also redundant in terms of attention catching (see below).

The person recognition module allows the robot to have a memory of some interaction values associated to each individual. The assumption is that the recognition ability has a significant impact on the individual, who perceives that the robot is aware of him/her as a particular individual. Section 5.6 covers the recognition of sporadic visitors, while Section 5.7 describes a robot owner detection method.

With respect to the auditive channel, sound localization should probably be a first ability that should be present. Its importance must not be underestimated. In the animal kingdom it is vital for survival. In the robot, sound localization allows it to respond to people that tries to catch its attention. This, in a sense, is a matter of survival in the robot niche. The sound localization module can trigger off reflex actions and also alter the current emotional status (because of loud sounds, for example). Sound localization can also serve for detecting if the individual in front of the robot (the one the robot is attending to) is talking to it.

When the robot is talking, sound localization events must be discarded. The sound localization module must receive that perception from the robot speech module (see below). Thus, some robot actions are treated as perceptions. In humans this is called *proprioception*, and is used as a feedback for motor control and posture. The sound localization module is described in detail in Section 5.2.

The use of auditive and visual perceptions is very problematic in the sense that a wealth of information is received. Reality imposes a bounded rationality constraint to any intelligent agent. That is, the agent can not process all the information that it possibly receives, if it is to produce a useful response in due time. Therefore, the robot also includes an attention module that acts as a filter. Besides, the fact that the robot can focus its attention to an individual is perceived by that individual (and others in the scene) as a minimum intelligence trait.

Attention is very much dependant on external stimuli (i.e. a loud sound) but it can also be "guided" by behaviour control, as can be seen in Figures 3.7 and 3.8. The current emotional status may also influence attention span. Section 5.3 covers the audio-visual attention module in detail.

We shall mention habituation effects here. When we experience a repetitive stimulus we end up ignoring it. Habituation is another type of filter that serves to discard uninteresting

stimuli. Living beings possess habituation mechanisms that allow them to ignore repetitive stimuli. If such stimuli were not gradually ignored, the continuous response would lead the living being to complete exhaustion. A large amount of information is continually received from the environment, and it has to be somehow filtered so that the agent can focus on the interesting data. Marsland [Marsland *et al.*, 1999] defines habituation as *"a way of defocusing attention from features that are seen often"*. Many animals, and humans too, have some kind of mechanism to filter uninteresting stimuli. Habituation is described in Section 5.8.

Action

Advanced sensory-motor activity not only requires extensive perception abilities. Thanks to a wide repertoire of efferent abilities humans can use high-level concepts such as "run" or "drive a car". Deaf people can not have the same concept "listen" as healthy people. Thus, actuators may also play a role in what we are calling high-level concepts. CASIMIRO has two main action channels: facial gestures and sound (voice), see Figure 3.9.

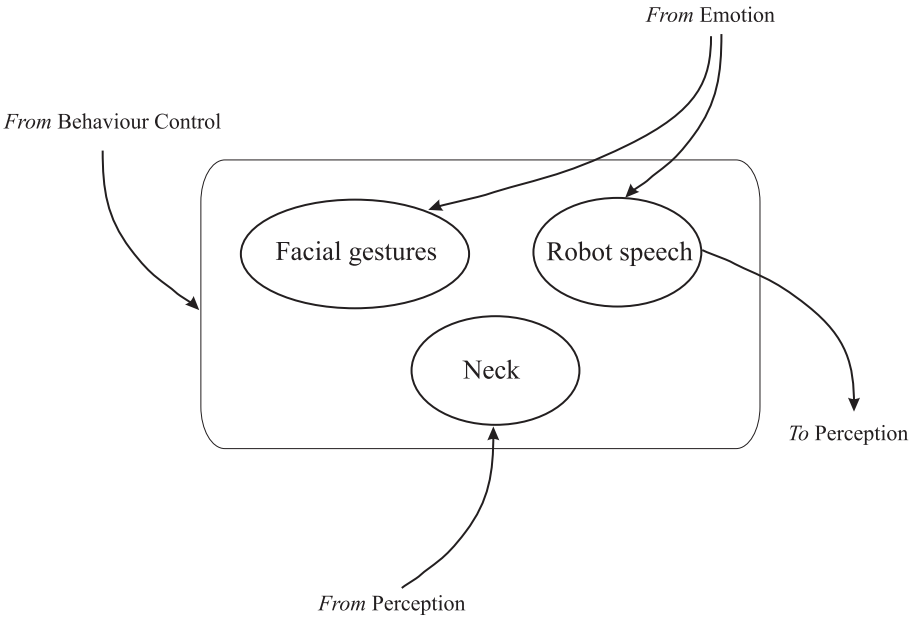


Figure 3.9: Action.

The use of a face for a social robot is not gratuitous. Faces are the centre of human-human communication [Lisetti and Schiano, 2000]. They convey a wealth of social signals, and humans are expert at reading them. They not only tell us identities but also help us to guess aspects that are interesting for social interaction such as gender, age, expression and more. That ability allows humans to react differently with a person based on the information

extracted visually from his/her face [Castrillón, 2003]. There is even a discipline called physiognomy that states that some personality traits can be deduced from the form of the face [Hassin and Trope, 2000]. These ideas suggest that the construction of a social robot must include a face.

In the literature of software agents anthropomorphic representations have been suggested too [Maes, 1994, Laurel, 1990, Oren *et al.*, 1990, Nass *et al.*, 1994]. The assumption is that the anthropomorphic representation allows for a rich set of easily identifiable behaviours and for social interaction [King and Ohya, 1996]. Besides, the work described in [Takeuchi and Naito, 1995, Koda and Maes, 1996, Keisler and Sproull, 1997] suggests that people tend to find agents with human faces more appealing and cooperative than those without it.

The facial expression module developed for CASIMIRO, which is essentially an expression to motor values decoder, is introduced in Section 6.1. The facial expression displayed by the robot is directly related to its current emotional status (according to Picard, in humans facial expression is not the emotion itself, though it is the main medium of manifestation [Picard, 1997]).

The robot also uses speech for expressing itself. Speech characteristics are "tuned" by the current emotional state. This way the message is conveyed with more fidelity. The robot speech module is described in Section 6.3. Within the limits of the available hardware, speech and mouth are synchronized. As indicated above, the robot will also need to know when it is talking so that sound perception can be adapted.

The robot neck is also controlled by an independent module, see Section 6.2. The main function of the neck is turning the robot head toward detected individuals. It can also track them smoothly. Again, its functioning is very much defined by the particular characteristics of the hardware.

Apart from the behaviour control module (see below), these action channels can be controlled directly by perceptions (reflexes) and "tuned" by the emotional state, see Figure 3.7. As an example, some aspects of the robot speech, like pitch, volume and speed, shall be controlled by the current emotional state.

Memory

Without memory no meaning or interpretation can be attributed to things and events that we perceive. Memory is needed to make use of concepts. For this robot, the most important concepts, which must reside in memory, are those related to the interaction evolution. These

data can be associated to each individual that the robot has interacted with (as indicated above, an identification module is able to identify individuals). Examples of this kind of concepts, which we also use in our everyday human-human interactions, may be "INDIVIDUAL_IS_UNCOOPERATIVE" or "I_HAVE_FELT_HAPPY_WITH_THIS_INDIVIDUAL". This allows the robot to have a means of adapting its actions to the evolution of the interaction session or the individual.

Emotions

Although the use of an emotions in a robot is still under debate, in the last years many authors have argued that the traditional "Dr Spock" paradigm for solving problems (eminently rational) may not be appropriate for modelling social behaviours. Rational decisions allow us to cope with the complex world that we live in. Thus, the rational selection among different options is crucial for survival and goal accomplishment. However, any agent whose actions are guided only by purely rational decisions would be in serious trouble. Weighing all the possible options would prevent the agent from taking any decision at all. There is evidence that people who have suffered damage to the prefrontal lobes so that they can no longer show emotions are very intelligent and sensible, but they cannot make decisions [Picard, 1997, Damasio, 1994]. A so-called "Commander Kirk" paradigm assumes that some aspects of human intelligence, particularly the ability to take decisions in dynamic and unpredictable environments, depend on emotions.

There is another interpretation, however, which makes clear the importance that emotion modelling may have in a robot. Social intelligence seems to obviously require emotions. People have emotions, recognize them in others and also express them. A wealth of information is conveyed through facial expressions, voice tone, etc. If robots can recognize, express and probably have emotions, the interaction with the user will be improved because the robot will be able to analyse the affective state of the user and choose a different action course depending on it [Hernández *et al.*, 2004]. Thus, it seems clear that any attempt to imitate human social abilities should consider modelling emotions or affective states. In fact, a field called *Affective Computing* is developing which aims at developing engineering tools for measuring, modelling, reasoning about, and responding to affect.

In CASIMIRO, the emotional module basically maintains an emotional state. The robot is at every moment in one of a predefined set of emotional states, like anger, happiness, surprise, etc. The emotions module has three important functions:

- It is a useful and conceptually simple way of modelling facial gestures and their tran-

sitions

- The emotional state can influence (and can be influenced by) the actions taken by robot
- The emotional state should affect perception, in restricting the attentional span, for example.

Behaviour Control

Having many sensors and actuators is a necessary condition for the robot to be perceived more intelligent or human. Nevertheless, an appropriate behaviour control mechanism must be present to link them at the highest abstraction level possible. The robot could have useful high-level concepts and still use only low-level concepts in its action selection module, which will make it useless from our point of view. Alternatively, it could have only conscious access to the high-level concepts, despite having the low-level concepts somehow represented in some part of it, which seems to be the case in humans.

Behaviour modelling is a central topic in robotics. A robot's observed actions will normally depend on many factors: the complexity and nature of the environment, its perception and action capacities, its goals and its internal processing abilities. When designing the robot, most of these factors are defined *a priori*: the robot is to work in a particular room or zone, its sensors and actuators are known, as well as its goals or preferences. These factors, which represent the robot niche, are therefore fundamental in defining a behaviour control strategy.

The way the robot processes available information internally defines the final observed behaviour. In this respect, two general distinctions can be made from the study of the related literature. First, designers may have to choose between:

- Representing explicitly all the possible cases that the robot may encounter and their associated actions, or
- Implementing complex learning and deduction algorithms so that the robot itself can adapt and improve its behaviour.

The first option is only feasible for the simplest systems, where the number of possible cases is low enough. A slight variation is possible: representing not all the possible perception-action cases, but only a number of important (albeit simple) cases. The complexity of the relationships and interactions between such cases or "mini-behaviours" may lead

to *emergent behaviour*. That is, the robot may show abilities not explicitly represented or considered by its designer.

The second option is particularly attractive because the designer does not need to specify a priori all the circumstances in which the robot may be. If properly programmed, the robot could adapt itself to changes in the environment and show more and more efficient behaviour. This option is thus particularly useful for dynamic and unpredictable environments. Here we should include for example the paradigm of developmental robotics, already mentioned in Chapter 1, which is based on children learning and development. However, in some cases useful behaviour may be obtained only after several trial and error or learning sessions. Also, it may not be clear what are the minimum requirements or algorithms needed to allow the robot achieve certain abilities.

On the other hand, observed behaviour may be considered reactive or deliberative. Reactivity is needed when the robot has to perform actions in a short time, such as those that are required for navigation. In social interaction, response times are also crucial to maintain engagement and believability. Aside from response time, Brooks' subsumption architecture [Brooks, 1986] showed that the interaction of simple reactive perception-action couplings can even lead to relatively intelligent behaviour.

Higher-level actions require deliberative processing. In this case, traditional artificial intelligence techniques are often used in the form of deduction and planning systems, where a search is carried out in an (often huge) space of possibilities. In a sense, the reactive/deliberate distinction is very similar to the representing-all/representing-only-the-minimum-necessary distinction explained above.

After outlining the overall robot architecture, it is important to note that the robot building process is actually hardware-software codesign. In order to understand aspects like facial expression generation and neck control it is essential to understand the robot's hardware, which is described in the next chapter.

Chapter 4

Hardware Overview

"Your face is a book, where men may read strange matters"

William Shakespeare.

Paraphrasing words of Kofi Annan: *intelligence is the only thing that is equally distributed all over the world*. Robotics is nowadays a topic more or less known by the general public. People all over the world get captivated by its appeal. They want to build their own robots, fascinated by Kismet or Cog. DYI books are also beginning to appear [Williams, 2004]. The ingenuity is in most of the cases in the way they manage to fulfil the hardware or software requirements:

> "The main objective of this project is to put all of the mentioned components together and show the possibility of developing such a complex platform from scratch with very elementary and low-cost components with aid of basic tools", H. Mobahi (builder of the robot Aryan [Mobahi, 2003]).

In some cases it is quite an achievement:

> "Well, after seeing all of this, this (my) robot is not so bad, if we bear in mind that it has not been sponsored by any firm, government or university. It is the result of work performed by a Spanish technician keen on robotics and with limited economical resources.", J.L. Martínez (builder of the robot SEGURITRON [Martínez, 2003])

CASIMIRO also shares that spirit. The implementation details are the most difficult part of any robotic project. In the abstract everyone can devise fancy architectures or capabilities, though the implementation is always the ultimate effort. This in fact is one of the ideas of the approach outlined in the previous chapter: fitting the implementation to the robot niche.

It is in the implementation where one has to choose between options and make a decision. For someone involved in developing a project, implementation details are the most rewarding information of the available bibliography or web resources. Often, that information is what keeps them excited:

> "Years ago when I was a wee laddie fresh out of college I would read academic papers on robotics projects procured at some expense and often taking weeks or months to arrive as crude photocopies from the British Library. Papers such as Brooks' *Elephants don't play chess*, together with others by Luc Steels, Inman Harvey, Cynthia Ferrel and many more were all fascinating to read, but there was always an absence of detailed hardware and software information which meant that getting started presented a big obstacle to overcome. It wasn't possible to reproduce the experiments and find out how other people's robots worked, because the details of particular implementations were usually kept secret, or were simply not described out of laziness. To me this lack of detailed description seemed very unscientific, opening up the possibility of researchers making exaggerated or misleading claims about the results of their experiments. So for my own robotics projects I'm determined to make the whole process as transparent as possible, making both source code and details of electronics and physical construction available wherever possible." - Bob Mottram (builder of the robot Rodney [Mottram, 2003]).

Again, CASIMIRO and particularly this document, were designed with those ideas in mind. The hope is that some day they can be useful to other people who want to go a little farther. The following sections describe the hardware of the robot. Details will be in general left out as the information is mainly technical data available elsewhere. It is important to introduce the hardware at this point of the document (instead of considering it an appendix). That helps to define in part the robot niche, to which the rest of the work will have to adhere.

4.1 Hardware

If we are to build an anthropomorphic face, would not it be easier to use a graphical face? The Vikia robot [Bruce *et al.*, 2001], for example, has a flat screen monitor in which an animated face is displayed. Kidd and Breazeal [Kidd and Breazeal, 2003] compared people's reactions to a robot and to an animated (flat) character. The results showed that the robot consistently scored higher on measures of social presence and engagement than the animated character. Subjects also rated the robot as more convincing, compelling and entertaining. On the other hand, Bartneck [Bartneck, 2003] found that embodiment had no significant influence on enjoyability. However, in the robotic character (as compared with a screen character) a social facilitation effect and a high forgiveness for speech recognition errors was observed.

In any case, a physical robot can be viewed from different angles, it can be touched and it is part of the space occupied by people. People expect that moving 3D objects require intelligent control, while flat images likely result from the playback of a stored sequence as in film or television [Mobahi, 2003, King and Ohya, 1996].

CASIMIRO is a physical face: a set of motors move a number of facial features placed on a metal skeleton. It also has a neck that moves the head. The current aspect of the robot is shown in Figure 4.1.

The study in [DiSalvo *et al.*, 2002] tries to look for certain external features of robot heads that contribute to people's perception of humanness. Humanness is desirable, though a certain degree of "robot-ness" is also needed to avoid false expectations. The study analysed 48 robot heads and conducted surveys to measure people's perception of each robot's humanness. The authors of the study give the following 6 suggestions for a humanoid robotic head:

- To retain a certain amount of "robot-ness" the head should be slightly wider than it is tall.
- The set of facial features (nose, mouth...) should dominate the face. Less space should be given to forehead, hair, jaw or chin.
- To project humanness the eyes should have a certain complexity: surface detail, shape, eyeball, iris and pupil.
- Four or more facial features should be present. The most important features are nose, mouth and eyelids.

- The head should include a skin, which can be made of soft or hard materials.
- The forms should be as stylized as possible.

CASIMIRO meets the first four suggestions (albeit without eyeballs). Skin was not used because the robot was always intended to be in a continuous state of development. The sixth suggestion has been also met to a certain degree. In the following sections the different parts of the head and neck and associated hardware are briefly described.

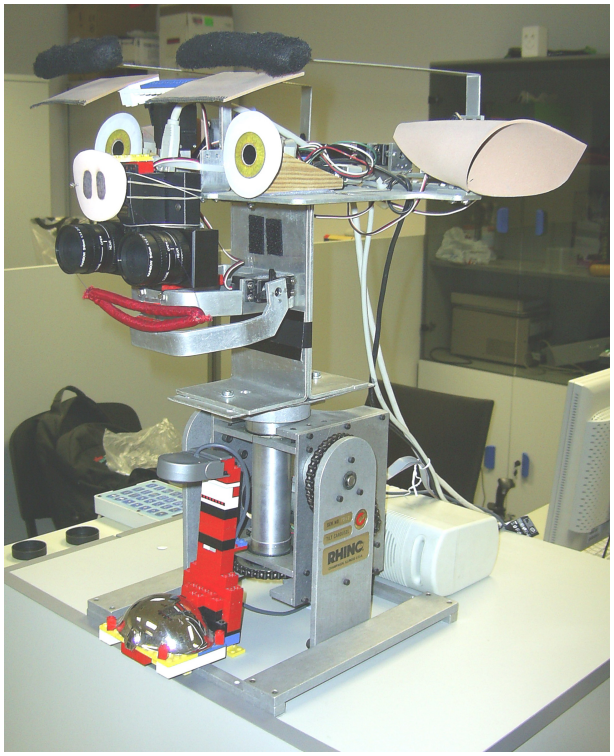


Figure 4.1: Current aspect of CASIMIRO.

Head Design

The mechanical design of the head was divided into four zones: mouth, eyelids, eyebrows and ears. For each zone different solutions were considered. As for the eyebrows, each one has a motion with two rotations in order to allow more expressivity, see Figure 4.2. Eyelids are directly connected to the shaft of one motor (Figure 4.3). The design for the mouth was deliberately made simple, only one motor is needed (Figure 4.4) (currently, a second motor is attached to the lips ends to form a smile). Each ear uses a motor, with the shaft near the centre (Figure 4.5).

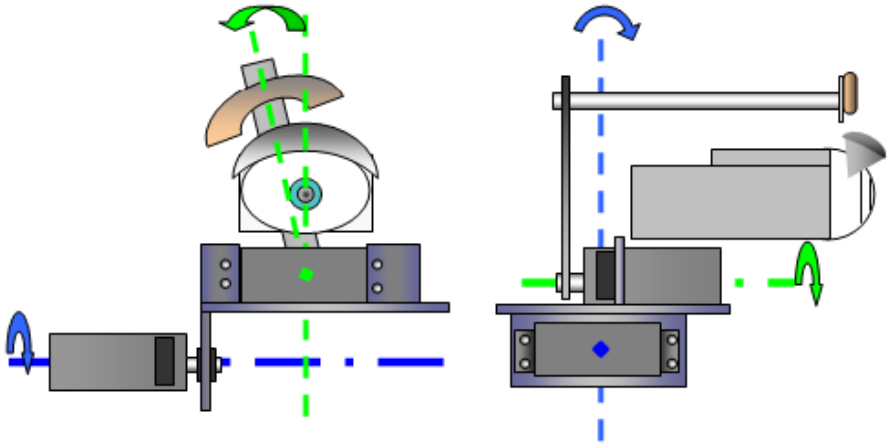


Figure 4.2: Mechanical design of the eyebrows, frontal and side views.

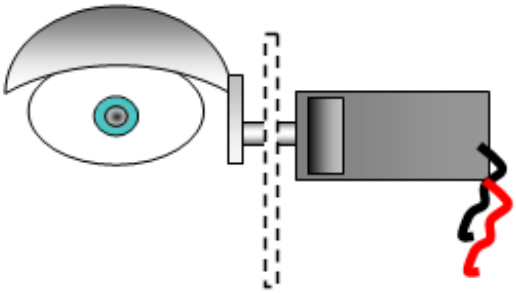


Figure 4.3: Design for the eyelids.

The wireframe design for the head is depicted in Figure 4.6. The head skeleton is made from aluminium. Additional data and the other options considered for each zone of the head are available in [Rageul, 2000].

Motor Control Board

Facial features are moved by 9 FUTABA S3003 servomotors, which have a maximum rotation angle of 180 degrees and a torque of 3.2kg/cm. Servomotors are controlled by an ASC16 control board from Medonis Engineering. It can control the position, speed and acceleration of up to 16 servomotors, and it also features digital and analog inputs. Commands can be sent to the board through a RS-232 interface. Additional details are available at [Medonis Engineering, 2003].

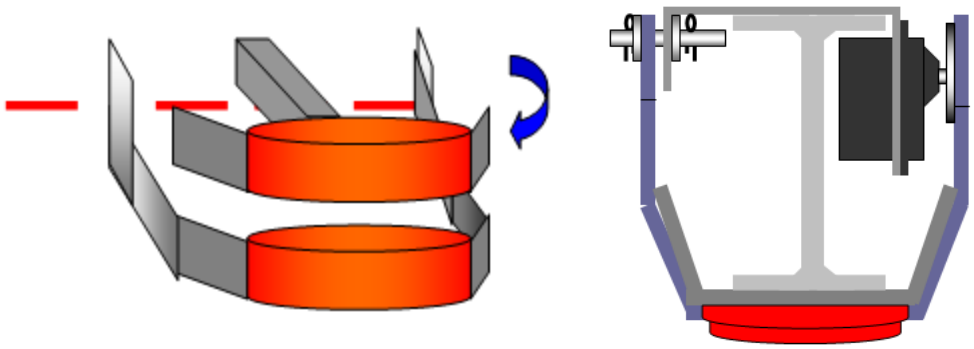


Figure 4.4: Design for the mouth.

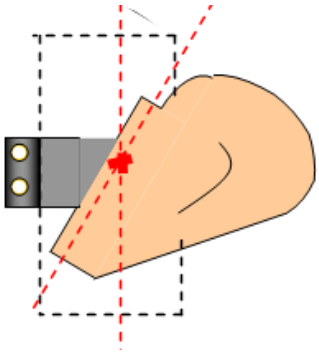


Figure 4.5: Design for the ears.

Neck

The neck of the robot uses hardware from RHINO Robotics Ltd. In particular, the MARK IV controller and the tilting carousel were used [Rhino Robotics Ltd., 2003]. The carousel provides the pan and tilt motion for the neck. Due to the construction of the tilting carousel, pan and tilt motions are not independent. In Figure 4.7 it can be seen that the tilt axis is under the panning surface. In the final version only the pan motion is used.

Omnidirectional Camera

As can be seen in Figure 4.1, there is an omnidirectional camera placed in front of the robot. The device is made up of a low-cost USB webcam, construction parts and a curved mirror looking upwards, see Figure 4.8. The mirror is in this case a kitchen ladle.

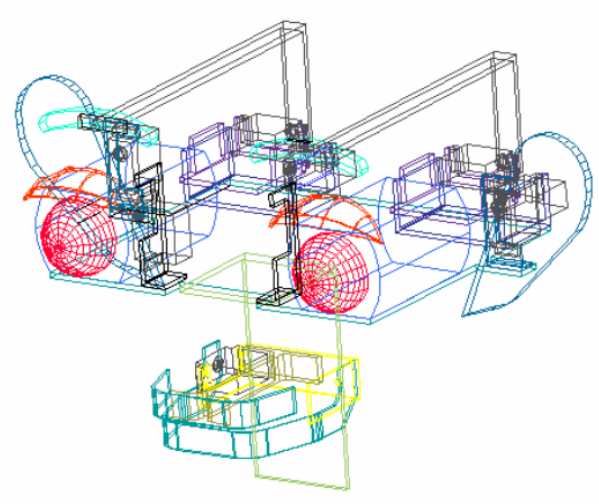


Figure 4.6: Wireframe design of the head.

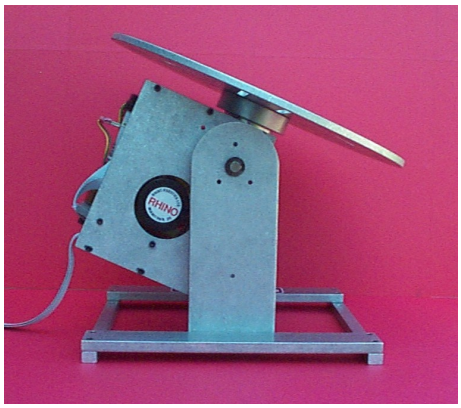


Figure 4.7: Tilting carousel used as a neck. Courtesy of Rhino Robotics Ltd.

Stereo Cameras

Most of the visual tasks accomplished by the robot depend on a pair of cameras placed just above the nose. It is a STH-MD1-C FireWire stereo head from Videre Design. The device was selected because of its high image quality and speed. Also, it includes a library for efficiently obtaining depth maps.

Microphones, Amplifiers and Sound Card

CASIMIRO has two omnidirectional microphones, placed on both sides of the head. Sound signals feed two amplifiers. An EWS88 MT audio system from Terratec is used to capture

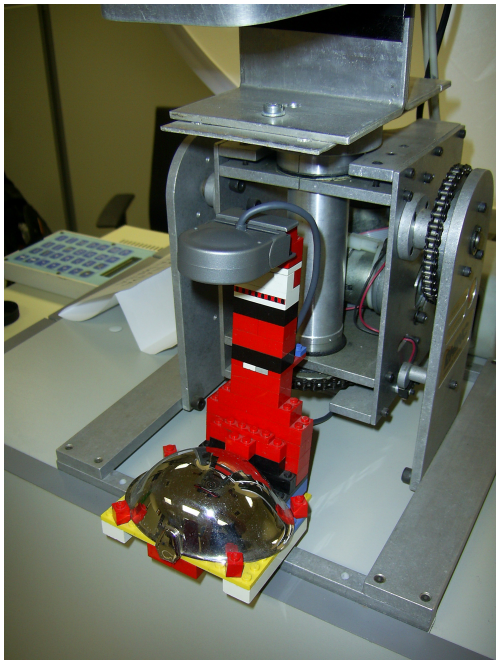


Figure 4.8: Omnidirectional camera.

the stereo sound signal.

Temperature Sensor

It is the case in a laboratory with many researchers and air conditioning that room temperature is not always comfortable for everyone. One of the environmental inputs of the robot is room temperature. The temperature measurement circuit is based on a standard LM35 sensor, and uses an analog input of the ASC16 control board.

4.2 Software Overview

CASIMIRO's software is divided into functional modules. Each module is a Windows 2000 independent application. Windows 2000 is not a real-time operating system, though it was chosen to facilitate software development and camera interfacing. Two PC computers are used for running the modules. The computers are Pentium 4 at 1.4Ghz and have 2 network cards each. They are connected to the local area network of the laboratory and also directly to each other through a 100Mbps hub. The internal connection uses private IP addresses and the routing tables of both PCs are established accordingly.

Each module can communicate with other modules through TCP/IP sockets. A configuration file stores the association between module, IP address of the machine in which it runs, and listening socket. A special module called Launcher performs all the necessary initialization steps and then runs all the modules in both computers. It can also stop all the modules.

Chapter 5

Perception

"I've seen things...seen things you little people wouldn't believe... Attack ships on fire off the shoulder of Orion bright as magnesium... I rode on the back decks of a blinker and watched c-beams glitter in the dark near the Tanhauser Gate... All those moments... they'll be gone."

- Blade Runner, Screenplay, by Hampton Fancher and David Peoples, 1981.

This chapter is devoted to the perceptual capabilities of CASIMIRO. Perception is one of the most important aspects of any robot. What we want is to endow the robot with perceptual intelligence:

> "perceptual intelligence is paying attention to people and the surrounding situation in the same way another person would"
> Alex Pentland, in [Pentland, 2000].

CASIMIRO has a number of modules that help it discern some aspects of the environment, especially those related to people. Section 5.1 describes the omnidirectional vision module. Sound localization is explained in Section 5.2. The products of these two modules are combined in an audio-visual attention system, which is described in Section 5.3. People's faces are detected by a facial detection module, Section 5.4. In Section 5.6 memory and forgetting mechanisms are described. An owner identification module is introduced in Section 5.7. The chapter is concluded with an study on habituation mechanisms.

5.1 Omnidirectional Vision

Typical interaction robots use two types of cameras: a wide field of view camera (around 70º), and/or a foveal camera. Recently, interest in omnidirectional vision has increased in robotics. Omnidirectional vision allows to capture images that span 360º. Four main techniques are being used to achieve this [Fraunhofer Institut AIS, 2004, Nayar, 1998]:

- Cameras with fish-eye lenses, which have a very short focal length (called *dioptric* systems.
- Cameras with curved (convex) mirrors mounted in front of a standard lens (called *catadioptric* systems). This is the most common variant, see Figure 5.1.
- Sets of cameras mounted in a ring or sphere configuration. The information flow is time consuming.
- An ordinary camera that rotates around an axis and takes a sequence of images that span 360º. Mechanical looseness can appear.

In ordinary cameras resolution is uniform, while in catadioptric systems image resolution is higher at the centre and lower in the external zones. Although the distortions introduced in the image are a factor that has to be considered, the advantages of a wider field of view are obvious, especially for certain applications like navigation [Gaspar, 2002, Winters, 2001], surveillance [Boult *et al.*, 1999, Haritaolu *et al.*, 2000] or meeting sessions [Stiefelhagen *et al.*, 2003, Stiefelhagen, 2002, Trivedi *et al.*, 2000]. Some major applications of omnidirectional vision are also described in [Nayar and Boult, 1997].

CASIMIRO is able to localize people entering the room using omnidirectional vision. The omnidirectional camera (a catadioptric setup) shown in Figure 4.8 provides CASIMIRO with a 180º field of view, which is similar to that of humans.

The most similar work to CASIMIRO's person detection and localization system is [Cielniak *et al.*, 2003]. That system is based on a single omnidirectional camera mounted on top of a mobile robot. Using background subtraction people surrounding the robot are detected. The images taken by the omnidirectional camera are warped to produce panoramic images. The angle of the person to the robot is extracted from the horizontal position in the panoramic image. A distance measure is also obtained by extracting three features from the panoramic image and using a trained neural network to produce a numeric value. The features are the person width (extracted from a horizontal histogram of the thresholded

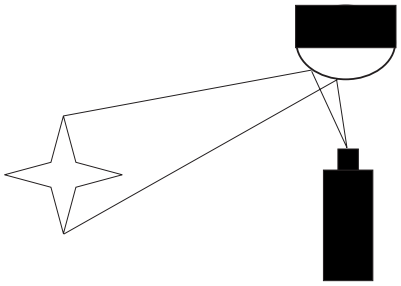


Figure 5.1: Typical omnidirectional vision setup.

panoramic image), the distance between the lower limit of the vertical histogram of the thresholded panoramic image and the bottom edge of the picture, and the total number of pixels in the detected blob. Additionally, a Kalman filter is used to track positions and angles. We believe, however, that this system is excessively complex, which does not fit well with the approach outlined in Chapter 3. We decided to look for a simpler system.

The implemented software is based on adaptive background subtraction. The first step is to discard part of the image, as we want to watch only the frontal zone, covering 180 degrees from side to side. Thus, the input image is masked in order to use only the upper half of an ellipse, which is the shape of the mirror as seen from the position of the camera.

A background model is obtained as the mean value of a number of frames taken when no person is present in the room. After that, the subtracted input images are thresholded and the close operator is applied. From the obtained image, connected components are localized and their area is estimated. Also, for each connected component, the Euclidean distance from the nearest point of the component to the centre of the ellipse is estimated, as well as the angle of the centre of mass of the component with respect to the centre of the ellipse and its largest axis. Note that, as we are using an ellipse instead of a circle, the nearness measure obtained (the Euclidean distance) is not constant for a fixed real range to the camera, though it works well as an approximation, see Figure 5.2. The peak of the 14th sample is due to the fact that distance is calculated from the nearest point of the component, which can produce abrupt variations.

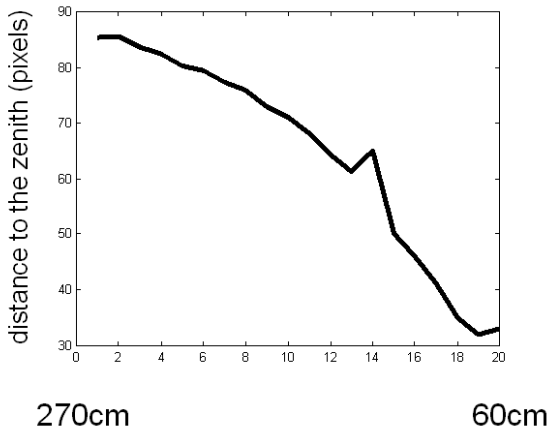


Figure 5.2: Approximate distance measure taken with the omnidirectional camera. In this situation, a person was getting closer to the robot, from a distance of 260cm to 60cm.

The background model M is updated with each input frame:

$$M(k+1) = M(k) + U(k) \cdot [I(k) - M(k)] \tag{5.1}$$

, where I is the input frame and U is the updating function:

$$U(k) = \exp(-\beta \cdot D(k)) \tag{5.2}$$

$$D(k) = \alpha \cdot D(k-1) + (1-\alpha) \cdot |I(k) - I(k-1)| \; ; \; \alpha \in [0,1] \tag{5.3}$$

where α and β control the adaptation rate. Note that M, U and D are images, the x and y variables have been omitted for simplicity. β directly controls the adaptation rate, whereas α controls the effect of motion in the image. For large values of β the model adaptation is slow. In that case, new background objects take longer to enter the model. For small values of β, adaptation is faster, which can make animated objects enter the model. Large values of α give less strength (to enter the model) to zones that are changing in the image. Figure 5.3 shows an example of object assimilation (these images were taken without the mirror of the omnidirectional camera).

Inanimate objects should be considered background as soon as possible. However, as we are working at a pixel level, if we set the α and β parameters too low we run the risk of considering static parts of animate objects as background too. This problem can be alleviated by processing the image D. For each foreground blob, its values in D are examined. The

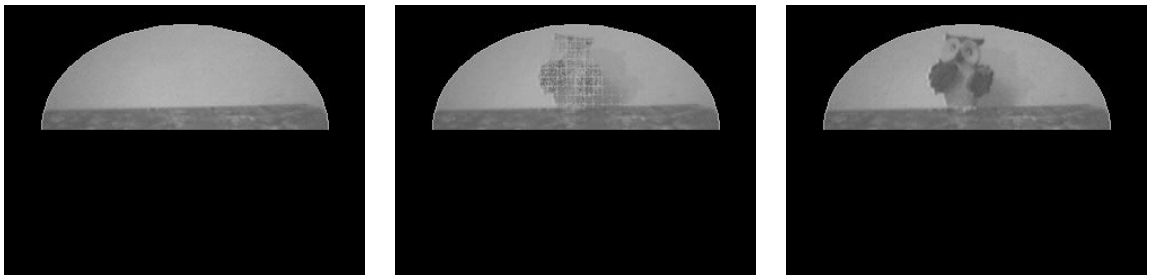


Figure 5.3: Example of how an object enters the background model.

maximum value is found, and all the blob values in D are set to that level. Let the foreground blobs at time step k be represented as:

$$B_i = \{x_{ij}, y_{ij}\} \ ; \ i = 1, .., NB \ ; \ j = 1, .., N_i \tag{5.4}$$

There are NB blobs, each one with N_i pixels. Then, after (5.3) the following is applied:

$$m_i = \max_{j=1,..,N_i} D(x_{ij}, y_{ij}, k) \ ; \ i = 1, .., NB \tag{5.5}$$

$$D(x_{ij}, y_{ij}, k) = m_i \ ; \ i = 1, .., NB \ ; \ j = 1, .., N_i \tag{5.6}$$

With this procedure the blob only enters the background model when all its pixels remain static. The blob does not enter the background model if at least one of its pixels has been changing. Figure 5.4 shows the omnidirectional vision module working.

In initial tests with the robot we saw that people tended to wave his/her hand to the robot. We decided to detect this and make the robot respond with a funny phrase. The contour of each foreground blob is obtained and compared with the contour of the same blob in the previous frame. The comparison is made using Hu moments [Hu, 1962]. When the comparison yields too much difference between the contours (i.e. a threshold is exceeded) a flag is activated. When the face detection module (Section 5.4) detects more than one skin colour blob in the image and a certain amount of motion and the flag is activated then the robot perception HANDWAVING is set to TRUE.

Omnidirectional vision works well in CASIMIRO. However, it is difficult to ascertain that located blobs are people (instead of chairs, for example), and the distance estimation can be too rough for certain tasks. One idea to explore is the use of a laser range finder. The laser

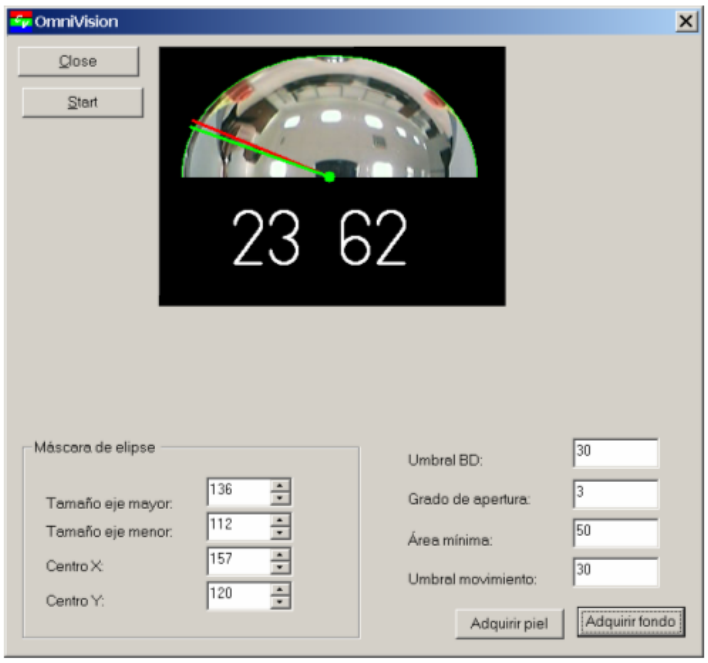


Figure 5.4: Omnidirectional vision module.

range finder is able to give a detailed and frequently updated depth map (a horizontal scan). Similarly to the omnidirectional camera, it could be placed on the table, in front of the robot. Or better still, it could be placed in a lower position. This way, people could be detected as leg pairs, as in [Lang *et al.*, 2003, Bruce *et al.*, 2001]. We believe that, for the task of turning the neck toward people, the use of a laser range finder would be the best option, as only a vector of values need to be processed. On the other hand, we would like to emphasize that the people localization approach used in CASIMIRO has a cost approximately 120 times lower than that of commercial laser range finders.

5.2 Sound Localization

Sound localization plays a crucial role in the perceptive function of living beings, being vital for the survival of many animal species. Barn owls, for example, can hunt in total darkness, thanks to their extraordinary sound localization abilities [HHMI, 1997]. Humans are no exception. Our ability to localize where a sound comes from warns us of potential danger. Sound localization is also an important attention fixing mechanism, especially in a verbal communication setting.

Our sound localization abilities stem from the fact that we have two ears. Although

we could distinguish different sounds with one ear alone, pinpointing where the sounds are coming from requires at least two ears. Reliably localizing a sound source in 3-D space requires even more hearing sensors. Sound differences between the signals gathered in our two ears account for much of our sound localization abilities. In particular, the most important cues used are *Interaural Level Difference* (ILD) and *Interaural Time Difference* (ITD). ILD cues are based on the intensity difference between the two signals. This intensity difference, which can be of up to 20dB, is caused mostly by the shading effect of the head. ITD cues are based on the fact that sound coming from a source will be picked up earlier by the ear nearest to the sound source. This difference will be maximum when the sound source is directly from one side, and minimum when it is in front of the head. Both ILD and ITD cues are dependent on the sound frequency. ITD cues are reliable for relatively low frequencies (up to 1 Khz, approximately), while ILD cues are better for higher frequencies (see [GCAT, 1999] for an explanation of this).

Humans use additional cues for sound localization. The shape of our head and outer ears affect received sounds in a manner dependent on arrival angle and frequency. A model of this process referred to in the literature is the *Head Related Transfer Function* (HRTF). HRTF-based cues allows us to obtain an estimate of the sound source elevation and also to distinguish between sound originating in front of and behind the listener. A more detailed description of sound localization mechanisms can be found in [Blauert, 1983, Yost and Gourevitch, 1987, Hartmann, 1999, GCAT, 1999].

These and other physiological findings have been emulated in computer-microphone systems with relative success. Sound localization can play an important role in human-machine interaction and robot interaction with its environment.

5.2.1 Previous Work

The first important work on a computer sound localization system is [Irie, 1995]. With a combination of hardware and software the system aims to learn to localize sounds in complex environments. The output of the system can be one three values: frontal, right and left. Both ILD and ITD cues are extracted from signals gathered from two microphones and a pre-amplifier circuit. Signals were previously high-pass filtered to remove background noise and then they were divided into segments. For each segment, the cues extracted are: difference of the two maximum positive values, difference in the positions of these maxima, delay between signals (computed by performing a cross-correlation of both signals), difference in the sum of magnitudes of the signals and filterbank-based cues. Filterbank-based cues are computed

by dividing the spectrum of the signals in a number of equally spaced banks, computing the sum of magnitudes in each bank. The cue itself is the difference between the sums of the two signals. 4 banks were used, so the complete feature set used had 8 cues. These cues were fed into a feedforward multi-layer perceptron with three outputs. This network was trained and tested using three sounds (hand clap, spoken "ahh" and door slam). This system is currently working on the *Cog* humanoid robot at MIT [1].

In [Alexander, 1995] a similar system is introduced. The input signals were divided into segments. The extracted cues were: difference between maximum values, difference in the positions of the maxima, correlation and difference in the sum of magnitudes. A classifier was not used, the output of the system was programmed (basically by means of comparing values). This can be a disadvantage in certain settings, because many thresholds have to be manually found (think for example that the difference in intensities could not be exactly zero for a perfectly frontal source, because the two microphones and/or pre-amplifier circuits could have different gains).

A work that used only one ITD cue is described in [Reid and Milios, 1999]. After performing high-pass and low-pass filtering, a signal level test was performed to discriminate between sound and silence. After that, correlation was performed to obtain the ITD estimate and another threshold test was performed on its result (based on the ratio peak/average correlation values). Finally, in order to discard outliers, many estimates were gathered before giving their median value as a response. Correlation was only computed for the possible range of temporal displacement values (as the sound speed is finite, there is a maximum delay possible in the ITD cue, and it depends on the distance between microphones), and this in turn allowed for a faster response. The output of the system was an angle, and it was tested with two types of sound (impulsive sound and speech).

For examples of simulated auditory models or systems that use more than two microphones or special-purpose hardware see [Rabinkin *et al.*, 1996, Härmä and Palomäki, 1999]. For the use of sound localization for robot positioning see [J.Huang *et al.*, 1999, Ryu, 2001].

5.2.2 A Study on Feature Extraction for Sound Localization

In this section the system described in [Irie, 1995] has been used as a base line for comparison, as it uses both ITD and ILD cues and has found practical use. We describe here a new cue extraction procedure that can eliminate some minor errors. The extracted cues for a computer sound localization system are always subject to error because of background noise,

[1] Prof. Rodney Brooks, personal communication

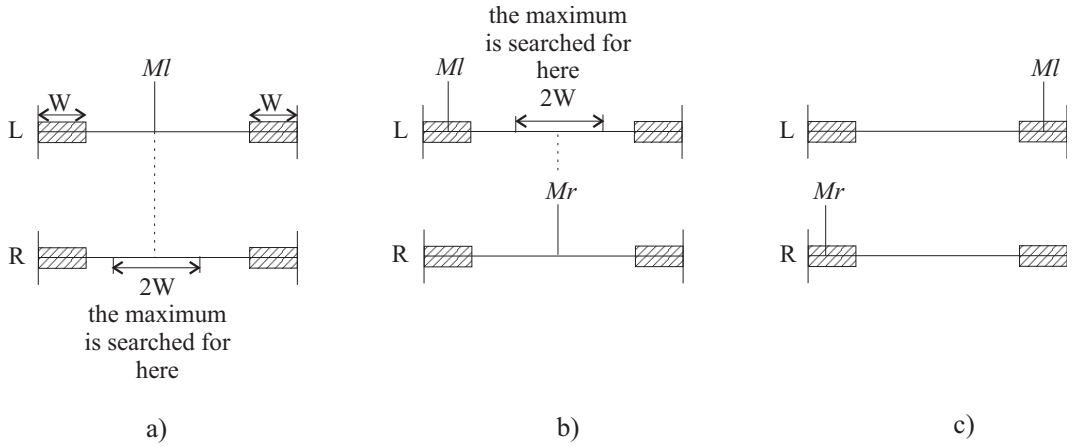


Figure 5.5: a) Ml does not fall in the initial or final "dangerous" zones, b) Ml falls in the "dangerous" zone, c) both Ml and Mr fall in "dangerous" zones. In the last case the sample is discarded.

electrical equipment noise, and specially echoes and reverberation. Echoes originate when sound signals reflect off planar surfaces. Often the effect of multiple reflective sound paths can be as loud or even louder than the sound travelling a direct path from the source.

An important fact to consider is the effect of using segments of the input signals. All systems described in Section 5.2.1 divide the input signal in segments, and extract features from these. However, none of the systems described consider problems that could arise at boundaries. If we consider for example the first extracted cue, difference of maximum positive values, the maximum of signal L (left) could be just at the beginning of the segment. If the source is on the right side, signal L will be delayed with respect to signal R (right). Thus the maximum of signal R is not associated with the maximum in signal L. This in turn affects the second extracted cue, the difference in maximum positions. We propose to extract the first cue as follows. The maximum of signal L is found, be it Ml. Then we search in signal R for the maximum in a zone around the position of Ml. The zone has a length of $2W$, where W is the maximum possible interaural delay. The value of W depends on the distance between microphones and the sound speed. Any (correct) ITD estimate must be equal or lower than W (in absolute value). If Ml falls in the initial zone of the segment, of length W, or in the final zone of the segment, also of length W, it is discarded and we repeat the procedure beginning with signal R. If the maximum of signal R, Mr, also falls in one of these "dangerous" zones, and the zone in which it falls is different from that of Ml, the segment is discarded (no cues are extracted from it). Figure 5.5 shows the three possible cases. This way, some segments are not used for localization, though the first (and second) cues extracted for other segments should be more reliable.

As for the third cue (temporal displacement as extracted from the maximum of the cross-correlation), we propose to use the option already described in [Reid and Milios, 1999], of considering only the maximum in the zone of possible temporal displacement, defined again by W.

Another characteristic of the proposed procedure is related to changes in the volume of the signals. Let us suppose that we want our system to give one of three possible outputs: frontal, left or right. We could fix by hand thresholds for the cue values that define the three regions. Alternatively, these regions could be inferred by training a classifier. In any case, what happens when the input signal has a different intensity (volume) than that used for training/fixing the thresholds? or, what happens when the source gets closer to or farther from the hearing system?. Figures 5.6 and 5.7 show that the value of the extracted ILD cues depend on the volume of the sound signal and the distance of the source.

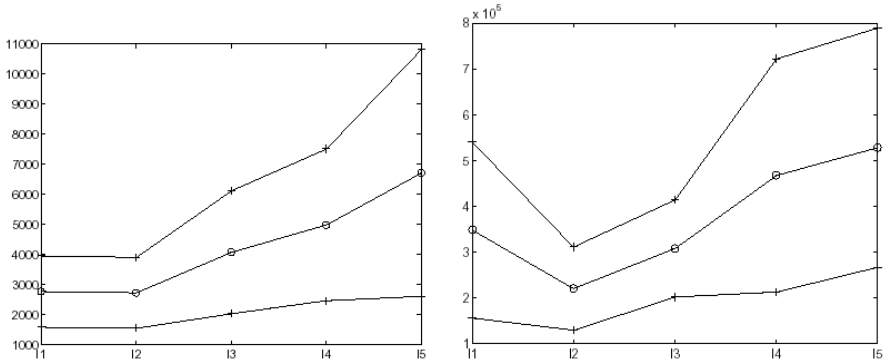


Figure 5.6: Effect of changes in the intensity of the sound signal. The sound source (a mobile phone) is located on the left side of the head at a constant distance. On the left: mean values obtained for cue 1. On the right: mean values obtained for cue 4. The upper and lower lines are the mean values plus and minus one standard deviation.

If we consider for example the first cue extracted, difference between maximum values, the obtained value could be incorrectly discriminated by the fixed thresholds/classifier, because the difference is dependent on the intensity of the input signal. Thus, ILD cues (cues 1 and 4) should be normalized. Let Sl and Sr be the sum of magnitudes for the left signal segment and right signal segment, respectively. Then the two ILD cues should be:

$$C_1 = \frac{Ml - Mr}{Ml + Mr} \;\; ; \;\; C_4 = \frac{Sl - Sr}{Sl + Sr} \tag{5.7}$$

However, normalization can be of advantage only if the differences in volumes are predominant over the error present in the signals. Otherwise it could be worse for the cues extracted. Let $\mathbf{x}$ be the difference between signals L and R. Any extracted ILD cue can

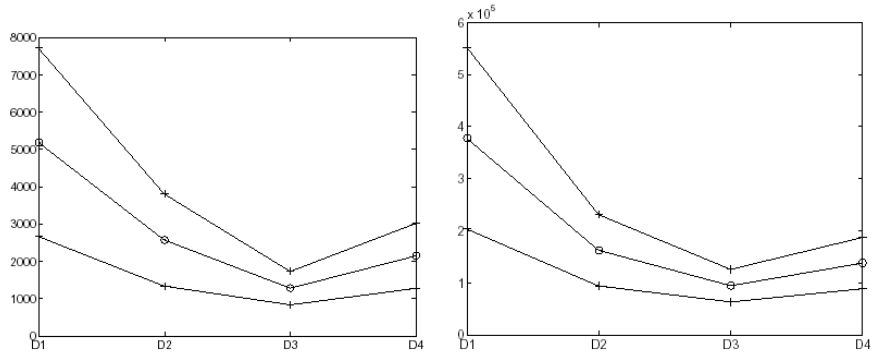


Figure 5.7: Effect of changes in the distance of the sound source. The sound source (a mobile phone) is located at distances such that Di>Di-1. The sound intensity is constant. On the left: mean values obtained for cue 1. On the right: mean values obtained for cue 4. The upper and lower lines are the mean values plus and minus one standard deviation.

be denoted as $C = f(\mathbf{x}) + \varepsilon_f(\mathbf{x})$, the error in the extracted cue being $e = |\varepsilon_f(\mathbf{x})|$. The normalized cue can be expressed as:

$$C_N = \frac{f(\mathbf{x}) + \varepsilon_f(\mathbf{x})}{g(\mathbf{x}) + \varepsilon_g(\mathbf{x})} \tag{5.8}$$

The error is, proportionally, "amplified" if $e_N > \frac{e_y}{|g(\mathbf{x})|}$. From (5.13), it can be shown that this occurs when:

$$|f(\mathbf{x}) - K(\mathbf{x})f(\mathbf{x}) - K(\mathbf{x})\varepsilon_f(\mathbf{x})| - |\varepsilon_f(\mathbf{x})| > 0, \tag{5.9}$$

where

$$K(\mathbf{x}) = \frac{g(\mathbf{x})}{g(\mathbf{x}) + \varepsilon_g(\mathbf{x})} \tag{5.10}$$

$\varepsilon_f(\mathbf{x})$ and $\varepsilon_g(\mathbf{x})$ can have any value and, supposing any volume is possible, so do $f(\mathbf{x})$ and $g(\mathbf{x})$. Thus there exists a possibility that the error be "amplified" after normalization. From (5.9) it can be seen that this possibility is smaller as $\varepsilon_f(\mathbf{x})$ and $\varepsilon_g(\mathbf{x})$ tend to zero. Also note that part of the error is caused by the use of signal segments in which the sound is beginning or ending.

With regard to spectral cues, the frequency banks used in [Irie, 1995] can be useful for modelling the frequency dependence of ILD cues. In any case, they should also be normalized. On the other hand, there exists a strong dependence between the reliability of ITD cues and the frequency of the sound signal, which can produce bad estimates. As

explained in Section 5.2.1, an additional test was used in [Reid and Milios, 1999] in order to reduce such (and other) errors in the ITD estimate. If the ratio between peak and average values of the correlation result was lower than a threshold, the sample was rejected. In our system that test is used too, though the sample is never rejected. If the obtained ratio is lower than the threshold, the value of the ITD cue for the sample is substituted by the last higher-than-the-ratio value obtained. As the sample is not discarded, this allow us to take advantage of the useful ILD information in the sample. The same mechanism was used for the second cue (difference in the positions of the maxima): if the correlation ratio is lower than the threshold, the value of the second cue is substituted by the last second cue value obtained in which the correlation ratio was higher than the threshold.

In order to test the method, sounds were recorded using two *Philips Lavalier* omni-directional microphones, pre-amplifier circuits and a professional sound card (see Section 4.1). A *DirectX* application was developed to integrate all the processing stages: low-pass filtering, sound source detection, feature extraction, data recording and playing (for off-line analysis), and classifying (see Figure 5.8).

In the experiments, the two microphones were placed 28 cm apart on both sides of a custom-made plastic head (see Figure 5.9).

Four different sounds were used in the experiments: hand claps, a call tone from a mobile phone, a maraca and a whistle, see Figure 5.10. The objective was to detect if the sound was coming from the left, right or frontal side. Sounds were recorded in front of the head and at between 35 and 45 degrees on the left and right sides, at a distance of at least one meter to the head. As indicated before, we have compared our feature extraction method with that used in [Irie, 1995], which will be referred to as 'Cog'.

In order to study the reliability of the extracted cues, the ratio between inter-class to intra-class variances will be used as a measure of overlap between samples:

$$r = \frac{\sum_{i=1}^{z}(\mu_i - \mu)^2/z}{\sum_{i=1}^{z}\sum_{j=1}^{n_i}(x_j - \mu_i)^2/n_i} \tag{5.11}$$

This is actually the Fisher criterion for feature evaluation. A classifier was not used because our interest is only in the error present in the individual extracted features. The larger the ratio the better the separation of the samples for a given feature. On the other hand, a number F of consecutive cue vectors was extracted and the mean of them was given as features. The results obtained for F=250 are shown in Table 5.1.

The results obtained with the proposed method achieve in general a higher separation

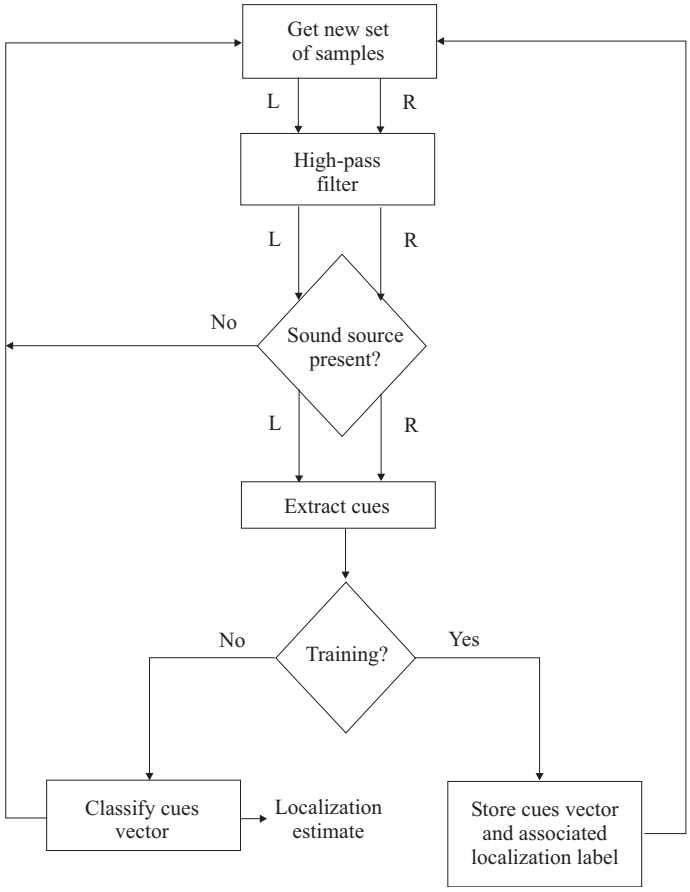


Figure 5.8: Steps performed by the developed sound localization module. The work described in this section focuses on the cue extraction stage.

ratio for the four features used. Note that this results are achieved with a high value for F. In Table 5.2 the same values are shown, for F=0 (cue values are not averaged). In this case, the ratio values for the normalized cues (1 and 4) are worse in the first and second sounds. As F is low, the error is higher and it could be amplified. This reflects negatively in the two first sounds because these sounds contain no significant changes in volume. The other two sounds still give a better ratio because they contain significant changes in volume, as can be seen in Figure 5.10.

The results using the four sounds together appear in Table 5.3, for both F=0 and F=250. Again, there is a significant improvement with the proposed method.

To summarize, this section describes a new method for feature extraction in the context of sound localization. Using the proposed procedure, extracted cues are more reliable, though a reject possibility is introduced. In typical environments changes in the volume of the sound signals are commonplace. Such changes are due to variations in the volume of the signal itself and changes in the distance to the sound source. In the proposed procedure

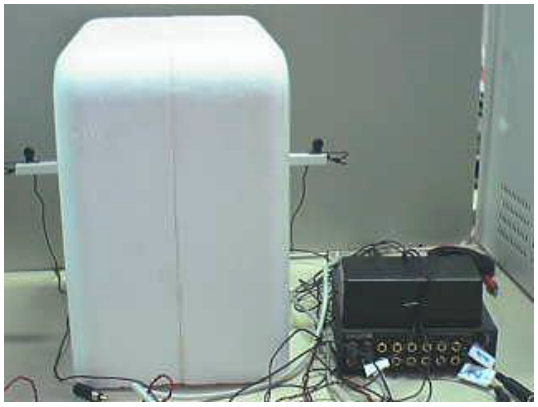


Figure 5.9: Plastic head used in the experiments, next to the sound card external rack and preamplifiers.

Sound	Cue1	Cue2	Cue3	Cue4
phone	6.986	0.561	0.418	4.016
claps	2.286	0.195	1.034	0.396
maraca	6.297	0.163	0.076	3.689
whistling	5.820	2.000	1.368	5.812
phone	**8.187**	**1.374**	0.100	**5.920**
claps	**3.546**	**0.687**	**1.175**	**0.409**
maraca	5.074	**0.890**	**1.884**	**4.093**
whistling	**16.71**	1.762	**2.347**	**17.09**

Table 5.1: Results obtained for F=250. The top half of the table shows the results obtained using *Cog's* system, while the bottom half shows the results obtained with the proposed method. Improved ratios appear in bold.

ILD cues are normalized, which allows for changes in the intensity of the sound signals. As sound intensity decreases with the square of the distance, moving sources are also addressed. Experiments confirm that the method is specially useful when the error in the signals is not too high. For all the sounds used in the experiments the extracted cues show a separation higher than the system used for comparison.

5.2.3 Implementation

Even though the procedure described above shows promising results, it was not implemented in CASIMIRO. In practice, the approach of extracting localization features from the sound signals, be it manually or through any automatic algorithm, does not produce a good performance. A simpler -and above all more robust- option was used in CASIMIRO. The angle of

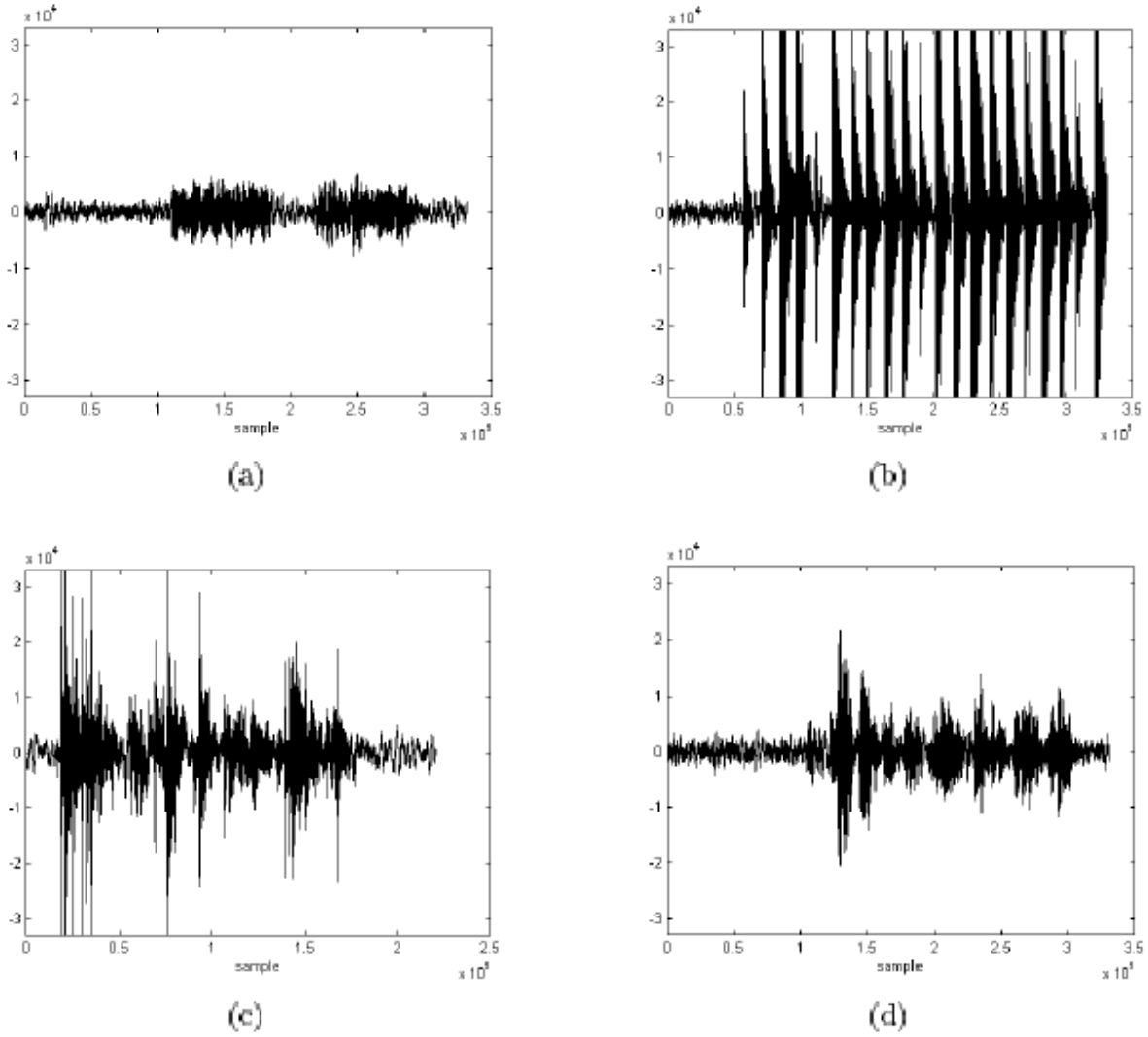


Figure 5.10: Sounds used in the experiments: a) mobile phone, b) hand claps, c) maraca, d) whistling.

the sound source is calculated with:

$$a = \frac{s \cdot \frac{j-M}{44100}}{d}, \tag{5.12}$$

where j is the delay calculated through cross-correlation, s is the sound speed, M is the maximum delay and d is the intermic distance (232 mm). 44100 is the sampling frequency in Hz.

This approach is better suited to CASIMIRO, for the following reasons:

- CASIMIRO's head does not produce a significant shading effect, as it mostly an skele-

Sound	Cue1	Cue2	Cue3	Cue4
phone	0.431	0.001	0.001	0.359
claps	0.015	7e-4	1e-4	0.004
maraca	0.078	0.002	0.004	0.116
whistling	0.141	7e-4	0.006	0.137
phone	0.300	**0.004**	**0.002**	0.296
claps	0.002	**0.004**	**0.001**	2e-4
maraca	**0.119**	**0.015**	**0.014**	**0.239**
whistling	**0.251**	**0.007**	**0.030**	**0.247**

Table 5.2: Results obtained for F=0.

Sound	Cue1	Cue2	Cue3	Cue4
F=0	0.053	2e-4	4e-4	0.071
F=250	0.458	0.097	0.059	0.429
F=0	**0.119**	**0.001**	**0.002**	**0.129**
F=250	**1.059**	**0.099**	**0.166**	**0.607**

Table 5.3: Results obtained considering the four sounds together.

ton.

➔ It gives an angle that can be used for an attentional mechanism (see Section 5.3).

A serious problem for sound localization is the fact that servomotors make noise. Actually, servomotors make noise even when not moving, and noise is also generated by computer fans nearby. In the SIG Humanoid this is solved by using four microphones, two inside the head cover and two outside, and cancelling components of the internal signals in the external signals [Nakadai *et al.*, 2000]. This ingenuous solution depends on the head cover attenuating the sounds. In earlier versions of CASIMIRO the microphones were located in the head, right below the eyes, and the sound localization module only produced results when no servomotor was moving. This is obviously the simplest option. Currently, the two microphones are located at a distance of approximately 80 cm from the servomotors. There is a wooden box (the head is on this box) between the head and the microphones. In their present position the microphones do not capture the noise of the motors, and still capture voice and other sounds coming from the interaction space.

5.3 Audio-Visual Attention

The most important goal for social robots lies in their interaction capabilities. An attention system is crucial, both as a filter to centre the robot's perceptual resources and as a mean of letting the observer know that the robot has intentionality. In this section a simple but flexible and functional attentional model is described. The model, which has been implemented in CASIMIRO, fuses both visual and auditive information extracted from the robot's environment, and can incorporate knowledge-based influences on attention.

In [Kopp and Gärdenfors, 2001] the authors argue that a robot with attention would have a minimal level of intentionality, since the attentional capacity involves a first level of goal representations. Attention is a selection process whereby only a small part of the huge amount of sensory information reaches higher processing centres. Attention allows to divide the visual understanding problem into a rapid succession of local, computationally less expensive, analysis problems. Human attention is divided in the literature into two functionally independent stages: a preattentive stage, which operates in parallel over the whole visual field, and an attentive stage, of limited capacity, which only processes an item at a time. The preattentive stage detects intrinsically salient stimuli, while the attentive stage carries out a more detailed and costly process with each detected stimulus. The saliency values of the attentive stage depend on the current task, acquired knowledge, etc [Heinke and Humphreys, 2001, Itti and Koch, 2001].

Kismet included an attention system based on Wolfe's *"Guided Search 2.0 (GS2)"* model [Wolfe, 1994]. GS2 is based on extracting basic features (color, motion, etc.) that are linearly combined in a saliency map. In a winner-take-it-all approach, the region of maximum activity is extracted from the saliency map. The focus of attention (FOA) will then be directed to that region.

It is a well accepted fact that attention is controlled both by sensory salient and cognitive factors (knowledge, current task) [Corbetta and Shulman, 2002]. The effect of the lower level subsystem (bottom-up influence) has been comprehensively studied and modelled. In contrast, the effect of higher level subsystems (top-down influence) in attention is not yet clear [Itti, 2003]. Hewett [Hewett, 2001] also suggests that volitive processes should control the whole attention process, even though some of the controlled mechanisms are automatic in the human brain. Therefore, high-level modules should have total access to the saliency map. This would allow the attention focus to be directed by the point that a person is looking at, deictic gestures, etc. Fixations to the point that a person is looking at are useful for joint attention. In [Scassellati, 2001] an additional feature map is used for the purpose of

assigning more saliency to zones of joint attention between the robot and a person.

In the third version of Wolfe's *Guided Search* [Wolfe and Gancarz, 1996] high-level modules act in two ways. On the one hand they can modify the combination weights. On the other hand, they can also act after each fixation, processing (recognizing, for example) the area of the FOA, after which an "inhibition of return" (IR) signal is generated. IR is a signal that inhibits the current FOA, so that it will not win in the saliency map for some time.

Top-down influences on attention are also accounted for in the *FeatureGate* model [Driscoll *et al.*, 1998]. In this model, a function is used to produce a distance between the low-level observed features and those of the interest objects. In [Milanese *et al.*, 1994] the top-down influence is embedded in the changing parameters that control a relaxation and energy minimization process that produces the saliency map. Also, in [de Laar *et al.*, 1997] a neural network, controlled by high-level processes, is used to regulate the flow of information of the feature maps toward the saliency map. A model of attention similar to that of Kismet is introduced in [Metta, 2001] for controlling a stereo head. Besides the feature maps combination (colour, skin tone, motion and disparity), space variant vision is used to simulate the human fovea. However, the system does not account for top-down influences. Moreover, it uses 9 Pentium processors, which is rather costly if the attention system is to be part of a complete robot.

In [Grove and Fisher, 1996] an attention system is presented where high-level modules do influence (can act on) the whole saliency map. When, after a fixation, part of an object is detected, saliency is increased in other locations of the visual field where other parts of the object should be, considering also scaling and rotation. This would not be very useful in poorly structured and dynamic environments. In the same system, a suppression model equivalent to IR is used: after a fixation the saliency of the activated zone is decreased in a fixed amount, automatically.

Following the methodology exposed in Chapter 3, the objective for our robot was not to achieve a biologically faithful model, but to implement a functional model of attention for a social robot. The next section describes the proposed attention system.

5.3.1 Attention Model

In all the citations made above, the effect of high-level modules is limited to a selection or guiding of the bottom-up influence (i.e. combination weights) and the modification of the relevance of the object in the FOA. We propose that the influence of high-level modules on attention should be more direct and flexible. Inhibition should be controlled by these

modules, instead of being an automatic mechanism. The following situation is an example of such case: if I look at a particular person and I like her, inhibition should be low, in order to revisit her soon. There could even be no inhibition, which would mean that I would keep on looking at her. Note that by letting other processes control the saliency map joint attention and inhibition of return can be implemented. Also, the mechanism explained before that increases saliency in the zones where other parts of objects should be can be implemented. In fact, any knowledge-directed influence on attention can be included.

As stated before, the objective of this work was to conceive a functional attention mechanism that includes sound and vision cues. Therefore, the model proposed here is simple to implement, being the most complex calculations done in the feature extraction algorithms. The activation (i.e. saliency) values are controlled by the following equation:

$$A(p,t) = \sum_i F_i(v_i \cdot f_i(p,t)) + \sum_j G_j(s_j \cdot g_j(p,t)) + K \cdot C(p,t) + T(p,t) \qquad (5.13)$$

where F and G are functions that are applied to the vision-based (f_i) and sound-based (g_j) feature maps in order to group activity zones and/or to account for the error in the position of the detected activity zones. Spatial and temporal positions in the maps are represented by the p and t variables. v_i, s_j and K are constants. C is a function that gives more saliency to zones near the current FOA: $C(p,t) = e^{-\gamma|p-FOA(t-1)|}$. $T(p,t)$ represents the effect of high-level modules, which can act over the whole attention field. The maximum of the activation map defines the FOA, as long as it is larger than a threshold U:

$$FOA(t) = \begin{cases} \max_p A(p,t) & if \max_p A(p,t) > U \\ FOA(t-1) & otherwise \end{cases} \qquad (5.14)$$

The model is depicted in Figure 5.11, using sound and vision for extracting feature maps. Note that a joint attention mechanism would use the component T of Equation 5.13, which for all practical purposes is equivalent to the approach taken in [Scassellati, 2001] that used a feature map for that end.

The implementation in CASIMIRO uses an auditive feature map: the localization of a single sound source. Notwithstanding, this scheme can be used with multiple sources, as long as they are separated by another technique.

The visual feature map is extracted from images taken with the omnidirectional cam-

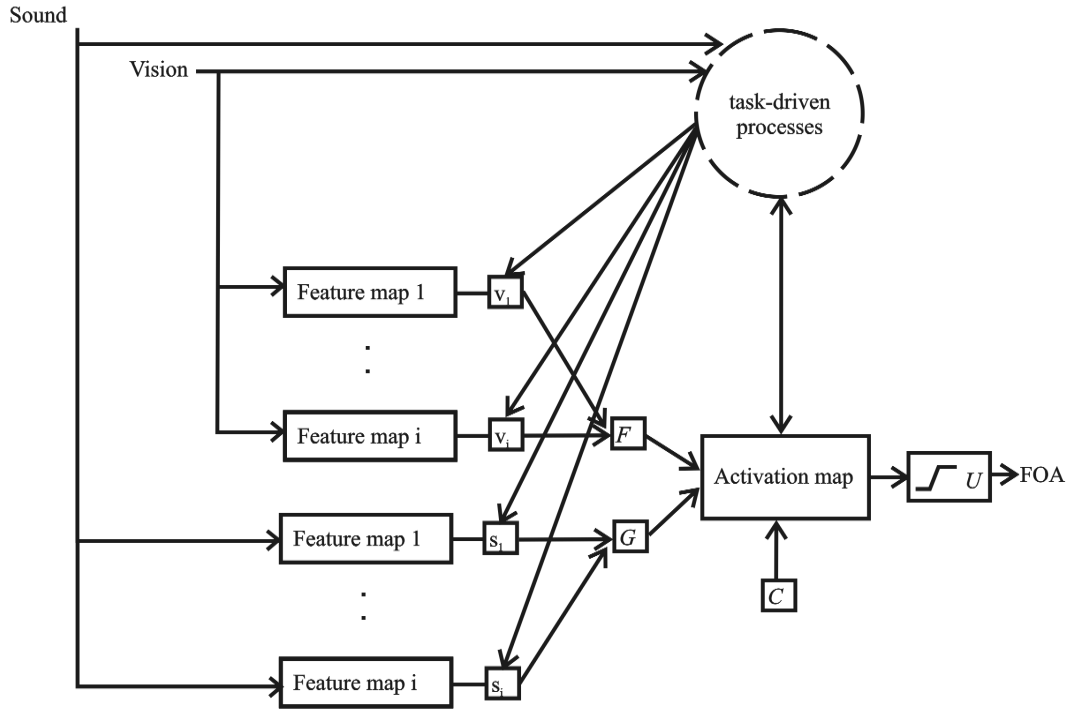


Figure 5.11: Model of attention. The feature maps must represent the same physical space than the activation map. If sensors do not provide such values, a mapping would have to be done.

era, see Section 5.1. As for the sound-based feature map, the aim was to detect the direction of sound sources (i.e. people talking or making noise), see Section 5.2.

5.3.2 Implementation and Experiments

The attention model shown above was implemented in CASIMIRO. The feature and activation maps represent a half-plane in front of the robot. The FOA is used to command the pan and tilt motors of the robot's neck. For our particular implementation it was decided that sound events should not change the FOA on their own, but they should make the nearest visual event win. Also, as a design decision it was imposed that the effect of sound events should have precedence over the effect of C.

In our particular case the variable p takes values in the range $[0, 180]$ degrees and F will not be used. $v_1 = 1, f_1 = \{0, 1\}$ represents the effect of a visual feature map that detects foreground blobs. The visual feature maps are not actually 1-D, but 1 1/2-D, as for

each angle we store the height of the blob, measured by the omnidirectional vision system. This height can be used to move the tilt motor of the robot's neck. $g_1 = \{0, 1\}$ represents the output of the sound localization routine. The vision and sound localization modules communicate with the attention module through TCP/IP sockets, see Section 4.2. In order to account for errors in sound localization, G is a convolution with a function $\exp^{(-D\cdot|x|)}$, D being a constant. In order to meet these conditions the following should be verified:

- $s_1 < 1$ (the FOA will not be directly set by the sound event).
- Suppose that 2 blobs are anywhere in the activation map. Then a sound event is heard. One of the blobs will be closer to the sound source than the other. In order to enforce the preferences mentioned above, the maximum activation that the farthest blob could have should be less than the minimum activation that the nearest blob could have. This can be put as $1 + K + s_1 \cdot e^{(-D\cdot a)} < 1 + K \cdot e^{(-180*\gamma)} + s_1 \cdot e^{(-D\cdot b)}$, b and a being the distances from the blobs to the sound source, the largest and the shortest one, respectively. That equation does not hold for $b < a$ but it can be verified for $b < a - \epsilon$, with a very small ϵ.

Operating with these two equations the following valid set of values was obtained: $D = 0.01, K = 0.001, s_1 = 0.9, \gamma = 0.15$. For those values $\epsilon = 0.67$ degrees, which we considered acceptable. The effect of high-level processes (T) will be described in Chapter 7, Section 7.3.

Additionally, a third feature map was implemented. It accounts for visual events that should change the robot's focus of attention, like a person who gets closer to the robot or is walking around. This third feature map has the same effect over attention than sound events. These events are detected by the omnidirectional camera system: the blob distance to the centre of the image (i.e. the rough measure of distance) is continually checked. When it changes significantly it produces an event of this type. This way, the cases of getting closer and raising the arm catch the robot's attention.

The simplicity of the model and of the implementation make the attention system efficient. With maps of 181 values, the average update time for the activation map was 0.27ms (P-IV 1.4Ghz). In order to show how the model performs, two foreground objects (two people) were brought near the robot. Initially, the FOA was at the first person. Then the second person makes a noise and the FOA shifts, and remains fixating the second person. In order to see what happens at every moment this situation can be divided into three stages: before the sound event, during the sound event and after the sound event.

Figure 5.12 shows the state of the feature maps and the activation map at each stage. Note that the vertical axis is shown in logarithmic coordinates, so that the effect of the C component, which is very small, can be seen. Exponential contributions thus appear in the figures as lines.

Before the sound event the FOA was at the blob on the left, approximately at 75 degrees, because it is the closest blob to the previous FOA (the robot starts working looking at his front, 90 degrees). This is shown in the first two figures. The two next figures show the effect of the sound event. The noise produces a peak near the blob on the right (the person). That makes activation rise near that blob, which in turn makes the blob win the FOA. The last two figures show how the FOA has been fixated to the person. In absence of other contributions the effect of the C component implements a tracking of the fixated object/person.

To summarize, an attentional system is a necessary module in a complex human-like robot. With it, the robot will be able to direct its attention to people in the environment, which is crucial for interaction. In this section a simple yet functional model of attention has been described, drawing upon previous attentional systems for robots. The model was implemented using both auditive and visual features extracted from a zone surrounding the robot. Visual features were extracted from video taken with an omnidirectional camera, which gives the robot a 180 degrees attentional span.

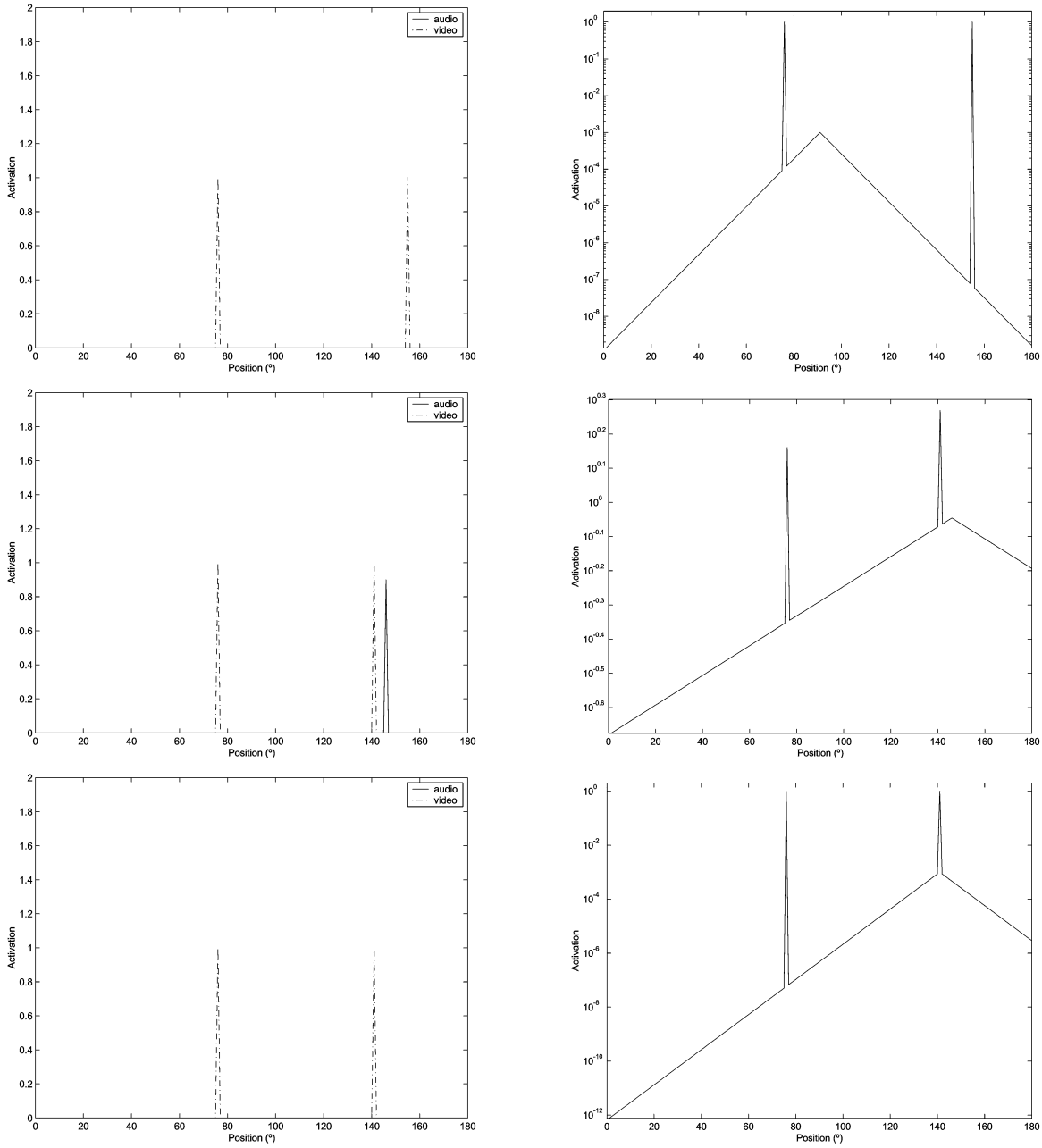


Figure 5.12: State of the feature and activation maps. On the left column the figures show the visual and auditive feature maps. On the right column the figures show the resultant saliency map.

5.4 Face Detection

Omnidirectional vision allows the robot to detect people in the scene, just to make the neck turn toward them. When the neck turns, there is no guarantee that omnidirectional vision has detected a person, it can be a coat stand, a wheelchair, etc. A face detection module that tries to detect people is described in this section. It uses colour images taken by the stereo camera described in Section 4.1. The module was originally developed for face detection and normalization integrated in a platform designed for general purpose artificial vision applications known as DESEO [Hernández Tejera *et al.*, 1999].

5.4.1 The ENCARA Face Detector

Face detection systems described in the literature can be classified by attending different criteria. One of them is based on the use of knowledge employed by these systems: implicit or explicit. The first group focuses on learning a classifier from a set of training samples, providing robust detection for restricted scales and orientations at low rates. These techniques perform with brute force, without attending to some evidences or stimuli that could launch the face processing modules, similar to the way some authors consider that the human system works [Young, 1998]. On the other hand, the second group exploits the explicit knowledge of structural and appearance face characteristics that could be provided from human experience, offering fast processing for restricted scenarios.

ENCARA [Castrillón, 2003], the face detection module integrated in CASIMIRO, merges both orientations in order to make use opportunistically of their advantages and conditioned by the need of getting a real-time system with standard general purpose hardware. ENCARA selects candidates using explicit knowledge for later applying a fast implicit knowledge based approach.

Classification is the crucial process in face detection. There are multiple possible solutions that, roughly speaking, can be divided into two groups: Individual and Multiple classifiers. The complex nature of the face detection problem is easily addressed by means of an approach based on multiple classifiers. The architecture for combination of classifiers used in ENCARA follows [Viola and Jones, 2001] and is sketched in Figure 5.13. However, there is a main difference in relation to that work where the classifiers are based only on rectangle features [Viola and Jones, 2001], in this model the different nature of the classifiers used is assumed and promoted.

Initially, evidence about the presence of a face in the image is obtained and the face

hypothesis is launched for areas of high evidence. A first classifier module confirms/rejects the initial hypothesis in the most salient area. If it is not confirmed, the initial hypothesis is immediately rejected and the classifier chain is broken, directing the system toward other areas in the current image or to the detection of new evidences. On the other side, if the hypothesis is confirmed, the following module in the cascade is launched in the same way. This process, for an initial hypothesis consecutively confirmed by all modules, is finished when the last module confirms also the face hypothesis. In this case, the combined classifier output is a positive face detection.

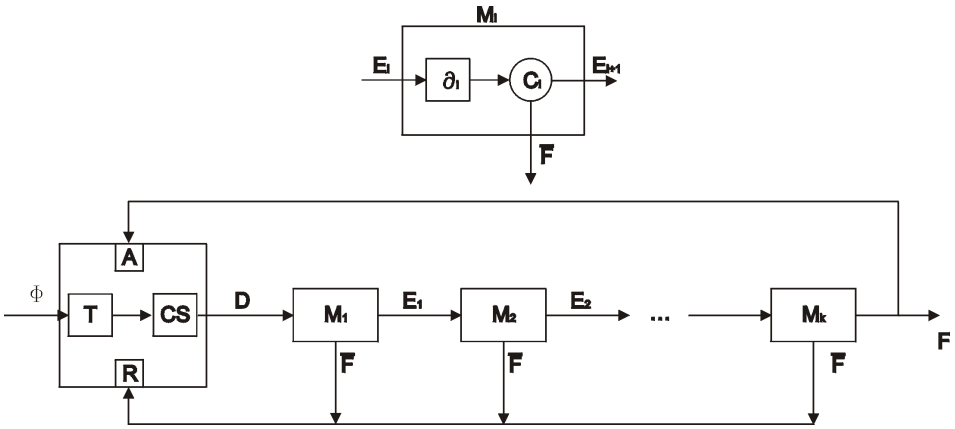


Figure 5.13: T means tracking and CS Candidate Selection, D are data, M_i is the i-th module, C_i the i-th classifier, E_i the i-th evidence, A accept, R Reject, $F/\bar{F}$ face/nonface, ∂_i the i-th evidence computation and Φ the video stream. (Courtesy of M. Castrillon)

The number of cascade stages and the complexity of each stage must be sufficient to achieve similar detection performance while minimizing computation. So, given a false positive rate, f_i, and the detection rate, d_i, for classifier module *i*, the false positive rate, FP, and the detection rate D of the cascade are respectively:

$$FP = \prod_{i=1}^{K} f_i \qquad D = \prod_{i=1}^{K} d_i \tag{5.15}$$

where K is the number of classifiers. These expressions show that cascade combination is capable of obtaining good classification rates and very low false positive rates if the detection rate of individual classifiers is good, close to 1, but the false positive rate of them is not so good, not close to 0. For example, for $K = 10$ and a false positive rate of individual classifiers of 0.3, the resulting false positive rate is reduced to 6×10^{-6}.

This classifier combination scheme can be considered as a kind of pattern rejection or rejecter, in the sense given by [Baker and Nayar, 1996], and can be interpreted in an analogy

with fluid filtering in a filtering cascade. In this case, each filtering stage rejects a fraction of impurity. The more stages with a rejection rate, the more pure fluid is obtained at the output.

How to select the individual classifier modules? Different options are possible. In ENCARA, an opportunistic criterion is employed to extract cues and to use, in a convenient fashion, explicit and implicit knowledge to restrict the solutions to a solution space fraction which can comply with real-time restrictions and have a flexible framework to test different solutions, adding modules or deleting others, allowing each module in the cascade to be also a combined classifier.

ENCARA is briefly described in terms of the following main modules, organized as a cascade of hypothesis confirmations/rejections:

M0.- Tracking: If there is a recent detection, the next frame is analysed first searching for facial elements detected in the previous frame: eyes and mouth corners. If the tracked positions are similar to the one in the previous frame and the appearance test is passed, ENCARA considers that a face has been detected.

M1.- Face Candidate Selection: The current implementation makes use of a skin colour approach to select rectangular areas in the image which could contain a face. Once the normalized red and green image has been calculated, a simple method based on defining a rectangular discrimination area on that colour space is employed for skin colour classification. Dilation is applied to the resulting blob image using a 3×3 structuring element.

M2.- Facial Features Detection: Frontal faces would verify some restrictions for several salient facial features. In the candidates areas selected by the *M1* module, the system removes heuristically elements that are not part of the face, i.e. neck, and fits an ellipse to obtain the vertical position of the blob. Later, this module searches for a first frontal detection based on facial features and its restrictions: geometric interrelations and appearance. This approach would first search potential eyes in selected areas taking into consideration that for Caucasian faces, the eyes are dark areas on the face. After the first detection of an individual, the detection process will be adapted to the individual's dimensions and appearance as a consequence of temporal coherence enforcement.

M3.- Normalization: In any case, the development of a general system capable of detecting faces at different scales must include a size normalization process in order to allow for a posterior face analysis and recognition reducing the problem dimensionality.

M4.- Pattern Matching Confirmation: A final confirmation step of the resulting normalized image is necessary to reduce the number of false positives. This step is based on an implicit knowledge technique. For eye appearance, a certain area (11×11) around both eyes is projected to a Principal Component Analysis (PCA) eigenspace and reconstructed. The reconstruction error provides a measure of its eye appearance [Hjelmas and Farup, 2001], and could be used to identify incorrect eye detections. If this test is passed, a final appearance test applied to the whole normalized image in order to reduce false positives makes use of a PCA representation that is classified using Support Vector Machines [Burges, 1998]. If the tests are passed, the mouth and nose are located in relation to eye pair position and their dark appearance in a face. In any other case, when no frontal face is detected, the system computes if there was a recent face detection in which at least one facial feature was not lost according to tracking process, and the possible face location is estimated with high likelihood.

5.4.2 Performance and Implementation

The main features of ENCARA are:

- The resulting system integrates and coordinates different techniques, heuristics and common sense ideas adapted from the literature, or conceived during its development.
- The system is based on a hypothesis verification/rejection scheme applied opportunistically in cascade, making use of spatial and temporal coherence.
- The system uses implicit and explicit knowledge.
- The system was designed in a modular fashion to be updated, modified and improved according to ideas and/or techniques that could be integrated.

ENCARA detects an average of 84% of the faces detected using the well-known Rowley-Kanade's detector [Rowley *et al.*, 1998], but 22 times faster using standard acquisition and processing hardware. ENCARA provides also the added value of detecting facial features for each detected face. More details of experiments carried out with ENCARA can be found in [Castrillón, 2003].

The ENCARA system was fully integrated in earlier versions of CASIMIRO. Currently, only the first filter of skin blob detection and a simple blob ratio filter are being used from ENCARA. The rest of the filters were too restrictive for our environment, with a significant amount of frontal faces being discarded by the system. Note that this is in line with

the approach described in Chapter 3: we are assuming here that skin-colour blobs with a certain width/height ratio are always faces. This constitutes a slight limitation of the domain in which we ourselves are able to detect faces, although that allows to get a good detection rate in the robot niche.

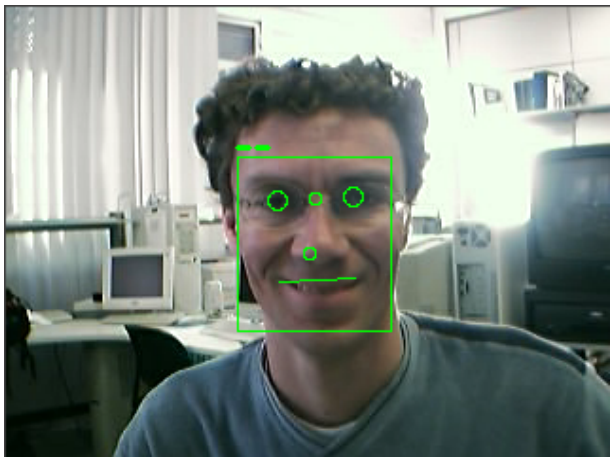


Figure 5.14: Example of face detected by ENCARA.

5.4.3 Use of Stereo Information and Shirt Elimination

ENCARA uses colour as a primary source for detection. Colour detection affects system performance as many faces labelled by humans as frontal are not processed by the system due to the fact that the skin colour blob does not seem to be correct. The great variety of colour spaces for detection seems to prove the impossibility of achieving a general method just using a simple approach. Detecting faces with skin colour is prone to false positive errors. Many objects in typical indoor environments tend to be categorized as skin, specially wooden furniture. Figure 5.15 shows how this problem affects detection.

In order to alleviate this problem, stereo information was used to discard objects that are far from the robot, i.e. in the background. A depth map is computed from the pair of images taken by the stereo camera. The depth map is efficiently computed with an optimized algorithm and library. The map is thresholded and an AND operation is performed between this map and the image that ENCARA uses. Fusion of colour and depth was also used in [Darrell *et al.*, 1998, Moreno *et al.*, 2001, Grange *et al.*, 2002]. The results are shown in Figure 5.16.

If the subject's shirt has a colour similar to skin (something relatively frequent) the system will detect the face rectangle with incorrect dimensions. This can cause the question

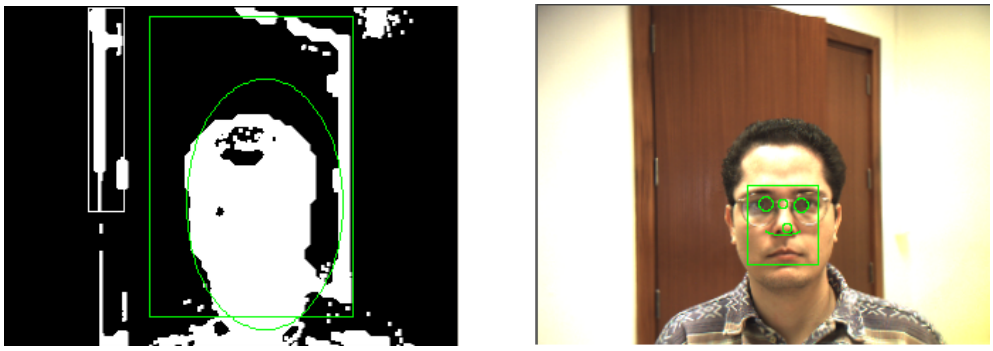


Figure 5.15: Skin colour detection. Note that wooden furniture is a distractor for facial detection.

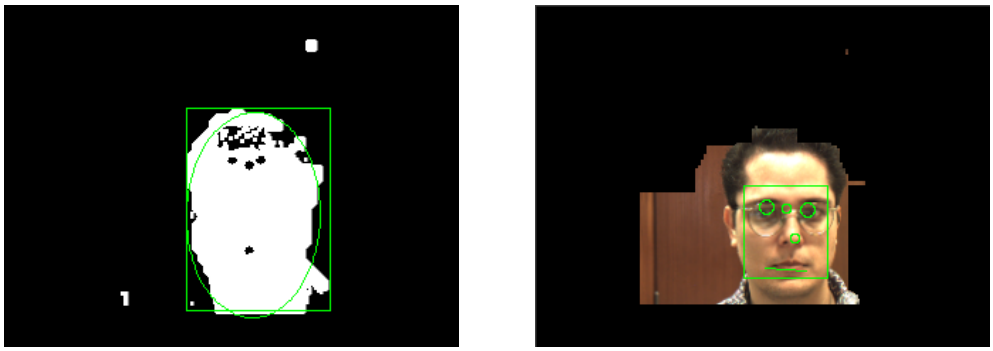


Figure 5.16: Skin colour detection using depth information.

system (Section 5.5) to work improperly. In order to alleviate this problem, the final face rectangle width is taken as the measured blob width at 1/4 of the blob height, see Figure 5.17. The final face rectangle height is (rectangle_height+final_rectangle_width/2).

5.5 Head Nod and Shake Detection

Due to the fact that (hands-free) speech feedback is very difficult to obtain for a robot, we decided to turn our attention to simpler input techniques such as head gestures. Kjeldsen contends that practical head gesture interactions fall into four categories [Kjeldsen, 2001]:

- Pointing: identifying an arbitrary location on the display.
- Continuous control: establishing the value of a continuous parameter.
- Spatial selection: identifying one of several alternatives distributed in either screen or image space.

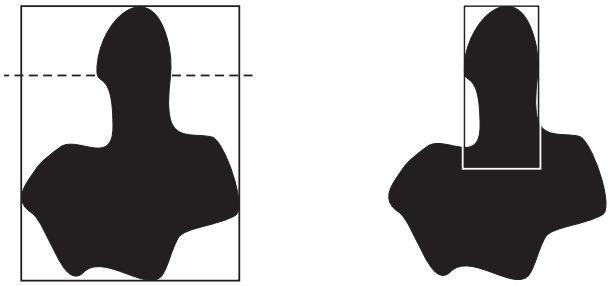


Figure 5.17: Face rectangles obtained without (left) and with (right) shirt elimination.

- Symbolic selection: identifying one of several alternatives by non-spatial means.

Head nods and shakes fall into the fourth category. They are very simple in the sense that they only provide yes/no, understanding/disbelief, approval/disapproval meanings. However, their importance must not be underestimated:

- The meaning of head nods and shakes is almost universal [2].
- They can be detected in a relatively simple and robust way.
- They can be used as the minimum feedback for learning new capabilities.
- YES and NO have been shown to be by far the two most common human inputs in conversations between humans and chatbots [Wallace, 2005].

The head nod/shake detection system described in [Tang and Nakatsu, 2000] achieved a recognition rate of around 90%. It used a Kanade, Luca and Tomasi tracker to follow facial features. Then, a vector is formed with the evolution of those features. The vector feeds a (previously trained) neural network that produces a decision.

The system for nod/shake detection described in [Kapoor and Picard, 2001] achieves a recognition accuracy of 78.46%, in real-time. However, the system uses complex hardware and software. An infrared sensitive camera synchronized with infrared LEDs is used to track pupils, and a HMM based pattern analyzer is used to the detect nods and shakes. The system had problems with people wearing glasses, and could have problems with earrings too. The same pupil-detection technique was used in [Davis and Vaks, 2001]. That work emphasized the importance of the timing and periodicity of head nods and shakes. However, in our view that information is not robust enough to be used. In natural human-human interaction, head

[2] In Bulgaria, a head nod signals a NO, while a head shake signals a YES.

nods and shakes are sometimes very subtle. We have no problem in recognizing them once the question has been made, and only YES/NO answers are possible. In many cases, there is no periodicity at all, only a slight head motion.

For our purposes, the nod/shake detector should be as fast as possible. We assume that the nod/shake input will be used only after the robot has asked something. Thus, the detector can produce nod/shake detections at other times, as long as it outputs right decisions when they are needed. We researched the problem from simple to more complex techniques, as we believe the solution should be as simple as possible.

The first proposal tested is shown in Figure 5.18. The algorithm should start (with v=h=0) immediately after the question has been asked. The larger the threshold t the better the response, although the delay before producing a response is longer. The algorithm was implemented using the centre of the estimated rectangle that contains the face.

```
repeat
   Compute absolute feature displacement in the horizontal and add it to h
   Compute absolute feature displacement in the vertical and add it to v
   if an output has not been given yet then
      if either v or h, or both, are above a threshold t then
         if if v > h then
            output=head nod
         else
            output=head shake
         end if
      end if
   end if
until an output is available
```

Figure 5.18: Simple head nod/shake detector.

The algorithm was tested by adding a button labelled "Ask" to a camera application. The algorithm starts when the user presses the button. Tests for yes/no responses were made for a given still position of the robot head. Taking into account only the response time we chose a value of t=3. For that value, 61 out of 78 tests (questions with alternate yes/no responses) were correctly recognized (78.2%).

The major problem of observing the evolution of simple characteristics like intereye position or the rectangle that fits the skin-colour blob is noise. Due to the unavoidable noise, a horizontal motion (the NO) does not produce a pure horizontal displacement of the observed characteristic, because it is not being tracked. Even if it was tracked, it could drift due to

lighting changes or other reasons. In practice, a horizontal motion produces a certain vertical displacement in the observed characteristic. This, given the fact that decision thresholds are set very low, can lead the system to error. The performance can be even worse if there is egomotion, like in our case.

The second proposal uses the pyramidal Lucas-Kanade tracking algorithm described in [Bouguet, 1999]. In this case, there is tracking, and not of just one, but multiple characteristics, which increases the robustness of the system. The tracker looks first for a number of good points to track, automatically. Those points are accentuated corners. From those points chosen by the tracker we can attend to those falling inside the rectangle that fits the skin-colour blob, observing their evolution. Note that even with the LK tracker there is noise in many of the tracking points. Even in an apparently static scene there is small motion in them. The procedure is shown in Figure 5.19.

```
repeat
   Compute the absolute displacement of each tracking point
   Let (Mv,Mh) be the mean absolute displacement of the points inside the skin-colour
   rectangle
   if an output has not been given yet then
      if Mv>thresholdv OR Mh>thresholdh then
         if if Mv > Mh then
            output=head nod
         else
            output=head shake
         end if
      end if
   end if
until an output is available
```

Figure 5.19: LK tracking-based head nod/shake detector.

The method is shown working in Figure 5.20. The LK tracker allows to indirectly control the number of tracking points. The larger the number of tracking points, the more robust (and slow) the system. The method was tested giving a recognition rate of 100% (73 out of 73, questions with alternate YES/NO responses, using the first response given by the system). It was implemented in CASIMIRO, using a resolution of 320x240 pixels. The horizontal and vertical thresholds should be different to compensate for the image aspect ratio and also for the distance to the individual.

What happens if there are small camera displacements (the case of CASIMIRO, for the camera used for head nod/shake detection is mounted on the head)? In order to see the

effect of this, linear camera displacements were simulated in the tests. In our particular case, only linear displacements are possible, for neck pan and tilt can not work simultaneously. In each frame, an error is added to the position of all the tracking points. If (D_x, D_y) is the average displacement of the points inside the skin-colour rectangle, then the new displacement is $D_x + e_x$ and $D_y + e_y$. The error, which is random and different for each frame, is bounded by $-e_{max} < e_x < e_{max}$ and $-e_{max} < e_y < e_{max}$. Note that it is no longer possible to use a fixed threshold, like in Figure 5.19, because the error is unknown.

Figure 5.20: Head nod/shake detector.

The error also affects to the tracking points that fall outside the rectangle. Assuming that the objects that fall outside the rectangle are static we can eliminate the error and keep on using a fixed threshold:

$$\begin{aligned}(D_x + e_x) - (F_x + e_x) &\approx D_x \\ (D_y + e_y) - (F_y + e_y) &\approx D_y\end{aligned} \tag{5.16}$$

For the system to work well it is needed that the face occupies a large part of the image. A zoom lens should be used. When a simulated error of $e_{max} = 10$ pixels was introduced, the recognition rate was 95.9% (70 out of 73). In this case there is a slight error due to the fact that the components F_x and F_y are not exactly zero even if the scene outside the rectangle is static.

A confidence measure for the system output can be obtained from the difference between Dx and Dy:

$$\frac{||Dx| - |Dy||}{||Dx| + |Dy||} \tag{5.17}$$

Another type of error that can appear when the camera is mounted on a mobile device is the horizontal axis inclination. Inclinations can be a problem for deciding between a YES and a NO. In order to test this effect, an inclination error was simulated in the tests (with the correction of (5.16) active). The error is a rotation of the displacement vectors **D** a certain angle α clockwise (the origin of the image plane is the top left corner):

$$\mathbf{D}' = \mathbf{D} \cdot \begin{bmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{bmatrix} \tag{5.18}$$

Recognition rates were measured for different values of α, producing useful rates for small inclinations, see Table 5.4.

α = 20º	90% (60 out of 66)
α = 40º	83.8% (57 out of 68)
α = 50º	9.5% (6 out of 63)

Table 5.4: Recognition results for different values of α.

The implemented module is continuously waiting for a signal indicating that a question has just been made. When the signal arrives, it starts the algorithm and, when the first output is obtained it is returned to the signalling module (the Talk module). If a certain time passes and no output is obtained an informative code of the situation is returned. Then the Talk module increases the robot's arousal and repeats the question.

Histogram equalization was applied to the face rectangle zone in order to alleviate the effect of illumination variations. After a question is answered CASIMIRO pronounces one or more "accepting" words like "ok" or "I see". We found that these words constitute a very positive feedback for the observer, who immediately recognizes that the robot has understood him/her.

5.6 Memory and Forgetting

Memory is a crucial component of intelligent beings. The best way to know the importance of memory is to look at cases where it is not present. Sacks describes in moving accounts the

effects of amnesia (also called the Korsakov Syndrome) in some of his patients [Sacks, 1995, Sacks, 1985]. When amnesia is severe, the patient is almost dead for daily life.

In [Schulte *et al.*, 1999] three characteristics are suggested as critical to the success of robots that must exhibit spontaneous interaction in public settings. One of them is the fact that the robot should have the capability to adapt its human interaction parameters based on the outcome of past interactions so that it can continue to demonstrate open-ended behaviour. CASIMIRO is intended to interact with people. Humans will be the most important "object" in its environment. Data associated to humans (gathered throughout the interaction) should be stored in memory, so that the robot could take advantage of previous experiences when interacting with them. Breazeal [Breazeal, 2002] argues that to establish and maintain relationships with people, a sociable robot must be able to identify the people it already knows as well as add new people to its growing set of known acquaintances. In turn, this capacity will be part of the robot's autobiographical memory.

In this work we consider memory important because it is a means of achieving an unpredictable and complex observed behaviour. Therefore memory will be used here merely from a functional point of view. Notwithstanding, person recognition (which needs memory) is in fact an ability that Gardner calls naturalist intelligence (see Chapter 1). On the other hand, most of us often wonder: given the same sensorial inputs, will I act always in the same manner? Obviously not, if we have memory. And without memory? In humans amnesia directly leads to repetitive behaviour.

In order to make this person memory possible, gathered data should be unambiguously associated to the correct person. Facial recognition would be the perfect approach. However, the experience of the author with face recognition (see the List of Publications) is somewhat negative: face recognition still does not work well in unrestricted scenarios. Recognition rates fall as more time passes since the training samples were taken (actually, performance decreases approximately linearly with elapsed time, [Phillips, 2002]). Illumination, pose and expression variations normally reduce recognition rates dramatically, see Section 1.2.

Colour histograms of (part of) the person's body could also be used as a recognition technique, see Figure 5.21. Colour histograms are simple to calculate and manage and they are relatively robust. The price to pay is the limitation that data in memory will make sense for only one day (at the most). Colour histograms of a person's body were used for short-term identification people in [Maxwell, 2003, Kahn, 1996, Maxwell *et al.*, 1999] and also for people tracking [Krumm *et al.*, 2000, Collins and Dennis, 2000].

CASIMIRO achieves person identity maintenance by using colour histograms in con-

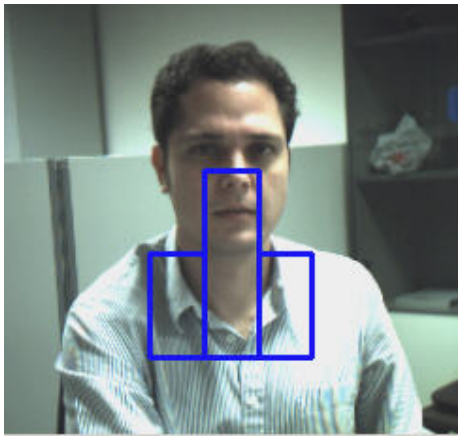


Figure 5.21: Region that could be used for person identification.

junction with a simple person tracking algorithm. Tracking is done in 1D, for the interesting position is the angle of the person with respect to the robot.

The implemented tracking algorithm is very simple. Each person is represented as a single point in two sets of horizontal positions (positions range from 0 to 180º) at times $t-1$ and t. The association of points between the two sets is obtained as that which minimizes the total sum of distances between points of the two sets. This minimization involves a factorial search, though it is practical for the number of people that will be expected to interact with the robot. Ties can appear, for example in the case of crossings, see the example of Figure 5.22. This ties are broken by selecting the association with lowest variance of distances, 1 with A and 2 with B in the case of the example. This always selects non-crossings.

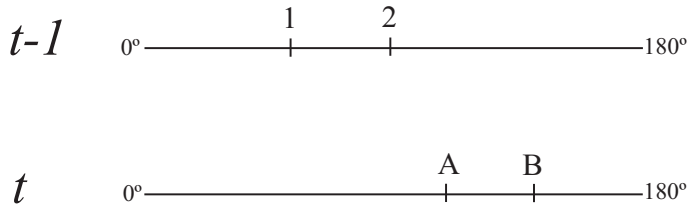


Figure 5.22: Tie in sum of distances. The sum of distances $|1-A|+|2-B|$ is equal to $|1-B|+|2-A|$. Without further information, we can not know if the two individuals have crossed or not.

Crossings are detected by considering that, in a crossing, there is always a fusion and a separation of person blobs. Person blobs are detected by the omnidirectional vision system (Section 5.1). Fusions and separations are detected as follows:

- A blob fusion is detected when the number of blobs in the whole omnidirectional

image decreases by one at the same time that one of the blobs increases its area significantly.

- A blob separation is detected when the number of blobs in the image increases by one at the same time that a fusioned blob decreases its area significantly.

The only way to know if a there is a crossing is by maintaining some sort of description of the blobs before and after the fusion. Histograms of U and V colour components are maintained for each blob. The Y component accounts for luminance and therefore it was not used. Whenever a separation is detected, the histograms of the left and right separated blobs are compared with those of the left and right blobs that were fusioned previously. Intersection [Swain and Ballard, 1991] was used to compare histograms (which must be normalized for blob size). This procedure allows to detect if there is a crossing, see Figure 5.23. The histogram similarities calculated are shown in Figure 5.24. A crossing is detected if and only if $(b + c) > (a + d)$. Note that in the comparison no threshold is needed, making crossing detection relatively robust.

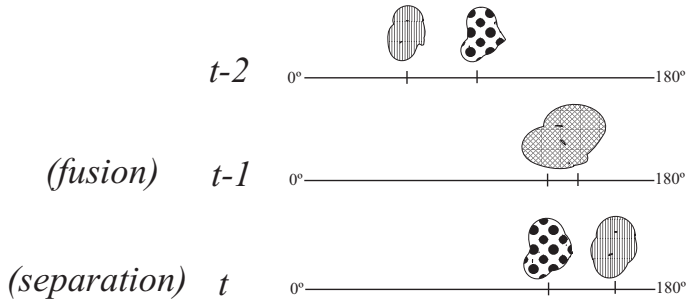


Figure 5.23: Crossings can be detected by comparing blob histograms at fusion and separation events.

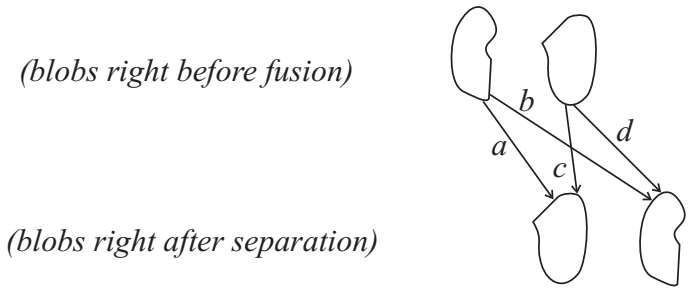


Figure 5.24: Blob similarities calculated.

In order to achieve person identification, a set of Y-U histograms are stored for each person detected. The zone from which these histograms are calculated is a rectangle in the

lower part of the image taken from the stereo camera (see Section 5.4 above). The rectangle is horizontally aligned with the centre of the face rectangle detected, and extends to the lower limit of the image (chest and abdomen of standing people will always occupy that lower part of the image). The upper edge of the rectangle is always under the lower edge of the face rectangle detected. The width of the rectangle is proportional to the width of the face rectangle detected.

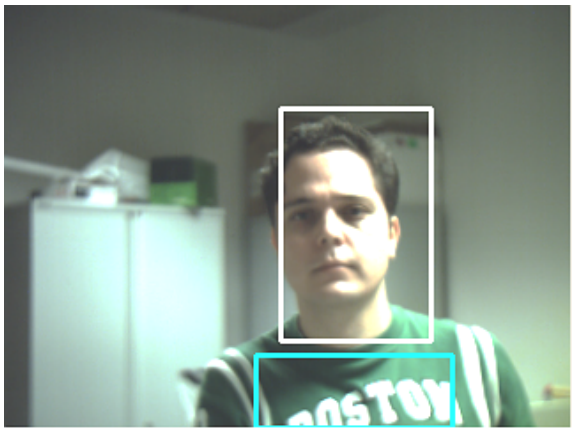


Figure 5.25: Region used for person identification.

When the robot fixates on a person that the tracking system has labelled as new (the tracking system detects a new person in the scene when the number of foreground blobs increases and no blob separation is detected), it compares the histograms of the fixated individual with those of previously met individuals. The Jeffrey divergence, which has been shown to give better results than intersection (although it is more computationally complex, see [Puzicha *et al.*, 1999]), was used for comparison. This search either gives the identity of a previously seen individual or states that a new individual is in the scene. In any case the set of stored histograms for the individual is created/updated.

Each person that the robot has focused on has data associated (the things that the robot has spoken to that person, the reactions of the person, etc.). Tracking allows to maintain this information. Whenever the robot focuses its attention on a person, the associated data (if any) is immediately available to other modules, mainly the Behaviour Module (Chapter 7).

CASIMIRO has a memory associated to individuals, but also a "global" memory. That memory retains information that can not be considered associated to any particular individual, like for example the fact of whether the robot has already said that it is a hot day.

Memory is of utmost importance for avoiding predictable behaviours. However, memorizing facts indefinitely leads to predictable behaviours too. Humans are not machines

with an incredible number of terabytes in which to store facts and events indefinitely. Behavioural changes occur when we memorize but also when we forget. Thus, a forgetting mechanism can also be helpful in our effort, especially if we take into account the fact that actions chosen by the action-selection module do not always produce the same visible outcome (i.e. the Talk actions).

Suppose that a behaviour is triggered by a certain state. As long as that state is present, the behaviour will execute the associated action over and over again, with a frequency imposed by the cycle of the action-selection implementation. With memory, the forgetting mechanism would have the control of the repetitions. Basically, actions should repeat only when:

- The robot forgets that it has executed them
- The robot do not forgets but a certain time has passed (in which it is reasonable to try again)

The first controlled studies of forgetting mechanisms were carried out by Ebbinghaus [Ebbinghaus, 1913]. Those experiments, replicated many times, concluded that the forgetting process is more accelerated (we tend to forget more information) in the first minutes and hours after memorization. This can be characterized by a power function (of the form $y = a^t - b$, where a and b are positive real numbers), as demonstrated by Wixted and colleagues [Wixted and Ebbesen, 1991, Wixted and Ebbesen, 1997, Kahana and Adler, 2002]. In [Rubin and Wenzel, 1996] over a hundred forgetting functions were compared and it was found that the power function was one of only four that provided a good fit to a wide range of forgetting data.

In CASIMIRO, forgetting is modelled in the following way. Let $f(t)$ be a forget function, which we use as a measure of the probability of forgetting something:

$$f(t) = \begin{cases} max(0, 1 - t^{-k}) & for\ t > l \\ 0 & for\ t \leq l \end{cases} \tag{5.19}$$

where k and l are constants. We apply the f function to the set of Boolean predicates that the robot retains in memory (both global and associated to individuals). There is evidence that some facts are forgotten earlier than others. Some facts are never forgotten. Interference effects are thought to be one of the factors that account for those differences. For simplicity, we do not model those aspects and consider them represented in the stochastic nature of the

forgetting function. When a predicate is forgotten, it takes the value it had at the beginning, when the system was switched on (i.e. it is removed from the current memory state).

In the implementation of this mechanism, t is the elapsed time in seconds. Different values of k were assigned to all the symbols that can be stored in memory. Those values were assigned with the help of some k values for which the probability of forgetting exceeds a value at a given time, see Figure 5.26. Table 7.3 in Section 7.2.3 shows the k values assigned to the symbols that can be stored in memory. A few predicates were given a value larger than zero for the constant l, which allows to avoid the case of too-early forgetting.

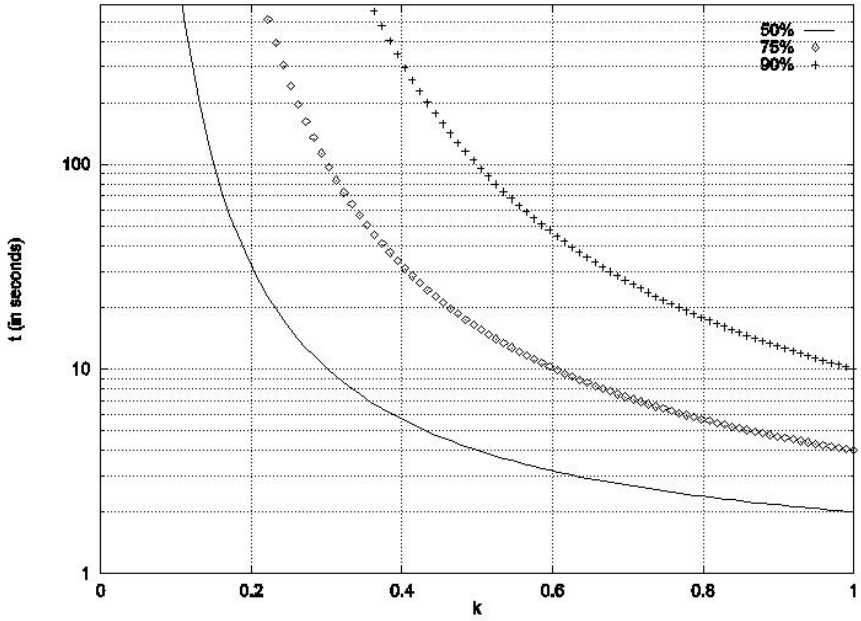


Figure 5.26: t values for given k values and forget probabilities.

5.7 Owner Identification

The recognition of the robot's owner or caregiver is a potentially important feature. The recognition may not affect the creator at all, but it could impress other people present in the room. Person recognition is difficult to achieve, the identification of the robot's owner may be more feasible, for the problem reduces to two classes (owner and not-owner). Still, using face or voice for owner recognition would lead to essentially the same lack of robustness of face and speech recognition for multiple individuals.

Who is the owner of the robot? The correct answers is: the person who buys it.

However, we are more interested in the role of the person who cares the robot and uses it more frequently (this person is generally the owner too). This person is the one who switches the robot on, which is usually done from a certain part of the robot or from a certain computer. That cue may be exploited to recognize the robot owner or caregiver.

Amazing Amanda [Playmates Toys Inc., 2005], a doll released in autumn of 2005, is able to recognize the girl that takes the mother role. Once the doll is activated, it starts asking questions. That way, the doll can "learn" the girl's voice patterns. From that moment on, the doll is able to recognize the utterances of its "mommy". Other voices can lead Amanda to say "You don't sound like Mommy".

Such technique may seem rather *ad hoc*. However, the approach finds striking examples in nature. Lorenz, one of the founders of ethology, found that, upon coming out of their eggs, geese follow and become attached to the first moving object that they encounter. He showed this by rearing the geese from hatching. From that moment on the geese would follow him. Such phenomenon, which also appears in mammals, is known as *imprinting* [Lorenz, 1981].

In the case of CASIMIRO, the main computer (from which the robot is switched on) is situated behind the robot, on the same table, see Figure 5.27. A camera was placed on top of that computer. The owner detection module uses that camera to search for a skin coloured blob in the image. When the robot is switched on this module will detect a skin coloured blob. The camera has a wide-angle lens, and a relatively low resolution of 160x120 is used.

When no blob is encountered in the image the module notifies the Attention module of that event. At that moment the owner detection module exits in order to free CPU resources. Once it has been notified by the owner detection module, the Attention module considers the owner as the first blob that "enters" the omnidirectional camera image from the left. The "owner" property is stored along with the individual in the tracking process.

This simple procedure is a form of imprinting. In a sense, the robot finds its owner-caregiver in the first human it sees. It does not stores any biometric features to recognize the owner after being switched on, only its position.

Face recognition researchers tend to measure performance in terms of the number of individuals that the system can recognize and *measured* error rate. A *measured* error rate of 5-10% can be considered very good under restricted conditions. The approach presented here recognizes a single individual with *guaranteed* zero error. No face recognition method would recognize the owner with such low error. Note that this is the result of following the approach introduced in Chapter 3: we have been able to devise the simplest algorithm (or

Figure 5.27: The computer from where CASIMIRO is started. The interaction space is on the left.

one of the simplest) that allows to recognize the owner. For that purpose we have fitted the solution to the robot niche.

5.8 Habituation

Habituation is a filtering mechanism that has received a lot of attention in physiology and psychology. In particular, some researchers have investigated the mechanisms of habituation in animals, being one of the most known works the study of the Aplysia's gill-withdrawal reflex [Castellucci *et al.*, 1970]. When the animal's siphon is touched, its gill contracts for a few seconds. If the siphon is stimulated repeatedly, the gill-withdrawal effect tends to disappear. Crook and Hayes [Crook and Hayes, 2001] comment on a study carried out on two monkeys by Xiang and Brown who identified neurons that exhibit a habituation mechanism since their activity decreases as the stimulus is shown repeatedly.

Stanley's model [Stanley, 1976] of habituation, proposed to simulate habituation data obtained from the cat spinal cord, has been widely used in the literature. This model describes the decrease efficacy y of a synapsis by the first-order differential equation:

$$\tau \frac{dy(t)}{dt} = \alpha(y_0 - y(t)) - S(t), \tag{5.20}$$

where y_0 is the normal, initial value of y, $S(t)$ represents the external stimulation, τ is a time constant that governs the rate of habituation and α regulates the rate of recovery. Equation (5.20) ensures that the synaptic efficacy decreases when the input signal $S(t)$ increases and returns to its maximum y_0 in the absence of an input signal.

The model given by (5.20) can only explain short-term habituation, so Wang introduced a model to incorporate both short-term and long-term habituation using an inverse S-shaped curve [Wang, 1995],

$$\tau \frac{dy(t)}{dt} = \alpha z(t)(y_0 - y(t)) - \beta y(t) S(t) \tag{5.21}$$

$$\frac{dz(t)}{dt} = \gamma z(t)(z(t) - l) S(t), \tag{5.22}$$

where α, y_0 and γ have the same meaning than in (5.20), β regulates the habituation and $z(t)$ decreases monotonically with each activation of the external stimulation $S(t)$, and models the long-term habituation. Due to this effect of $z(t)$ after a large number of activations, the recovery rate is slower.

Note that novelty detection is a concept related to habituation. Novelty detection is the discovery of stimuli not perceived before and so habituation serves as a novelty filter [Stiles and Ghosh, 1995]. From a engineering viewpoint, perceptual user interfaces, like human-like robots, should be endowed with a habituation mechanism. The interest is twofold. First, it would be a filtering mechanism, discarding (or minimizing the importance of) repetitive information while paying attention to new experiences. This is in part motivated by the desire to distinguish between artificial and human signals. Artificial signals are often static or repeat with a fixed frequency. We do not want our robot to pay much attention to the hands of a wall-mounted clock. Instead, it would be more interesting to detect non-repetitive stimuli, such as a conversation or a sudden loud noise. Note that we generally consider monotonous signals as those having a fixed frequency or frequencies (which can be zero, that is, the signal does not change) but signals whose frequency changes in a periodic pattern could also be considered monotonous. Higher scales are also possible but we do not consider them in this work because they are very hard to visualize and real examples of them are not so common.

Second, habituation would lead to a more human-like behaviour, as perceived by users of the interface. As an example of this, consider Kismet. Someone can catch the eye of the system while waving a hand in its visual field of view, but if the stimulus is repetitive for a long time the system can show a lack of interest in it. Many aspects of Kismet's mental

architecture are directly or indirectly influenced by the detection of monotonous sensory signals: stimulation and fatigue drives and the arousal dimension of its affect space (and in turn some emotional states, like surprise, boredom or interest).

Although we focus our work on the abilities described above, many other applications are also imaginable. In the robotics field, habituation mechanisms have been used to reduce oscillations caused by collision-avoidance behaviours when navigating through a narrow corridor [Chang, 2000]. Marsland [Marsland *et al.*, 2000] uses a SOM neural network as a memory for novelty detection. To add short-term habituation to the original network, each neuron of the SOM is connected to an output neuron with habituable synapses based on the model (5.20). Habituation is also used in [Stoytchev and Arkin, 2003] for controlling reactivity strength, visual attention [Peters and Sowmya, 1998, Breazeal and Scassellati, 1999], and general learning [Damper *et al.*, 1999]. On the other hand, there is considerable interest in the field of musicology in Beat Tracking Systems (BTS) [Goto and Muraoka, 1997]. BTS systems aim to find the tempo of an audio signal, which is basically the rate of repetitions. The main applications of BTS systems are audio/video editing, synchronization of computer graphics with music, stage lighting control and audio content searching.

If we use the model of Equation (5.20) we can obtain undesired effects with certain stimuli. A periodic input signal (with frequency greater than zero) can produce a response that does not exhibit habituation. This is due to the fact that the model does not account for changing stimuli, but for continuous ones. In order to include this fact in the model, we propose to use an auxiliary signal which will be zero when the stimulus is stationary or with a fixed frequency, and one otherwise, and use this signal as an input to the habituation model (5.20).

The auxiliary signal, which basically detects monotonous stimuli, is obtained from the spectrogram of the stimulus itself. The spectrogram is a time-frequency distribution of a signal, and it is based on the Fourier Transform with a sliding window [Holland *et al.*, 2000]. The equation

$$\Phi(t,f) = \left| \int_{-\inf}^{inf} x(\tau) e^{(t-\tau)^2/T^2} e^{-j2\pi f\tau} d\tau \right|^2 \tag{5.23}$$

gives the definition of a spectrogram with a Gaussian window function of half-width T, and it is the power spectrum of a signal which corresponds to the squared magnitude of the Fourier transform of the windowed signal. The window can have other forms apart from the Gaussian one. In Figure 5.28 we show an audio signal and its corresponding spectrogram, where brighter areas correspond to higher power. Two well defined frequency spectra can

be distinguished, for there is a change in the input signal at time 0.5 s. Temporal patterns of the stimulus signal have a specific pattern in the spectrogram. A fixed frequency signal corresponds to a straight line parallel to the time axis in the spectrogram, and the length of this line indicates how long has been the stimulus present.

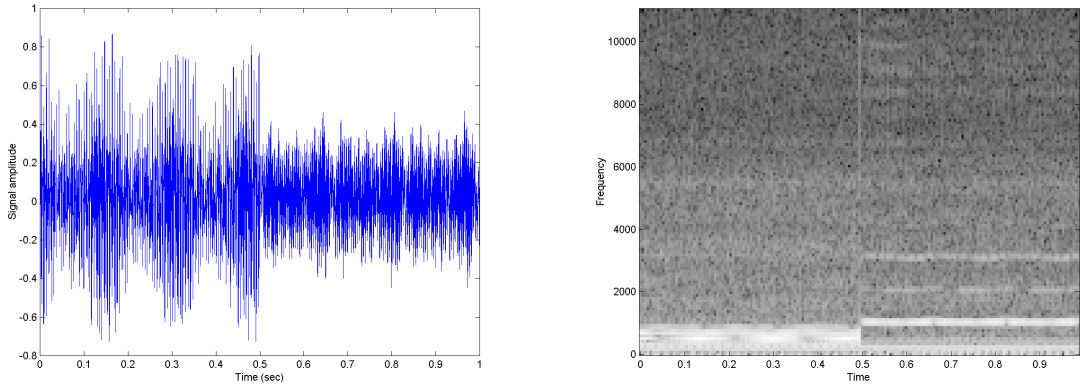


Figure 5.28: Audio signal (left) and its corresponding spectrogram (right).

Spectrograms are computed from windows of the input signal. These windows, of length l, overlap by $l - 1$ samples. Let each spectrogram be represented as a matrix M, in which rows represent frequencies and columns represent time. We calculate the variance of each row of M, which produces a column vector $\mathbf{v}$. The norm of this vector $\mathbf{v}$ is a measure of how monotonous the input signal is. The norm will be high when the signal is changing, and low otherwise. Thus, the auxiliary signal needed is simply the thresholded norm of $\mathbf{v}$. The amplitude of the input signal affects the power content of the spectrograms, and in turn the norm of $\mathbf{v}$. Thus, prior to calculating the FFT the input signal must be normalized dividing each input window by the sum of its absolute values. A value of 1 for the auxiliary signal will mean that there are changes in the input signal, while a value of 0 indicates that the input signal is monotonous. Once the auxiliary signal is available, the model (5.20) is used to get the desired habituation behaviour, as controlled by parameters τ and α.

Formally, let N and l be the number of rows and columns of M, respectively, and let $m_{i,j}$ represent the element in row i and column j of M. Vector $\mathbf{v}$ is calculated as:

$$v_i = \frac{\sum_{j=1}^{l}(m_{i,j} - \mu_i)^2}{l} \quad ; \quad i = 1, \ldots, N \tag{5.24}$$

where:

$$\mu_i = \frac{\sum_{j=1}^{l} m_{i,j}}{l} \quad ; \quad i = 1, \ldots, N \tag{5.25}$$

The auxiliary signal is then, for a given threshold T:

$$A = \begin{cases} 1 & \text{if} \quad |\mathbf{v}| > T \\ 0 & \text{if} \quad |\mathbf{v}| \leq T \end{cases} \tag{5.26}$$

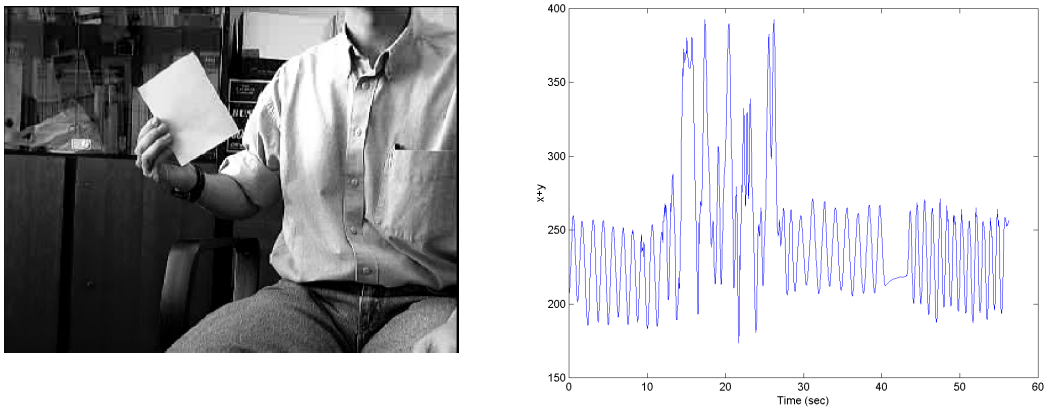
With this method both static an fixed frequency stimuli can be detected. However, there are stimuli that change their frequency according to a periodic pattern. These stimuli should also be considered as monotonous. The hissing sound of a siren, for example, is a signal whose frequency changes in a repeated pattern. After few repetitions the signal will be considered monotonous. One way to detect these kind of stimuli is to use the same method with the auxiliary signal. If the input signal changes its frequency content in a repeated pattern, the auxiliary signal will be periodic with a fixed frequency, and that can be detected as explained in the previous paragraph. Thus, two thresholds will be needed, one for the "first level" and one for the "second level". Higher levels could conceivably be used, but we have not considered them because they are very difficult to visualize and encounter in the physical world. Note that the second-level auxiliary signal will be 1 when there are changes in the first-level auxiliary signal, and thus when there are changes in the input signal, and 0 otherwise. Thus, the final input to the habituation model (5.20) will be the second-level auxiliary signal. Note that this second level introduces additional computation, and in some cases we could consider it unnecessary, if we decide to detect only simple monotonous signals.

There is only one detail left. If the first-level auxiliary signal is 1 (meaning that the input signal is changing), and this remains for a while, the second-level auxiliary signal will be 0 (because the second-level norm of the variance vector will be 0) which is not the correct value. In order to correct this, the second level must detect when the norm is 0 and, if so, use the value of the first-level auxiliary signal, instead of the second-level auxiliary signal. Note that if the first-level auxiliary signal is periodic the second-level variances obtained should theoretically be 0, which would prevent the use of this correction. However, in all the experiments carried out this never happened, because there is always an unavoidable amount of fluctuations in the input signal, which makes the variances larger than 0.

A previous version of the method proposed here has been already published elsewhere [Lorenzo and Hernández, 2002*b*, Lorenzo and Hernández, 2002*a*]. That version used only the frequency associated to the maximum power. Habituation should be present when the plot of that frequency versus time is a straight line. Changes are detected by fitting a line to the last k values of the frequency and computing the difference between the current value and the predicted value with the fitted line. That approach, however is too simplistic

in the sense that it assumes that the input signal is entirely represented by the frequency of maximum power.

The algorithm described above was implemented to test it with different input signals. The first experiments that we present use only the first level mentioned above. In order to gather signals from the visual domain, we recorded video containing a yellow bright stimulus (a yellow card) that was moved in a repetitive fashion, see Figure 5.29-a). Using simple segmentation techniques we extracted the centroid of the card on each frame (384x288) and summed the x and y pixel coordinates to form the one-dimensional signal of Figure 5.29-b). The sequence of card movements throughout the recording was: horizontal movement, random (aperiodic) movement, vertical movement and vertical movement at a different frequency than the previous one.



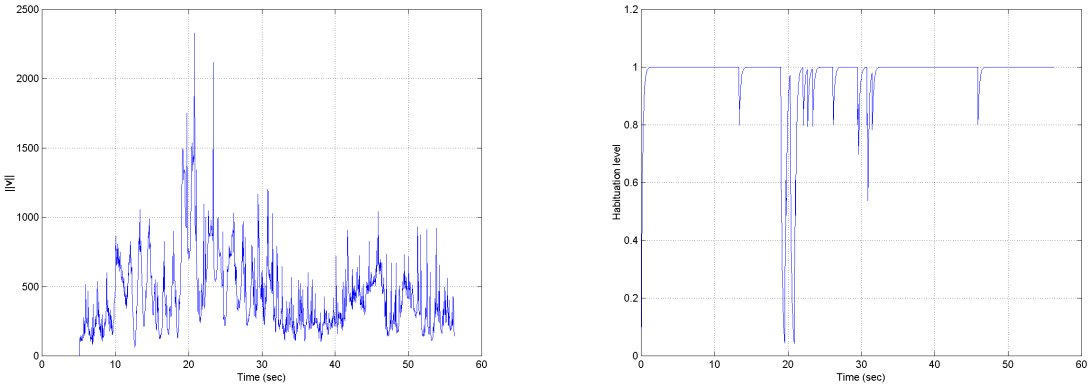
a) b)

Figure 5.29: a) Video recording used for the visual habituation experiment, b) one-dimensional signal extracted from it.

The results appear in Figure 5.30. Windows of 128 samples were used, and the variance threshold was set at 1000.

As for the audio domain, we recorded signals with a standard PC microphone, at a 22050 Hz sample rate, 8 bits. Figure 5.31 shows the results obtained for an audio signal that contains three sequential parts: silence (0-0.5s), people speaking (0.5-1s) and a tone played from an electric piano (1-1.4s). Note that there is an initial delay due to the need to fill the input window, here of length $l = 5120$. The habituation level, obtained using the model of (5.20), shows a satisfactory response.

Figure 5.32 shows the results obtained for an audio signal that contains another three sequential parts: a tone played from an electric piano (0-0.5s), silence (0.5-1s) and another tone (1-1.4s). The same window length $l = 5120$ was used, and again the habituation level



a) b)

Figure 5.30: a) Evolution of the (l_2) norm of the variance vector $\mathbf{v}$, b) habituation level, using $\tau = 5, \alpha = 1$.

shows a satisfactory behaviour.

In order to test both the first and second levels of the method, we built an audio signal containing three sequential parts: a beep repetitive sound from a mobile phone, people speaking and a tone played from an electric piano. This signal was accelerated to reduce computation time, which does not alter the qualitative results of the experiments. Results are shown in Figure 5.33. The window length was $l = 5120$ for the first level and $l = 2148$ for the second. In this case the repetitive beeps (clearly observed as a repetitive pattern in the first part of the spectrogram) are correctly considered as monotonous. This would not have occurred if we had used the first-level auxiliary signal alone, for numerous changes are detected (see Figure 5.33-d).

Next, we discuss a few aspects of practical interest. Particularly, we will comment on the effect of the values of the different parameters to use:

- Length of the input window, l: It should be the largest possible, in order to detect stimuli with large period. However it cannot be too large because that would introduce an unacceptable delay in the response to stimuli with smaller period. Thus, it depends on the type of stimuli. A flexible solution would be to implement multiple instances of the problem, each one with a different size for this parameter, in a multiscale fashion.
- Tau, τ: It controls the rate of habituation.
- Alpha, α: It controls the rate or recovery.
- Number of discrete frequency levels, N: Dependent on the type of input stimulus, it should normally be the largest possible. For the case of auditive signals, the minimum

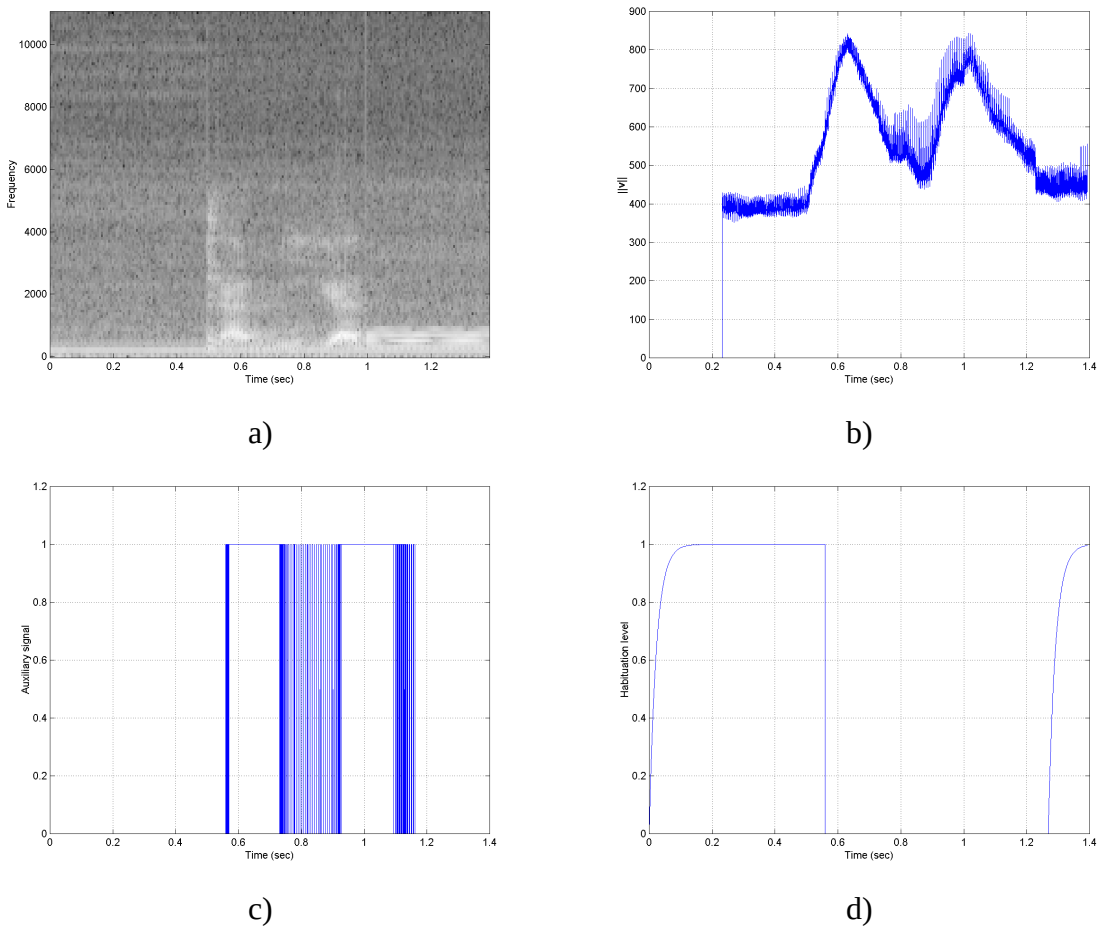


Figure 5.31: a) Spectrogram of the audio signal, b) evolution of the (l_2) norm of the variance vector $\mathbf{v}$, c) auxiliary signal, obtained using a threshold of 600, d) habituation level, using $\tau = 1, \alpha = 0.002$.

noticeable difference that people can distinguish is as low as 1.3Hz. Other input stimuli could be sonar data, blob positions, pressure readings, etc.

- Variance thresholds: They refer to the minimum change in the frequency spectrum to detect a change in the signal. If set too high, we run the risk of ignoring signal changes. If set too low, "hypersensitive" responses could be obtained. The appropriate values depend both on the type of input signal and the number of discrete frequency levels. These thresholds could be changed depending on the amount of available resources. If available resources are high, a lower threshold could be appropriate (producing more sensitivity or attention). Otherwise, a higher threshold would produce a more believable response.

The computational cost of the proposed method is basically dependent on the calculus

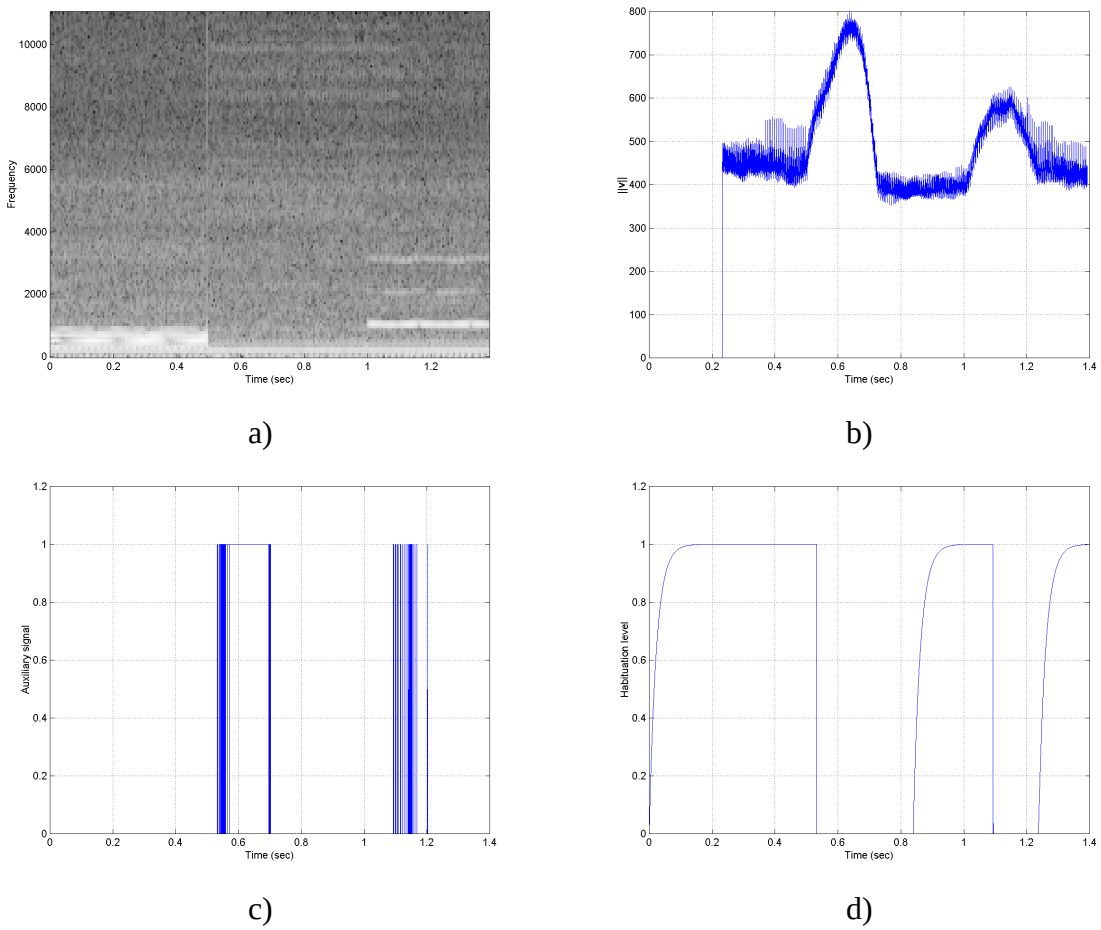


Figure 5.32: a) Spectrogram of the audio signal, b) evolution of the (l_2) norm of the variance vector $\mathbf{v}$, c) auxiliary signal, obtained using a threshold of 600, d) habituation level, using $\tau = 1, \alpha = 0.002$.

of the spectrogram. This, in turn, basically depends on the FFT. Thus, the total cost, for a window of length l is $l \log_2 l$ (for the first-level alone). This is therefore the cost of producing a new value of the auxiliary signal for each input sample. If a multiscale (multiple values for l) approach is used, the multiple instances of the problem can use parallel computation. Also, the second-level part of the problem can be solved in parallel with the first-level.

The habituation mechanism described here was implemented in CASIMIRO, for signals in the visual domain only, i.e. images taken by the stereo camera (see Section 4.1). The difference between the current and previous frame is calculated. Then it is thresholded and filtered with Open and Close operators. Also, blobs smaller than a threshold are removed. Then the centre of mass of the resultant image is calculated. The signal that feeds the habituation algorithm is the sum of the x and y components of the centre of mass. This way when the image does not show significant changes or repetitive movements are present for a while

the habituation signal grows. When it grows larger than a threshold, an inhibition signal is sent to the Attention module, which then changes its focus of attention. Neck movements produce changes in the images, though it was observed that they are not periodic, and so habituation does not grow.

For signals in the audio domain the algorithm performs too slow to work with the raw signal. It can be applied to the audio domain if higher level signals are formed from the raw input. Segments of the raw signal could be categorized first, in a fast way, producing a signal of longer period.

Summarizing, in this section a simple spectrogram-based algorithm has been described for detecting monotonous input signals, independent of their sensory origin (auditive, visual, ...). Signals that repeat with constant frequency or frequencies are considered monotonous. Signals that present a periodic changing pattern in their frequency content can also be considered monotonous. The usefulness of the algorithm was evaluated in experiments with signals gathered both from the visual and the auditive domains.

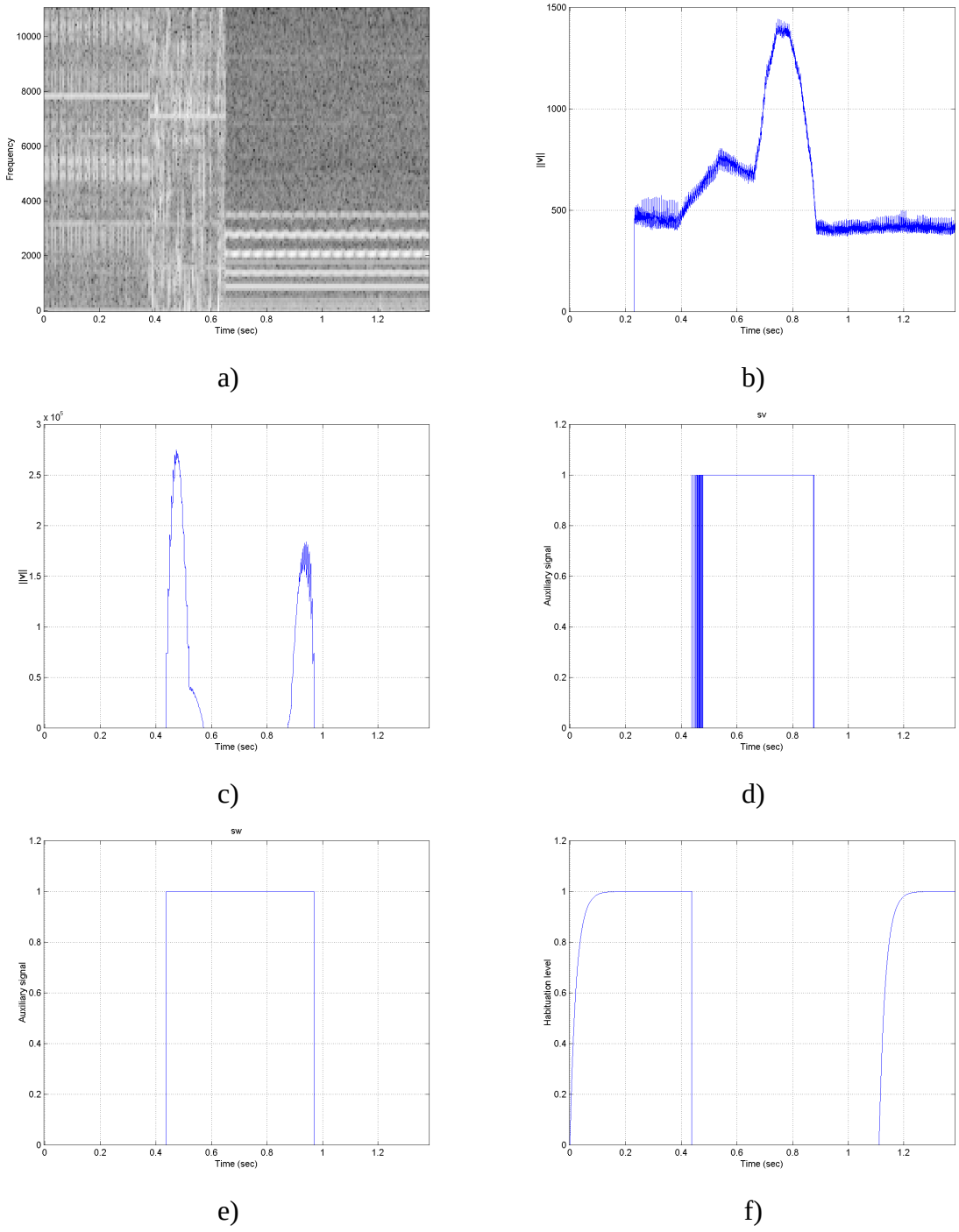


Figure 5.33: a) Spectrogram of the audio signal, b) evolution of the first-level (l_2) norm of the variance vector, c) evolution of the second-level (l_2) norm of the variance vector, d) first-level auxiliary signal, obtained using a threshold of 600, e) second-level auxiliary signal, obtained using a threshold of 1000, f) habituation level, using $\tau = 1, \alpha = 0.002$.

Chapter 6

Action

"CHRISTOF: We've become bored with watching actors
give us phony emotions. We're tired of
pyrotechnics and special effects. While
the world he inhabits is in some
respects counterfeit, there's nothing
fake about Truman himself. No
scripts, no cue cards...It isn't always
Shakespeare but it's genuine. It's a life."

- The Truman Show, Screenplay, by Andrew Niccol.

This chapter describes the actions that CASIMIRO can perform. Section 6.1 explains how facial expressions have been implemented in CASIMIRO. Voice generation is described in Section 6.3, along with brief descriptions of expressive talk and aspects of humour in the language.

6.1 Facial Expression

We begin this section by describing previous work on robots that are able to display facial expressions. A very simple robot face is that of Minerva [Thrun *et al.*, 2000], already mentioned in Chapter 1. It has four degrees of freedom, one for each eyebrow and two for the mouth. Eyebrows rotate over their centres. The mouth has a red elastic band. Each side of the band is hooked to a leg of the servomotor, being motion limited by three pins. Despite its simplicity, Minerva's face can produce a significant effect on a human observer. It can adopt

four basic expressions: neutral, smile, sadness and anger.

Another robotic face than can display four basic expressions (neutral, surprise, anger and fear) is Aryan (see Chapter 1). The face has 6 degrees of freedom: jaw, left/right eye pan, eyes tilt, eyebrows elevation and eyebrows arcing. Despite (or perhaps because of) its simplicity, experiments showed that expression recognizability is clearly better than in other two robotic faces. Aryan's face has not the blinking function. Apart from keeping the eye-balls clean and moist, blinking is associated to mental capacities such as concentration and nervousness. There is also a relationship between blink frequency and emotional state. Someone who is tired will blink more frequently and for a longer duration than someone who is well rested [Stern, 2004].

Kismet's face has eyebrows (each one with two degrees of freedom), eyelids (one degree of freedom) and mouth (one degree of freedom). The mouth has gained additional degrees of freedom in the latest versions. Kismet can adopt expressions of anger, fatigue, fear, disgust, excitation, happiness, interest, sadness and surprise, all of them easily recognizable by a human observer. Kismet's facial motion system was divided into three levels. In a first level there are processes that control each motor. In the next level there are processes that coordinate the motion of facial features, like for example an eyebrow. In the third level, there are processes that coordinate the facial features to form the basic expressions. This three-level framework allows to decompose the modelling work in a natural and scalable way. Kismet also uses an intensity for each expression, which is a degree with respect to a pose considered neutral.

The *WE-3RIV* robot [Miwa *et al.*, 2001, Takanishi Laboratory, 2003] is very different than Minerva and Kismet in the sense that it has a total of 21 degrees of freedom for the face alone: 4 for the eye balls, 4 for the eyelids, 8 for the eyebrows, 4 for the lips and 1 for the jaw. The basic expressions that it can adopt are: neutral, happiness, anger, surprise, sadness, fear, disgust, "drunk" and shame. As in Kismet, it uses an intensity measure for the expressions.

6.1.1 Functional Model

A level hierarchy like that described in the previous paragraphs was used to model facial expressions in CASIMIRO. Groups of motors that control a concrete facial feature are defined. For example, two motors are grouped to control an eyebrow. For each of the defined motor groups, the poses that the facial feature can adopt are also defined, like 'right eyebrow raised', 'right eyebrow neutral', etc. The default transitions between the different poses uses the straight line in the space of motor control values.

Expression: "Surprise"		
Group: Mouth	Pose: Open	Degree: 90
Group: Right eyebrow	Pose: Raised	Degree: 90
Group: Left eyebrow	Pose: Raised	Degree: 90
Group: Right ear	Pose: Raised	Degree: 100
Group: Left ear	Pose: Raised	Degree: 100
Group: Right eyelid	Pose: Raised	Degree: 80
Group: Left eyelid	Pose: Raised	Degree: 80

Table 6.1: Typical definition of an expression.

The modeller is given the opportunity to modify these transitions, as some of them could appear unnatural. A number of intermediate points can be put in all along the transition trajectory. Additionally, velocity can be set between any two consecutive points in the trajectory. The possibility of using non-linear interpolation (splines) was considered, although eventually it was not necessary to obtain an acceptable behaviour.

The first pose that the modeller must define is the neutral pose. All the defined poses refer to a maximum degree for that pose, 100. Each pose can appear in a certain degree between 0 and 100. The degree is specified when the system is running, along with the pose itself. It is used to linearly interpolate the points in the trajectory with respect to the neutral pose.

In another level, facial expressions refer to poses of the different groups, each with a certain degree. "Surprise", for example, could be represented by Table 6.1.

The facial expression is specified while the system is running along with a degree that allows, by multiplication, to obtain the degree to apply to the poses of the different groups. For more control, the modeller can also specify a time of start for each group. This way, "surprise" could be achieved by raising first the eyebrows and then opening the mouth. Now it is easy to see that the level hierarchy allows to move individual features, like for example winking, blinking or opening the mouth for talking. With respect to this particular case, the motion of the mouth can be combined with poses of other facial features, producing combinations like for example "talking with expression of surprise".

6.1.2 Transitions Between Expressions

Earlier it was mentioned that the degree of a pose can be specified at run time, while the defined poses referred to the maximum degree. How to obtain the trajectory from a pose A with a degree G_j to a pose B with a degree G_f? In other words, at a given time a group

is at pose A with degree G_i and we want the system to adopt pose B with degree G_f. For a two-motor group, the procedure is depicted in Figure 6.1, where N represents the neutral pose, X the point corresponding to the initial degree and Y the point corresponding to the final degree. To obtain the expression that gives the transition trajectory the following limit conditions are imposed:

a) If $G_i = 0 \Rightarrow T = XNY$

b) If $G_f = 0 \Rightarrow T = XNY$

c) If $G_i = 1 \Rightarrow T = (1 - G_f) \cdot XNY + G_f \cdot XABY$

d) If $G_f = 1 \Rightarrow T = (1 - G_i) \cdot XNY + G_i \cdot XABY$

The trajectory that fits these restrictions is:

$$T = [1 - ((G_i \geq G_f)\ ?\ G_f : G_i)] \cdot XNY + ((G_i \geq G_f)\ ?\ G_f : G_i) \cdot XABY \qquad (6.1)$$

Symbols ? and : correspond to the IF-THEN-ELSE instruction (as in the C language). The trajectory equation is continuous in the values of G_i and G_f. The same relationship is used to obtain velocities along the trajectory.

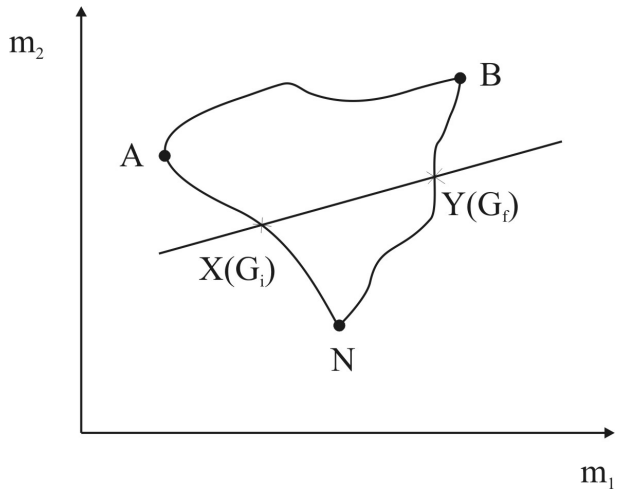


Figure 6.1: Transitions between expressions in motor space.

6.1.3 Implementation

The modelling framework described was implemented in the form of a pose editor. The editor gives the modeller a simple and interactive way to define and test poses, transitions, etc. This editor uses a low-level library for control of the servomotors. The specifications of poses and transitions can be saved to a file. Later, this file will be the only thing necessary to reproduce (and generate) the different movements and trajectories. The pose editor can thus work both at design and run time, and it can also be controlled by other modules.

The modeller has to follow a few simple steps: connect the ASC16 board (see Section 4.1), establish the physical motion limits and specify the poses and transitions. The physical limits represent the minimum and maximum values that can be given to a motor, and they are also saved in the file.

The central part of the editor allows to define groups of motors, poses and transitions, see the following figure. On the top left side there are scrollbars to move motors. On the central left side motor groups are defined. On the central right side poses and transitions can be defined for a selected group. The big STOP button stops all motors immediately.

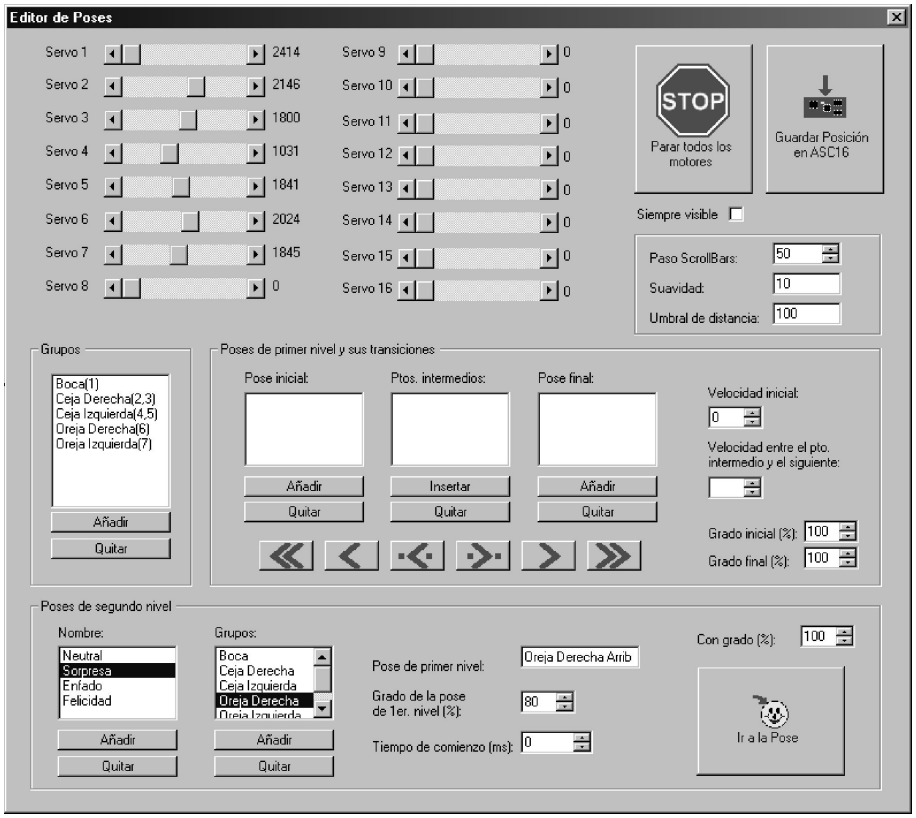


Figure 6.2: Main window of the pose editor.

Two parameters control motion continuity. Each part of the trajectory can have a different velocity (recall that intermediate points can be defined) and, moreover, all the velocities are adjusted so that the motors stop at the same instant. The motor control board sends a signal when the motor has reached the final destination (or an intermediate point). However, there is a slight delay between the sending of this signal, its processing in the PC, and the next motion command, which causes the motion to be discontinuous. There is a "softness" parameter that makes the control board send the final of motion signal ahead of time. There is also a distance threshold to eliminate from trajectories too close consecutive points.

The pose editor uses a buffer of requests that momentarily stores "Go to pose X with degree Y" petitions coming from other modules. The buffer is necessary, as requests take a certain time to complete. The buffer checks all the time the pending requests. Those compatible with the request currently being served (i.e. they do not have any motor in common) are also served, always in the order in which the requests were made.

Blinking has great importance for the robot. Blinking human forms have been appraised to have significantly higher accuracy and to be more intelligent than other human-like forms [King and Ohya, 1996]. In our implementation, blinking is treated as a special case. In principle, with the framework described it would have to be implemented using to consecutive commands sent to the pose editor, one for each eyelid. However, the eyelids must return to the original position, no matter what, and the velocity should always be set at maximum. Therefore, the system automatically converts two motion commands into four, and maximum velocity is imposed. In the case of winking, one motion command produces two commands internally. Also, to obtain maximum velocity and continuity in the blinking motion, the robot only blinks when no other motor on the face is working.

In conclusion, the pose editor is the module that allows a modeller to define how the facial features will move to adopt the different expressions. The pose repertoire can be upgraded (if motors are added) easily. More flexibility could have been given to the program. That was discarded because the main objective is to get the face to move well, as observed by humans, and because in practice the number of motors is relatively reduced. Figure 6.3 shows the facial expressions modelled in CASIMIRO using the pose editor.

6.2 Neck

As commented in Section 4.1 there is a 2-DOF neck under the robot's face. The neck module of CASIMIRO's software is very simple. A velocity PID controller is used for small move-

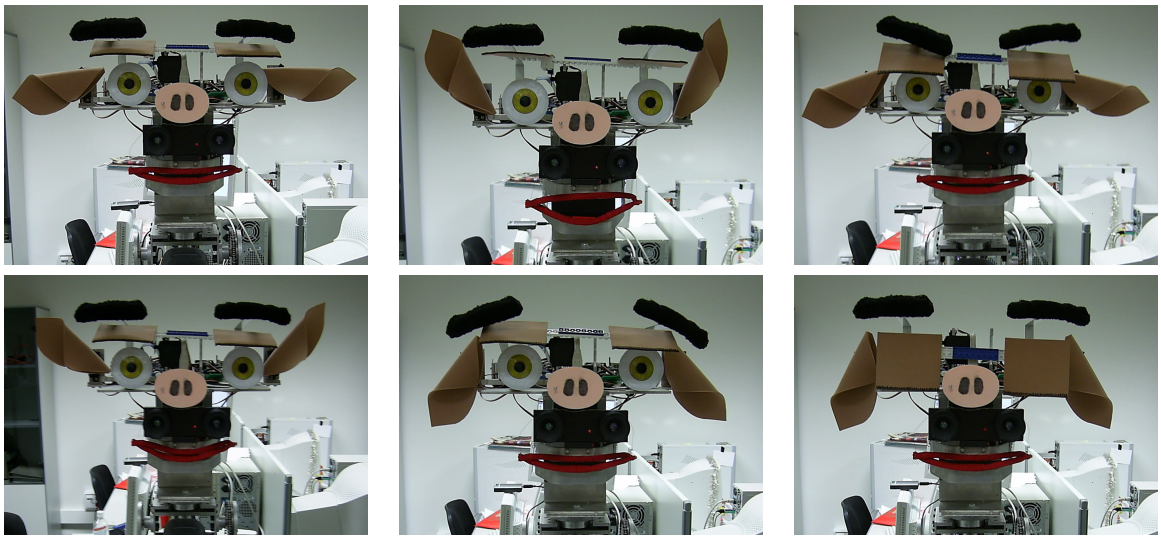


Figure 6.3: Facial expressions modelled in CASIMIRO. From top to bottom, left to right: Neutral, Surprise, Anger, Happiness, Sadness, Sleep.

ments. For large neck movements, only position commands are used. This way the neck is able to do smooth tracking and also respond rapidly in case of large displacements (when the robot fixates on a different person).

There was, however, a major problem that had to be tackled. The omnidirectional camera (which controls the neck through the attention module) is not aligned vertically with the axis of the pan motor. This is due to the physical construction of the robot (the camera would have to be placed under the table or on top of the robot, being both positions unsightly).

The situation is depicted in Figure 6.4. The omnidirectional camera was placed on the table, in front of the robot. Thus, the camera is at a distance f of the pan motor axis, which is $13cm$ in our case. The robot must turn its head β degrees. That angle, which is smaller that α, is given by:

$$\beta = arcsin(\frac{a}{d}) \tag{6.2}$$

, where

$$d = \sqrt{(b+f)^2 + a^2} \tag{6.3}$$

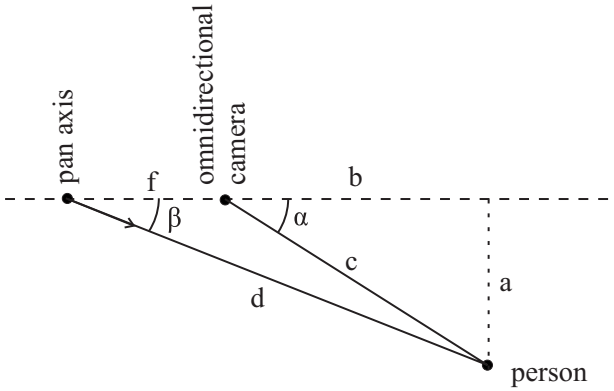


Figure 6.4: Position of the camera and the pan axis of the neck.

, and

$$a = c \cdot sin(\alpha) \tag{6.4}$$

$$b = c \cdot cos(\alpha) \tag{6.5}$$

Once an angle α has been measured by the omnidirectional vision module, there must be a correction of that angle to get a value for the neck angle β, so that the robot can look toward the detected person. The exact value for β depends on the distance c to the person, which is unknown.

As an approximation, we may use the typical interaction distance value of $100cm$ in Equation 6.2 and obtain an expression for β that does not depend on c. Besides, we approximated the trigonometric function by a (much faster) quadratic. The quadratic fitting to the trigonometric function was obtained using the least-squares method. These approximations are good enough, as can be seen from Table 6.2.

6.3 Voice Generation

In natural language generation four basic schemes are used: canned text, templates, cascaded items and features [School of Computing, University of Leeds, 2003, FLUIDS Project, 2003]. In the canned text approach, specific fixed phrases are used, without modifications whatsoever. This is very easy to implement, although it is rather inflexible. To avoid repeatability a large number of phrases would have to be introduced beforehand. Templates are phrases that

c	α	error
75cm	0º	0.10º
	45º	1.32º
	90º	1.94º
100cm	0º	0.10º
	45º	0.10º
	90º	0.49º
125cm	0º	0.10º
	45º	0.99º
	90º	1.96º
150cm	0º	0.10º
	45º	1.60º
	90º	2.94º

Table 6.2: Theoretical angle errors using the implemented approximation.

include empty slots that are filled with data, like a mail merge. Cascaded items and features are even more complex models.

CASIMIRO uses canned text for language generation, not least because of its simplicity. A text file contains a list of labels. Under each label, a list of phrases appear. Those are the phrases that will be pronounced by the robot. They can include annotations for the text-to-speech module (a commercially available TTS system was used. Annotations allow to change parameters like word intonation, speed, volume, etc.). Labels are what the robot wants to say, for example "greet", "something humorous", "something sad", etc. Examples of phrases for the label "greet" could be: *"hi!"*, *"good morning!"*, *"greetings earthling"*.

The Talk module, which manages the TTS system, reads the text file when it starts (Appendix A shows the phrase file. The text is in Spanish). It keeps a register of the phrases that haven been pronounced for each label, so that they will not be repeated. Given a label, it selects a phrase not pronounced before, randomly. If all the phrases for that label have been pronounced, there is the option of not saying anything or start again. The text file include annotations that had to be added to each particular phrase to enhance the naturalness of the utterance.

In [Koku *et al.*, 2000] the humanoid ISAC is used to investigate an interesting interaction technique. When the robot detects a person, it tries to initiate a conversation with him/her by using a simple phrase based on daily facts. Typical phrases are *"Did you know that Tutu calls for abolition of death penalty"* or *"Listen to this, Fed in focus on Wall Street"*. This system was based on the idea that when people share the same environment for a while, they do not start a conversation by greeting each other, but instead they generally initiate

the conversation based on some daily facts that the other side might be interested in. In CASIMIRO, a somewhat similar technique has been used. It was considered that it would be very useful that the robot could say things that are very popular at present, like for example something about Prince Felipe's fiancée (which was one of the main pieces of gossip in Spain at the time of writing). That would be interpreted positively by the observer. As an example, consider the phrase *"Prince Felipe's fiancée is definitely beyond my expectations!"*. The phrase is fixed, though it will only pronounced if the thing, event or person is popular at present. Otherwise, it would not make sense to pronounce the phrase. These cases are implemented by preprocessing the text file mentioned above. The file contains those X's inserted in the phrases. The substitutions are made when the system starts, producing a file ready to be used by the Talk module.

Another example: if "Beckham" is popular at present, then the phrase *"I'm more famous than Beckham!"* can be pronounced. As another example, in our region of Spain, it is humorous and relatively common to greet somebody with the name of a person who is extremely popular at the time. How to determine if something or someone is popular? The solution implemented looks in electronic newspapers of the last K days for mentions of a term (or phrase). If the term has a frequency of apparition that exceeds a threshold then it is considered popular. In the text file, "Popular" is a special label that gives the list of terms to look for. In the rest of the file those terms appear as labels, with their phrases to pronounce.

If many terms are popular they all will be used, in a random order. Once a popular term is used it will not be used again. If the threshold is carefully chosen and the phrases are not too specific then it would seem to be in context. For each newspaper, terms are counted only once (i.e. whether they appear in it or not).

Note that the robot should know when it is appropriate to speak. Turn-taking is a basic form of organization for any conversation. Before speaking, the Talk module interrogates the sound localization module (Section 5.2) in order to know if any sound has been detected in the last K milliseconds (currently K=2000). If so, the robot assumes that someone is speaking and postpones its action.

6.3.1 Expressive Talk

The Talk module pronounces phrases with an intonation that depends on the current facial expression. Before pronouncing a phrase, the Talk module interrogates the pose editor to obtain the current facial expression of the robot. The mapping from expression to voice parameters is based on Table 6.3.

	fear	anger	sorrow	joy	disgust	surprise
speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower	much faster
pitch average	very much higher	very much higher	slightly lower	much higher	very much lower	much higher
pitch range	much wider	much wider	slightly narrower	much wider	slightly wider	
intensity	normal	higher	lower	higher	lower	higher
voice quality	irregular voicing	breathy chest tone	resonant	breathy blaring	grumbled chest tone	
pitch changes	normal	abrupt on stressed syllable	downward inflections	smooth upward inflections	wide downward terminal inflections	rising contour
articulation	precise	tense	alluring	normal	normal	

Table 6.3: Effect of emotions on human speech [Breazeal, 2002].

Facial expression	*s*	*b*	*f*	*v*	*h*
Surprise	50	70	ND_f	60	0
Sadness	35	50	40	45	0
Anger	50	75	70	60	0
Happiness	50	70	70	60	0
Sleep	35	ND_b	30	45	50

Table 6.4: Values for NM used. s=speed, b=pitch baseline, f=pitch fluctuation, v=volume, h=breathiness.

With the TTS system used four voice parameters can be controlled: speech rate, pitch average, pitch range and intensity. Besides, the rising contour for the surprise expression can be achieved. Are these four parameters sufficient to convey the emotion represented in the facial expression? In [Schröder, 2001] it is stated that some emotions can be recognized reasonably well through average pitch and speed.

Each of the four voice parameters have a value N associated that indicates how strong is the effect or characteristic. As the facial expression has a degree (between 0 and 100, 0=neutral), the following value for N was used:

$$N = round\left(ND\left(1 - \frac{g}{100}\right) + NM\left(\frac{g}{100}\right)\right), \tag{6.6}$$

where g is the degree of the facial expression, ND is the default value for N (supposedly corresponding to a neutral intonation) and NM is the maximum (or minimum) value that N can have. Values for NM for each expression and parameter were established manually, starting from the values of Table 6.3 and the indications in [Montero *et al.*, 1998], see Figure 6.4. An additional parameter set was included for the expression "Sleep". In this case, the intonation is monotonous and slow, and breathiness is high.

6.3.2 Local Accent

CASIMIRO has been built in the Canary Islands, a region of Spain where people has a softer accent that in the mainland. For local people, that accent sounds more familiar and promotes empathy. The most significant variation in the Canarian accent is that the "c" (as in "hacer") is pronounced as a "s" (see [Hernández, 1999] for more details). Mexican Spanish, also available in our TTS system, is very similar, although phrase intonation profiles are different.

The voice generated by the TTS system has a standard Spanish accent. Simple substitutions were made in order to get a local accent, see Table 6.5. Not all the cases are covered, though the effect was easily noticed.

Substring	*Substitution*
"ce"	"se"
"ci"	"si"
"z"	"s"
"ado"*	"áo"
"edo"*	"éo"
"ido"*	"ío"
"udo"*	"úo"
"eda"*	"éa"
"ido"*	"ía"
"oda"*	"óa"
"uda"*	"úa"

Table 6.5: Substitutions for getting the local accent (*=only when it appears at the end of a word).

This section has described the simple -yet functional- language generation system implemented in CASIMIRO. Sophisticated language generation entails a number of requirements. First, the robot has to have something meaningful to say. It could certainly say many things, though they should make sense. Second, there should be variety in the vocabulary and phrases. Otherwise the robot will appear repetitive and dumb. Third, we believe that this effort should be accompanied by speech recognition capabilities, so that conversation can be rich and lasting. Some techniques have been proposed for compensating speech recognition deficiencies in dialog systems, such as reducing the vocabulary.

In principle, it would be desirable that voice synthesis produced a more human voice.

However, we believe that, given the external aspect of the robot, it is more appropriate to use a somewhat artificial, metallic voice. Otherwise the robot would provoke false expectations and would seem strangely balanced in its capabilities.

Additional improvements to the pose editor could make mixture of expressions possible. That is, the possibility of making the face adopt two or more expressions at the same time. Of course, the practical implementation of this option would require a higher number of degrees of freedom on the face. On the other hand, variability in expressions was considered (i.e. using randomness as in [Quijada *et al.*, 2004] or Perlin noise [Perlin, 1995]). However, it was discarded because it would complicate excessively the implementation and above all it would increase the time in which motor noise is present (which may affect sound localization).

Principles of animation could also be applied to the pose editor. Many simple techniques for enhancing the perceived motion are considered in [van Breemen, 2004]. Some of them caught the attention of the author:

- Squash and strecht: things made of living flesh change their shape while moving
- Secondary actions: they make scenes richer and more natural. For example blinking while turning the head
- Timing: a fast eye blink makes the character alert and awake, whereas a slow eye blink makes the character tired

Chapter 7

Behaviour

"A good name, like good will, is got
by many actions and lost by one"

Lord Jeffery.

This chapter describes the modelling of behaviour in CASIMIRO, which includes both reactive and deliberative aspects. A reflex system (Section 7.1) implements direct perception-action associations, while the higher-level aspects of the robot behaviour are modelled with an action selection module (Section 7.2). An emotional module was also implemented to account mainly for the external observable aspects of emotions, like facial expressions (Section 7.3).

7.1 Reflexes

Reflexes are inborn, automatic -involuntary- responses to particular stimuli. Examples are: the patellar reflex, the knee-jerk reflex commonly tested during routine physical examinations, and the eyeblink reflex, where the eyes closes in response to a puff of air striking the cornea. Most reflexes are simple, but even fairly complicated activities can be reflexive in nature.

In humans, a repertoire of reflexes appear soon after birth. Some of these are reflexes of approach (to increase contact with the stimulus), while others are reflexes of avoidance (to avoid intense or noxious stimuli). The behaviours of young infants are very much confined to reflexes, and are gradually replaced with voluntary actions. Adult behaviour is dominated

by voluntary actions, though reflexes are still present.

With respect to robotics, the most influential work related to reflexes is that of Brooks' Subsumption Architecture [Brooks, 1986], already mentioned above. Brooks argues that instead of building complex systems and use them in simple worlds, we should build simple systems and use them in the real, unpredictable world. Reflexive responses are necessary in a complex, dynamic environment. Response to stimuli is reflexive, the perception-action sequence is not modulated by cognitive deliberation.

Implementing fixed reflexes in a robot is relatively straightforward. The following 3 simple reflexes were implemented in CASIMIRO:

- If the robot hears a loud sound, it increases its arousal and valence, especially arousal. This in an example of primary emotion (see Section 7.3).
- If the robot sees very intense light, it closes the eyes and then frowns. Intense light is present when the number of pixels with value greater than a threshold exceeds another threshold. This is carried out by the face detection module (Section 6.1).
- If the robot sees an object that is too close, it closes the eyes, frowns a little and pans for a while so as to look in other direction. The module Omnivision can detect close objects by detecting when the estimated distance of any object in the image is lower than a threshold (see Section 5.1).

These actions are directly triggered by the corresponding stimuli, except for the first one that exercises its effect through the Emotions module (see below). The first two are reflexes of avoidance.

Reflexes were implemented as fixed perception-action couplings. If a particular percept is present that triggers the reflex, the robot will always perform the same action. That mechanism can make the robot appear too predictable. The next section describes an action selection module that allows the robot to show a more elaborate behaviour.

7.2 Action Selection

Action selection is the process whereby an action or output is produced, given the current inputs and state of the system. It is thus sort of a brain for the robot, that is continuously deciding what to do next. Action selection systems usually specify sets of possible actions, possible inputs and goals that should guide the selection. The way in which those sets are

used defines the action selection method. Goals, at times conflicting, may be pursued in parallel.

The action selection problem has inherent difficulties. Robots generally assume certain states of the environment that may not be true, not least because it can be dynamical and unpredictable. Sensors are imperfect in terms of quantization and updating frequencies. Also, the robot's actuators may not work as they are supposed to do. The selected action may not be optimal because the response must be given in a limited time. Finally, some circumstances require high reactivity while others need deliberative actions (planning). All these possibilities have given rise to many models, see [Tyrrell, 1993, §8] for a survey.

Humphrys [Humphrys, 1997] divides action selection methods into three categories: hierarchical, serial and parallel. Hierarchical models are multi-level systems where the modules are built into some kind of structure. Some modules have precedence over others, and control flows down to their submodules. Kismet's behaviour system has a hierarchical architecture, with three main branches, each specialized to fulfil a drive: the social drive, the fatigue drive and the stimulation drive. In serial models some modules must terminate before others can start. Finally, in parallel models modules can interact with and interrupt each other.

Pirjanian [Pirjanian, 1997] proposes a different taxonomy: reactive, deliberative and hybrid. Deliberative systems consist of a series of processing blocks. Information flows from sensors to actuators through these blocks, generally following a sequential order of execution. This approach is inappropriate for dynamical environments in which timely responses are needed. Reactive systems combine a set of behaviours, each connecting perception to action. The combination of these behaviours produces the emergent global behaviour of the system. Hybrid systems try to combine the advantages of the two approaches.

7.2.1 Behaviour Networks

Maes' Behaviour Networks [Maes, 1990, Maes, 1989] constitute an action selection model which can be categorized as serial and hybrid. Behaviour networks have interesting features such as reactivity, robustness (the failure of a single module or behaviour does not necessarily have catastrophic effects), management of multiple goals and its cheap and distributed calculations. Moreover, it is not necessary to have explicit representations of plans to achieve goal-orientedness. For these and other reasons this action selection model was initially chosen for CASIMIRO.

In this type of network, nodes, which have an associated action, are described with

Drink:		
	Preconditions:	*At-location-A*
	Add list:	Not-thirsty
	Delete list:	
Eat:		
	Preconditions:	*At-location-B*
	Add list:	Not-hungry
Approach food:		
	Preconditions:	
	Add list:	At-location-B
	Delete list:	At-location-A
Approach water:		
	Preconditions:	
	Add list:	At-location-A
	Delete list:	At-location-B

Table 7.1: Example behaviours. Not-hungry and Not-thirsty are goals.

predicate lists:

- A list of preconditions that should be true in order to execute the action
- A list of predicates that will be true after executing the action (Add list).
- A list of predicates that will be false after executing the action (Delete list).

These lists conceptually link nodes with each other. Each network node is called behaviour module and it represents one of the available actions. For illustrative purposes, we provide in Table 7.1 an example of a simple behaviour network taken from [Goetz, 1997].

The node activation mechanism depends on the concept of "energy" flow through the network. Goals to be fulfilled and Add lists introduce positive energy into the network. Delete lists introduce negative energy. After energy has flowed through the network, the node with the maximum energy is selected to be executed, provided that it exceeds a certain threshold. Otherwise the threshold is reduced by 10% and the cycle is repeated until some behaviour is selected. Two parameters are important: ϕ is the amount of energy injected into the network for a valid predicate of the current state of the environment and γ is the amount of energy injected into the network by a goal.

Behaviours compete and cooperate to select the action that is most appropriate given the environment state and system goals. Actions are selected in a sequence that favours goal fulfilment. The system is thus deliberative, although not entirely so because at each

step environment inputs can make the system select any behaviour. Therefore, both cases of planning and reactivity are dealt with. The weight given to one or other capacity depends on the ratio of ϕ to γ.

7.2.2 ZagaZ

CASIMIRO's action selection module is based on ZagaZ [Hernández-Cerpa, 2001]. ZagaZ is an implementation of PHISH-Nets [Rhodes, 1996], an enhanced version of Maes' Behaviour Networks. It has a graphical interface that allows to execute and debug specifications of PHISH-Nets. Specifications have to be compiled before they can be executed. There are two compilation modes: Release and Debug. Figure 7.1 shows the main windows of the application.

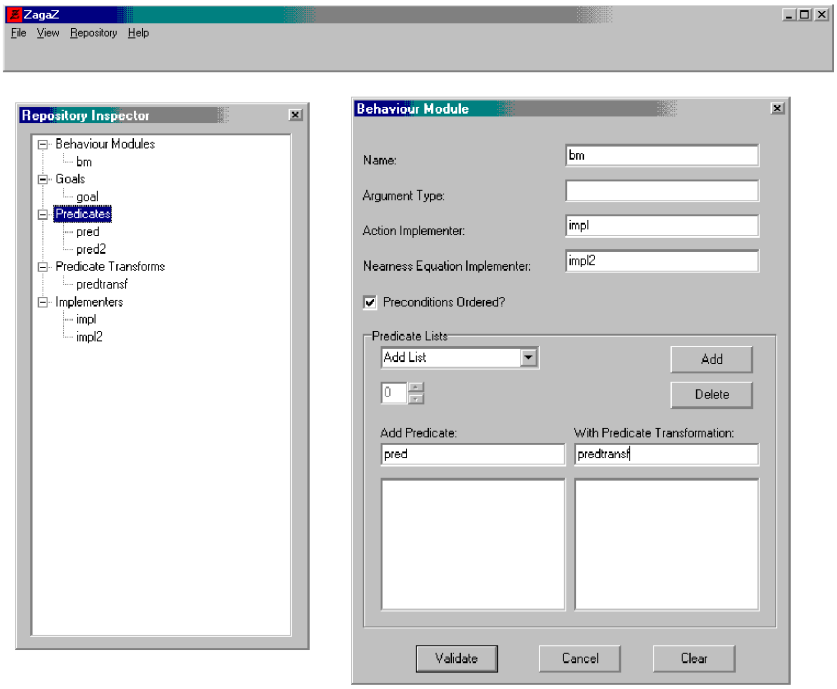


Figure 7.1: Main windows of the ZagaZ application. (Courtesy of D. Hernández)

ZagaZ capabilities are not fully exploited in CASIMIRO, although it is thought they will be used in the future as more complex behaviours are added. Currently, the system works as a priority-guided rule system. The behaviour that is finally selected for execution is chosen among those executable according to preset priorities (see below). Also, the Add and Delete lists are not used, for they are actually implemented through the memory system. This way a better control can be exercised over the memorized and forgotten items. Only

one goal was used, and it is present in the Add list of each behaviour. The implementation guarantees that the goal is never achieved.

7.2.3 Implemented Behaviours

CASIMIRO's behaviour was made intentionally simple. It engages in interaction with the subjects that enter its interaction space (roughly a 180^o zone in front of it). It tries to tell the subjects poems, jokes and data about itself, in that order. Basic interaction stability is maintained: each person is greeted, and at times the robot also makes "continuity", "funny" comments. It can also detect when the person is distracted, which makes it change topic (between poems, jokes and data about itself). It can also ask the subject whether he/she wants to hear more phrases of the current topic. When the subject does not collaborate (i.e. when the robot can not accomplish its task as stated above) the robot gets angry.

Table 7.2 shows a list of the perceptions involved in the specification of CASIMIRO's behaviours. All of the perceptions are Boolean predicates. Some predicates are just perceptions of the world at a given moment. Others take their values from facts stored in memory (either in the individual's or in global memory), see Section 5.6. That is, they are true or false depending on the predicate being present in memory or not. For example, the perception Greeted is true only when the Greeted fact is present in memory (the Greet action inserts it). Memory is initially empty, the execution of behaviours adds predicates to it. From time to time a predicate is forgotten (i.e. removed from memory). The forgetting parameters associated to these predicates are shown in Table 7.3. These forgetting parameters are very important for the correct sequence of behaviours.

Since some predicates take their values from facts stores in memory, perceptions like "LikesMe" and "NotLikesMe" are not quite the opposite of each other. "LikesMe" is true when the robot has asked a "Do you like me" question and the answer was Yes, and false otherwise. "NotLikesMe" is true when the robot has asked a "Do you like me" question and the answer was No, and false otherwise.

Table 7.4 shows a list of the available actions. Note that, as commented in Section 3.5, actions are high-level. The TalkAbout[X] actions can turn into different phrases to say (see Section 6.3). Other actions like Greet can also produce different final actions.

Instead of using the energy value to select a behaviour, a priority scheme was used. Each behaviour action has an assigned priority (Table 7.5) that allows to pick only one behaviour when more than one is executable.

Table 7.6 shows a summary of the behaviours defined in ZagaZ.

Predicate	*Description*	*From memory?*
FrontalFace	true when a roughly frontal face of an individual has been detected	
PersonNotGreeted	true if the individual has not been greeted yet	yes, individual
NotClose	true if the individual is not close to the robot (a depth estimation obtained from the stereo camera is used)	
TooClose	true if the individual is too close to the robot	
NotTooClose	true if the individual is not too close to the robot	
Moving	true if the individual is moving around	
NotMoving	true if the individual is not moving around	
Familiar	true if the individual has been observed a minimum of time by the robot	
LikesMe	true if the robot considers that the individual likes it	yes, individual
NotLikesMe	true if the robot considers that the individual does not like it	yes, individual
Owner	true if the individual is the robot's owner	
NotOwner	true if the individual is not the robot's owner	
NotLookedAtOther	true if the robot has not looked at other individual	yes, global
TooManyPeople	true if more than 2 individuals are being detected around the robot	
FrontalSound	true if the robot has heard a sound coming from in front	
WavingHand	true if the Omnidirectional vision module has detected a waving hand (see Section 5.1)	
WavingHand1	true if the face detection module has detected a waving hand	
EmotionHighArousal	true if average arousal experienced with the individual exceeds a threshold (see Section 7.3)	
EmotionLowArousal	true if average arousal experienced with the individual is below a threshold (see Section 7.3)	
EmotionHighValence	true if average valence experienced with the individual exceeds a threshold (see Section 7.3)	
EmotionLowValence	true if average valence experienced with the individual is below a threshold (see Section 7.3)	
EmotionNotAngry	true if the robot is not angry	
NotIncreasedValence	true if the robot has not increased its valence	yes, global
NotIncreasedArousal	true if the robot has not increased its arousal	yes, global
NotDecreasedValence	true if the robot has not decreased its valence	yes, global
NotDecreasedArousal	true if the robot has not decreased its arousal	yes, global
ColdRoom	true if room temperature is below a threshold (the sensor is described in Section 4.1)	
HotDay	true if local temperature exceeds a threshold (local temp. is obtained from a web page)	
MorePoemsAvailable	true if there are more poems for the robot to say	
NotMorePoemsAvailable	true if there are not more poems for the robot to say	
WantPoem	true if the individual wants to hear poems	yes, individual
NotWantPoem	true if the individual does not want to hear (more) poems	yes, individual
OneItemSaid	true if the robot has said at least one item (either a poem, a joke or a phrase about itself)	
MoreJokesAvailable	true if there are more jokes for the robot to say	
NotMoreJokesAvailable	true if there are not more jokes for the robot to say	
WantJoke	true if the individual wants to hear jokes	yes, individual
NotWantJoke	true if the individual does not want to hear (more) jokes	yes, individual
WantRobotData	true if the individual wants to hear data about the robot	yes, individual
NotWantRobotData	true if the individual does not want to hear (more) data about the robot	yes, individual
MoreRobotDataAvailable	true if there are more robot data for the robot to say	
NotMoreRobotDataAvailable	true if there are not more robot data for the robot to say	
NotTalkedAbout[X]	true if the robot has not talked about [X]	yes, individual
Questioned[X]	true if the robot has asked the individual about [X]	yes, individual
NotQuestioned[X]	true if the robot has not asked the individual about [X]	yes, individual
IAmAlone	true if there is no one is the interaction area	

Table 7.2: List of available high-level perceptions.

Predicate stored in memory	*Assigned* k	*Assigned* l
Greeted	0	0
TalkedAboutTooManyPeopleAround	0.12	0
TalkedAboutColdRoom	0.05	0
TalkedAboutHotDay	0	0
TalkedAboutILikeYou	0.20	0
TalkedAboutIDontLikeYou	0.20	0
TalkedAboutSthingFunny	0.05	0
TalkedAboutSthingToAttractPeople	0.14	0
TalkedAboutTooClose	0.5	0
TalkedAboutMoving	0.7	4
TalkAboutSpeaking	0.7	4
TalkAboutWavingHand	0.7	4
TalkAboutPoem	0.15	0
TalkAboutJoke	0.15	0
TalkAboutRobotData	0.15	0
LookAtOther	1	0
IncreaseArousal	0.05	0
DecreaseArousal	0.05	0
IncreaseValence	0.05	0
DecreaseValence	0.05	0
QuestionedDoYouLikeMe	0.005	20
QuestionedDoYouWantPoem	0	0
QuestionedDoYouStillWantPoem	0.015	20
QuestionedDoYouWantJoke	0	0
QuestionedDoYouStillWantJoke	0.015	20
QuestionedDoYouWantRobotData	0	0
QuestionedDoYouStillWantRobotData	0.015	20
TalkAboutOwner	0.20	0
TalkAboutIAmAlone	0.30	30
TalkAboutIAmFedUpWithYou	0.20	10

Table 7.3: k and l forgetting parameters for predicates stored in memory.

Action	*Description*
Greet	The robot greets the individual
TalkAboutILikeYou	The robot says the individual that it likes him/her
TalkAboutColdRoom	The robot says sthing. when the room is too cold
TalkAboutHotDay	The robot says sthing. when the day is hot
TalkAboutIDontLikeYou	The robot says the individual that it does not like him/her
TalkAboutTooManyPeopleAround	The robot says sthing. when there are too many people around
TalkAboutSthingToAttractPeople	The robot says sthing. to attract people
TalkAboutSthingFunny	The robot says sthing. funny ("continuity" comment)
TalkAboutTooClose	The robot reprimands the individual for being too close
TalkAboutMoving	The robot reprimands the individual for being moving around
TalkAboutSpeaking	The robot reprimands the subject for being speaking
TalkAboutWavingHand	The robot reprimands the individual for waving his/her hand
IncreaseValence	The robot increases its valence level
DecreaseValence	The robot decreases its valence level
LookAtOther	The robot changes its focus of attention to other individual
IncreaseArousal	The robot increases its arousal level
DecreaseArousal	The robot decreases its arousal level
TalkAboutPoem	The robot tells a poem
TalkAboutJoke	The robot tells a joke
TalkAboutRobotData	The robot tell sthing. about itself
TalkAboutOwner	The robot tells sthing. to its owner
TalkAboutIAmAlone	The robot complains of its loneliness
QuestionDoYouLikeMe	The robot asks if the user likes it. This action activates either the LikesMe or NotLikesMe perceptions
QuestionDoYouWantPoem	The robot asks the subject if he/she wants to hear poems. This action activates either the WantPoem or NotWantPoem perceptions
QuestionDoYouWantJoke	The robot asks the subject if he/she wants to hear jokes. This action either the WantJoke or NotWantJoke perceptions
QuestionDoYouWantRobotData	The robot asks the subject if he/she wants to hear data about itself. This action activates either the WantRobotData or NotWantRobotData perceptions
QuestionDoYouStillWantPoem	The robot asks the subject is he/she wants to hear more poems
QuestionDoYouStillWantJoke	The robot asks the subject is he/she wants to hear more jokes
QuestionDoYouStillWantRobotData	The robot asks the subject is he/she wants to hear more data

Table 7.4: List of available high-level actions.

Action	*Priority*
Greet	1
WantPoem	1
WantJoke	1
WantRobotData	1
TalkAboutIDontLikeYou	1
TalkAboutILikeYou	1
LookAtOther	2
TalkAboutOwner	3
QuestionDoYouStillWantPoem	3
QuestionDoYouStillWantJoke	3
QuestionDoYouStillWantRobotData	3
QuestionDoYouWantJoke	4
QuestionDoYouWantRobotData	4
TalkAboutWavingHand	5
TalkAboutPoem	5
TalkAboutJoke	5
TalkAboutRobotData	5
TalkAboutSpeaking	6
TalkAboutMoving	7
TalkAboutTooClose	7
TTalkAboutIAmAlone	8

Table 7.5: Priorities assigned to actions. Actions that do not appear in the table have all an equal priority value of 0.

Name	*Action to execute*	*Preconditions*
TalkAboutTooClose	TTalkAboutTooClose	TooClose,NotTalkedAboutTooClose
TalkAboutMoving	TTalkAboutMoving	NotTalkedAboutMoving,Moving
Greet	TGreet	NotOwner,NotMoving,NotTooClose,PersonNotGreeted,FrontalFace
TalkAboutColdRoom	TTalkAboutColdRoom	ColdRoom,NotTalkedAboutColdRoom
TalkAboutHotDay	TTalkAboutHotDay	HotDay,NotTalkedAboutHotDay
TalkAboutIDontLikeYou	TTalkAboutIDontLikeYou	NotTalkedAboutIDontLikeYou,EmotionLowArousal,Familiar,EmotionLowValence
TalkAboutSthingFunny	TTalkAboutSthingFunny	NotTalkedAboutSthingFunny,NotMoving,NotTooClose,Familiar
LookAtOther	TLookAtOther	Owner,Familiar,NotLookedAtOther
TalkAboutSthingToAttractPeople	TTalkAboutSthingToAttractPeople	NotTalkedAboutSthingToAttractPeople,NotClose
QuestionDoYouLikeMe	TQuestionDoYouLikeMe	NotQuestionedDoYouLikeMe,QuestionedDoYouWantPoem,NotTooClose,NotMoving,Familiar
IncreaseValence	TIncreaseValence	NotIncreasedValence,LikesMe
TalkAboutTooManyPeopleAround	TTalkAboutTooManyPeopleAround	TooManyPeople,NotTalkedAboutTooManyPeopleAround
IncreaseArousal	TIncreaseArousal	LikesMe,NotIncreasedArousal
DecreaseArousal	TDecreaseArousal	NotDecreasedArousal,NotLikesMe
TalkAboutILikeYou	TTalkAboutILikeYou	EmotionHighArousal,Familiar,EmotionHighValence,NotTalkedAboutILikeYou
DecreaseValence	TDecreaseValence	NotLikesMe,NotDecreasedValence
TalkAboutSpeaking	TTalkAboutSpeaking	NotIAmAlone,NotTalkedAboutSpeaking,Familiar,FrontalSound
TalkAboutWavingHand	TTalkAboutWavingHand	WavingHand1,NotTalkedAboutWavingHand,WavingHand
QuestionDoYouWantPoem	TQuestionDoYouWantPoem	NotQuestionedDoYouWantPoem,NotTooClose,NotMoving,Familiar
TalkAboutPoem	TTalkAboutPoem	EmotionNotAngry,MorePoemsAvailable,WantPoem,NotTalkedAboutPoem
QuestionDoYouStillWantPoem	TQuestionDoYouStillWantPoem	OneItemSaid,NotQuestionedDoYouStillWantPoemOREmotionLowValence,WantPoem, MorePoemsAvailable,NotTooClose,NotMoving,Familiar
QuestionDoYouWantJoke	TQuestionDoYouWantJoke	NotQuestionedDoYouWantJoke,NotWantPoemORNotMorePoemsAvailable, Familiar,NotMoving,NotTooClose
TalkAboutJoke	TTalkAboutJoke	EmotionNotAngry,MoreJokesAvailable,WantJoke,NotTalkedAboutJoke
QuestionDoYouStillWantJoke	TQuestionDoYouStillWantJoke	OneItemSaid,NotQuestionedDoYouStillWantJokeOREmotionLowValence,Familiar, NotMoving,NotTooClose,MoreJokesAvailable,WantJoke
QuestionDoYouWantRobotData	TQuestionDoYouWantRobotData	NotQuestionedDoYouWantRobotData,NotWantJokeORNotMoreJokesAvailable, Familiar,NotMoving,NotTooClose
TalkAboutRobotData	TTalkAboutRobotData	NotTalkedAboutRobotData,WantRobotData,MoreRobotDataAvailable
QuestionDoYouStillWantRobotData	TQuestionDoYouStillWantRobotData	OneItemSaid,NotQuestionedDoYouStillWantRobotDataOREmotionLowValence,WantRobotData, MoreRobotDataAvailable,Familiar,NotMoving,NotTooClose
TalkAboutIAmAlone	TTalkAboutIAmAlone	IAmAlone,NotTalkedAboutIAmAlone
TalkAboutIAmFedUpWithYou	TTalkAboutIAmFedUpWithYou	NotTalkedAboutIAmFedUpWithYou,NotWantRobotDataORNotMoreRobotDataAvailable, Familiar,NotTooClose,NotMoving
TalkAboutOwner	TTalkAboutOwner	Owner,NotTalkedAboutOwner,NotTooClose,NotMoving,Familiar

Table 7.6: Behaviours implemented in ZagaZ.

7.3 Emotions

Many emotional models have been proposed both within the AI community and in psychology (see the Emotion Home Page [E. Hudlicka and J.M. Fellous, 2004]). The most well-known model is perhaps that of Russell [Russell, 1980], which considers that emotions fall in a bidimensional space, with orthogonal valence and arousal components, see Figure 7.2.

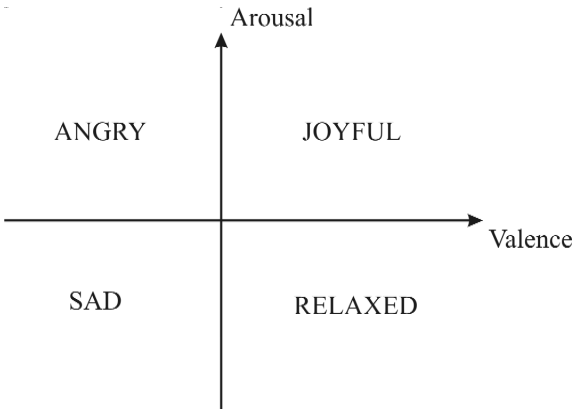


Figure 7.2: Arousal and valence emotional space.

This bidimensional space (also called circumplex structure) has received wide support in the related literature [Carney and Colvin, 2005]. Many forms of human emotional experience (judgement of the similarity between pairs of affect terms, self-reports of current emotion and from perceptions of similarity between static photographs of expressed emotion) point to an ordering of basic emotions around the perimeter of a circle with arousal and valence axes.

The central zone of that space would correspond to "no emotion". It is a sort of neutral state, where there is no feeling of being well or bad, excited or calmed. In this unemotional state, it is like emotions are nonexistent. They do not influence behaviour, attention or perception. This state is much like that of being "a machine", in which behaviour tends to detailed calculi and deliberation, without time restrictions. On the other hand, zones that are far from the centre of the emotional space correspond to normal emotional states in humans, though they are rarely contemplated in machines.

For Sloman there are only three types of emotions: basic, secondary and tertiary [Sloman, 2001]. Picard [Picard, 1997] and Damasio see only two types. Basic emotions come directly from certain stimuli. Other emotions arise after a cognitive appraisal. These two types of emotions are present in CASIMIRO:

- Basic emotions: Direct influence from sensors: If the robot hears a loud sound, it increases its arousal and valence, especially arousal.

- Secondary emotions: Influence in the Emotions module from ZagaZ (see Section 7.2).

The Emotions module maintains a position in a 2D valence and arousal space. The module receives messages to shift the current position in one or the two dimensions. The 2D space is divided into zones that correspond to a facial expressions. In order to simplify the module, it is assumed that the expression is given by the angle in the 2D space (with respect to the valence axis), and the degree is given by the distance to the origin. The circular central zone corresponds to the neutral facial expression. When the current position enters a different zone a message is sent to the pose editor so that it can move the face, and to the Talk module so that intonation can be adjusted. The facial expressions are assigned to the 2D space as shown in Figure 7.3. Values of arousal and valence are not always inside the exterior circle, though the expression degree is maximum for values that lie outside the circle.

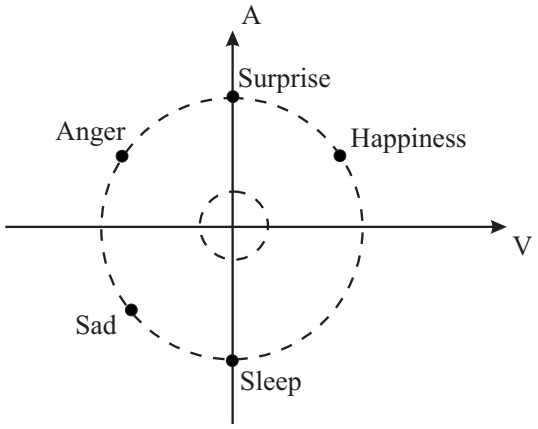


Figure 7.3: Assignment of facial expression according to the emotional state.

Relative displacements in the arousal and valence axes need a correction. Consider the case depicted in Figure 7.4 in which the current position in emotional space is P. If we want to lead the robot to Anger, we increase arousal and decrease valence with a displacement $\vec{d}$. However, the resulting position will be Q, which is associated to Surprise. Obviously, the effect of changes in arousal and valence depends on the current position, which is undesirable.

The correction, which we have not seen previously in the literature, is as follows. Given a displacement $\vec{d} = (v, a)$, the components of the new displacement vector $\vec{d'}$ are

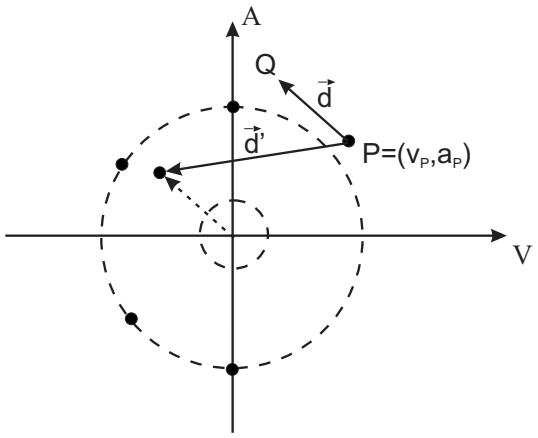


Figure 7.4: Effect of an increase in arousal and decrease in valence.

given by:

$$\vec{d'} = \left(\frac{v - v_P}{l}, \frac{a - a_P}{l} \right) \cdot min(l, m) \tag{7.1}$$

, where $l = sqrt((v - v_P)^2 + (a - a_P)^2)$ and $m = sqrt(v^2 + a^2)$.

Figure 7.5 illustrates the effect of the correction.

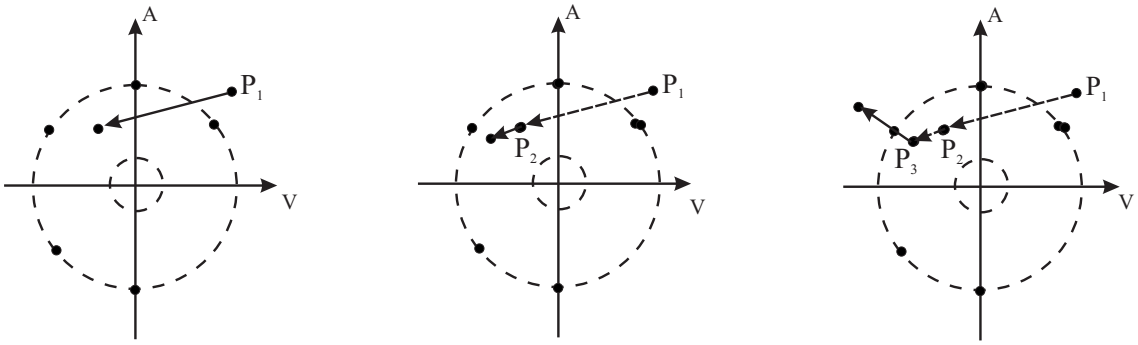


Figure 7.5: Effect of the correction in emotional space when three (arousal-increase,valence-decrease) displacements are submitted to the system. Note that the position in the emotional space tends to the desired expression. When the current position is at the angle of the desired expression only the distance to the centre increases, which in turn increases the degree of the expression.

A very simple decay is implemented: every once in a while arousal and valence are divided by a factor. This does not change the angle in the 2D space, and thus the facial expression does not change, only the degree. This procedure is in accordance with the fact that emotions seem to decay more slowly when the intensity is lower [Bui *et al.*, 2002]. In

[Reilly, 1996] a scheme is used in which each emotion can have its decay rate, being for example slower in anger than in startle. In our implementation each emotion can have a decay factor associated, by default set at 2. The use of decay factors (and other parameters like the magnitude of the displacements in the emotion space) allow the designer to define personalities [Hernández *et al.*, 2004].

The emotions that the robot has experienced while interacting with an individual are stored in the memory associated to that individual (see Section 5.6). Actually, memory is updated periodically with the mean values of arousal and valence experienced with that individual (a running average is used).

As for sleep, when the position in the 2D space has been for a certain time in the neutral state arousal is lowered by a given amount (valence will be zero). Besides, sleep has associated a decay factor below 1, so that it tends to get farther the centre instead of closer. This way, the emotional state will eventually tend to neutral and, in time, to sleep. When the robot is asleep the neck (see Section 4.1) stops working and the robot snores.

The Euclidean distance from the current position to the centre of the emotional space is used to send the Attention module a message to reduce the fixation in the current focus of attention. The Attention module periodically interrogates the Emotions module to obtain the current Euclidean distance, which affects the T component in Equation (5.13) (see Section 5.3). The larger the distance the larger the probability P that this influence makes the current FOA change. In particular, $P = max(0, min(d, R)/R - K)$, being $K \leq 1$ a constant to specify. The larger K, the less effect of the emotional state over attention.

Chapter 8

Evaluation and Discussion

"The important thing is not to stop questioning."

Albert Einstein.

The modules that make up CASIMIRO have been evaluated independently in previous chapters. The evaluation of those subsystems is relatively straightforward. Basically, each technique or subsystem is compared with other reference techniques or subsystems. The performance of the subsystems can be evaluated through numerical results, statistical techniques can be applied and relatively sound conclusions may be extracted.

How to evaluate the performance of a social robot as a whole? In an ideal situation we would have two different robots available and would evaluate them on the same tasks and under the same conditions. That would allow to obtain objective comparative measures. However, that approach cannot be taken due to the complexity and cost of the robots and their accompanying physical and human support. As we will see throughout this chapter, more subjective techniques may be required in the evaluation.

This chapter is devoted to observing and analysing how CASIMIRO works as a whole. Being a written account, an effort shall be made to illustrate with figures and tables how the robot behaves in typical interaction sessions. Section 8.1 shows a review of the available literature on the topic. Then, examples of short interaction sessions are described. These examples will allow the reader to know how a typical interaction session develops. Then, in Sections 8.4 to 8.6 we show the results of a series of interviews in which people showed their impressions about the robot.

8.1 Evaluating Social Robots, a Review

For industrial or navigation robots performance evaluation must include aspects like precision, repeatability, speed, autonomy, degrees of freedom or battery consumption. It is no wonder that these measures turn out to be inappropriate in our context of social robots. In fact, it would seem that they are rather contrary to what should be measured. Above all, social/interactive robots (still) do not have a well defined task as is the case for industrial robots. Scholtz and Bahrami [Scholtz and Bahrami, 2003] note that this fact makes typical human-computer interaction evaluation measures such as efficiency, effectiveness and user satisfaction inappropriate.

One of the most comprehensive evaluations of a social robot was made by Shibata and colleagues [Shibata and Tanie, 2001]. Their robot Paro was already mentioned in Chapter 1. Paro is a baby harp seal that was evaluated in terms of its positive mental effect on humans, such as joy and relaxation through physical interaction. Shibata already notes that contrary to the case of industrial robots, in which evaluation is objective, for Paro and other "mental commit" robots evaluation is mainly subjective [Wada *et al.*, 2003, Shibata *et al.*, 2003]. In their studies people's moods were rated using face drawings and Profiles of Mood States (POMS) [McNair *et al.*, 1971]. Comments of nursing staff were collected as extra information. When questioned, people had to select the face that best showed their emotional state. POMS is a popular questionnaire which measures people's mood, and it is mainly used in medical therapy and psychotherapy. It can measure 6 moods simultaneously: depression-defection, tension-anxiety, anger-hostility, vigour, confusion and fatigue. There are 65 items related to moods, and each item is punctuated on a 0-4 scale: 0=not at all, 1=a little, 2=moderately, 3=quite a bit and 4=extremely.

Shibata and colleagues were able to confirm that physical interaction with the robot improved the subjective evaluation, and that a priori knowledge has much influence in the results of the subjective evaluation. Principal Component Analysis was used to evaluate the answers of the subjective evaluation, and that allowed the researchers to extract the most meaningful factors.

In [Kanda *et al.*, 2001] a psychological experiment is described on people's impressions of a mobile robot designed for interaction. People answered a questionnaire to rate the robot in 28 adjective pairs (kind-cruel, distinct-vague, interesting-boring, etc.) with a 1-7 scale. Factor analysis was used to evaluate the results, confirming the ideas that gaze control promotes human-robot interaction and the computer skills of the subjects affect their impressions of the robot.

The study of Dautenhahn and Werry [Dautenhahn and Werry, 2002] is focused on the reactions of autistic children to robots. They propose a technique for quantitatively describing and analysing robot-human interactions. The technique is based on observing low-level behaviours in children, like eye contact, touching, handling, moving away, speech, etc. The procedure consists of recording videotapes of the sessions and then manually marking the behaviours. The posterior analysis is based on frequencies of apparition and span of the behaviours.

The task of observing sessions, as in the previous study, can be tedious even if recording equipment is used. An interesting approach is to automate the extraction of session outcomes. The system of [Littlewort *et al.*, 2003] describes a robot that uses computer vision to classify facial expressions with respect to 7 dimensions in real time: neutral, anger, disgust, fear, joy, sadness and surprise. When compared with human judgements of joy, the system achieved a correlation coefficient of 0.87. Obviously, this approach is useful not only for robot evaluation purposes, but also for improving the robot's observed behaviour.

8.2 Quantitative/Qualitative Measures

Desirable characteristics for performance evaluation of any robot are systematicity and, especially, "numerical results" that can be analysed with existing statistical techniques. Those objectivity characteristics would allow to extract solid conclusions about the robot and to compare it with others. However, as mentioned in the previous section, the evaluation of social/interactive robots implies a strong and unavoidable subjective component.

The natural approach is to try to systematize the inherently subjective evaluation of the robot. As seen above, questionnaires seem to be the preferred method. However, many aspects remain critical, such as the exact questions to be made and the scales for the answers. There is always the risk that the question does not exactly cover the aspect to be analysed. Also, the scales can left out some cases. This introduces errors that add to the inherent errors of the statistical techniques (especially for low number of cases. Factor analysis for example, generally requires more than 100 data points [Kanda *et al.*, 2001]), which makes extracting solid conclusions feasible only for certain precise aspects of the robot.

Another difficult point is the control of the biases in the experiments. At least two effects are present: the social desirability response bias (where people respond to studies in a way that presents them in a positive light), and the interviewer bias (where the interviewer influences the responses) [Dautenhahn and Werry, 2002].

Duffy and colleagues [Duffy *et al.*, 2002] argue that the field of social robotics inherits a methodology problem of sociology, i.e. whether quantitative methods capture the complexities in the social domain. They stand for a qualitative approach, for two main reasons:

- Information is lost and discarded when situations and scenarios are reduced to numbers.
- Quantitative methods and data often have qualitative bases:

> "A survey that collects people's opinions of Da Vinci's Mona Lisa painting (love, like, dislike, hate) is qualitative. However, when the survey requests the opinions to be on a scale from 0 to 10 is considered quantitative -is this quantised data better or more valid?", [Duffy *et al.*, 2002].

Despite the inherent limitations, the use of questionnaires and numerical scales seems to be the preferred method. The observation and analysis of interaction sessions has been also widely used, especially in the work of Dautenhahn. Both techniques have been used here. The following sections show examples of interaction sessions with CASIMIRO, as well as the results of a series of questionnaires completed by the individuals who interacted with the robot.

8.3 Interaction Examples

Table 8.1 shows an example interaction session with the robot. In this case one person enters the interaction area and takes an uncooperative attitude. Only the most significant interaction data are shown in the table (in particular, memory and predicate values have been omitted). Figure 8.1 shows the moments in which the subject was moving around and when he got too close to the robot.

Table 8.2 shows another example interaction with a more cooperative subject. In this case the interaction develops with less disruptions and the robot tells poems and jokes to the individual. The robot's emotional state is in this case more positive than in the previous interaction example (see Figure 8.2).

Table 8.3 shows another interaction example. In this case, only the executed behaviours are shown, along with the predicates that enter memory and those that are forgotten. First the robot fixates on a person. Then, at around time 56, the robot's owner enters

Time (s)	Observations	Executable behaviours	Emotional state
0	robot start		Neutral 100%
18	robot greets the subject	**Greet**,TalkAboutSthingToAttractPeople	
23	the subject is speaking, the robot reprimands him	**TalkAboutSpeaking**,QuestionDoYouWantPoem, TalkAboutSthingFunny	Anger 70%
29		**TalkAboutSthingFunny**,QuestionDoYouWantPoem	
31			Anger 35%
35	the subject is too close, the robot reprimands him	**TalkAboutTooClose**,QuestionDoYouWantPoem	
39			Anger 17%
40	the subject answers Yes	**QuestionDoYouWantPoem**	
47			Neutral 100%
50	the subject is speaking	**TalkAboutSpeaking**,TalkAboutPoem	
51			Anger 70%
57		**TalkAboutPoem**	
59			Anger 35%
67			Anger 17%
70	the subject is moving	**TalkAboutMoving**	Anger 88%
74	the subject is waving his hand	**TalkAboutWavingHand**	
75			Anger 100%
80		**TalkAboutMoving**	
86	the robot tries to change topic. Answer=Yes	**QuestionDoYouStillWantPoem**, TalkAboutSthingFunny,QuestionDoYouLikeMe	
93		**TalkAboutSthingFunny**,QuestionDoYouLikeMe	
97			Anger 57%
98		**TalkAboutPoem**,QuestionDoYouLikeMe	
106			Anger 28%
108	the subject answers No	**QuestionDoYouLikeMe**	
114			Anger 14%
115		**TalkAboutSthingFunny**,DecreaseValence, DecreaseArousal	
120		**TalkAboutPoem**,DecreaseValence,DecreaseArousal	

Table 8.1: Example interaction: one person interacting with the robot. Behaviours in bold were the ones executed by the robot.

the interaction area and calls out the robot. The robot fixates on the owner and tells him something. Then, at time 68, the robot fixates again on the first person and continues the previous interaction. Note that, in memory, owner data are treated as global. This can be done because there is only one owner.

In order to show the effect of person recognition, the robot was led to have short interactions with two individuals. Initially, individual A enters the interaction area, see Figure 8.3. The valence values show that he adopts an uncooperative attitude. The robot tries to ask him if he wants to hear poems, but the individual keeps moving around and the robot has to abort the questions. At time 55 individual A begins to leave the interaction area. The robot is then alone. At time 67 another individual enters the interaction area. This individual B is

Figure 8.1: Person moving around and getting closer to CASIMIRO.

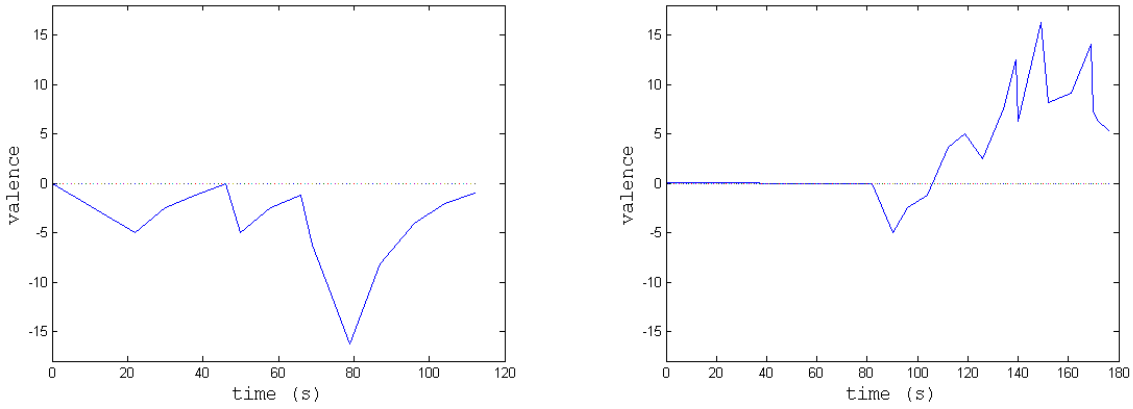


Figure 8.2: Valence values in the first (left) and second (right) example interactions.

more cooperative and answers affirmatively to the two questions made by the robot. Individual B leaves the area at around time 126. Then, individual A comes back at time 145 and is recognized by the robot, which avoids greeting him again. Note that upon seeing individual A the robot emotional state turns very negative, for its previous interaction with him ended unsatisfactorily. Table 8.4 shows what the robot says in each case.

The interaction examples shown in this section allow the reader to get a taste of how a typical interaction develops. Still, it would be very valuable to get an impression of the opinions of those who interacted with the robot. The following sections are devoted to it.

Time (s)	Observations	Executable behaviours	Emotional state
0	robot start		Neutral 100%
19	robot greets the person	**Greet**, TalkAboutSthingToAttractPeople	
25		**TalkAboutSthingFunny**,QuestionDoYouWantPoem	
31	the person gets too close	**TalkAboutTooClose**, QuestionDoYouWantPoem	
36	the person answers Yes	**QuestionDoYouWantPoem**	
45		**TalkAboutPoem**,QuestionDoYouLikeMe	
56	the person answers Yes	**QuestionDoYouLikeMe**,TalkAboutSthingFunny	
63		**TalkAboutPoem**,IncreaseArousal,IncreaseValence, TalkAboutSthingFunny	
73		**IncreaseArousal**, IncreaseValence, TalkAboutSthingFunny TalkAboutSthingFunny	
74			Surprise 50%
76	the person answers Yes	**QuestionDoYouLikeMe**, IncreaseValence, TTalkAboutSthingFunny	
82			Surprise 25%
86		**TalkAboutPoem**, QuestionDoYouStillWantPoem,IncreaseArousal, IncreaseValence,TalkAboutSthingFunny	
90			Surprise 12%
96	the person is moving, the robot reprimands him	**TalkAboutMoving**,QuestionDoYouStillWantPoem,IncreaseArousal, IncreaseValence,TalkAboutSthingFunny	Anger 70%
100	robot tries to change topic	**QuestionDoYouStillWantPoem**,IncreaseValence, TalkAboutSthingFunny	
104			Anger 35%
105	the person answers Yes	**QuestionDoYouWantJoke**, IncreaseArousal, IncreaseValence, TalkAboutSthingFunny	
112		**TalkAboutJoke**,IncreaseArousal, IncreaseValence, TalkAboutSthingFunny	Anger 17%
119			Happiness 50%
121		**TalkAboutSthingFunny**,QuestionDoYouLikeMe, IncreaseArousal	
126		**TalkAboutJoke**, IncreaseArousal, IncreaseValence, QuestionDoYouLikeMe	
127			Happiness 25%
132			Happiness 75%
133			Happiness 100%
134		**TalkAboutSthingFunny**, IncreaseArousal,QuestionDoYouLikeMe	
140		**TalkAboutJoke**, IncreaseArousal,QuestionDoYouLikeMe	
147	the person answers Yes	**QuestionDoYouStillWantJoke**,IncreaseArousal, QuestionDoYouLikeMe,TalkAboutSthingFunny	
154			Happiness 81%
155	The person answers Yes	**TalkAboutJoke**,IncreaseArousal, IncreaseValence, QuestionDoYouLikeMe,TalkAboutSthingFunny	
163			Happiness 90%

Table 8.2: Example interaction: one person interacting with the robot. Behaviours in bold were the ones executed by the robot.

8.4 Interviews. What to Evaluate?

Many researchers tend to think that the best evaluation for a social robot is the effect that it has on people. A typical measure they resort to is interaction time: the larger the interaction sessions the better the robot. Also, people are interviewed about aspects like engagement or entertainment in their interactions with the robot. Most of robot evaluations emphasize the effect of the robot social abilities on humans. This suggests that the main goal of the robot was solely to produce a good effect on people (or at least that is what it was being measured).

Interaction distances are also commonly viewed as a central measure. Generally, it is assumed that the shorter the interaction distances the better the appeal of the robot.

Time (s)	Executed behaviour	Global mem.	Individual's mem.
0	robot start		
15	Greet		Greeted[1]
21	TalkAboutSthingFunny	TalkedAboutSthingFunny[1]	
28	QuestionDoYouWantPoem		QuestionedDoYouWantPoem[1], QuestionedDoYouStillWantPoem[1]
31		TalkedAboutSthingFunny[2]	
47	TalkAboutPoem		TalkedAboutPoem[1]
59	TalkAboutSthingToAttractPeople	TalkedAboutSthingToAttractPeople[1]	
60			TalkedAboutPoem[2]
64	TalkAboutOwner	TalkedAboutOwner[1]	
68	LookAtOther	LookedAtOther[1]	
71		TalkedAboutOwner[2]	
72	TalkAboutPoem		TalkedAboutPoem[1]
73		TalkedAboutSthingToAttractPeople[2]	
74		LookedAtOther[2]	

[1] inserted into memory
[2] forgotten

Table 8.3: Example interaction: one person interacting with the robot plus the owner.

While being important, it is our view that such engagement aspects should not be the central part of the evaluation. That approach would be the equivalent of establishing if a film has been well shot exclusively on the basis of a group of spectator's opinions.

It is extremely difficult, if not impossible, to isolate aspects like engagement or interaction time from population values like age, profession, observer's mood, level of knowledge, etc. Quoting Brooks: *Intelligence is in the eye of the observer* ([Brooks *et al.*, 1998]). Experiments described in [Schulte *et al.*, 1999] showed that children under 10 attributed intelligence to a museum robot (Minerva, see Section 6.1), whereas children aged 11+ attributed the robot the intelligence of a dog or a monkey. This result is in accordance with the effect of the Paro robot (see above): the group of interviewed subjects aged under 20 and over 50 rated the robot higher than the rest of subjects. More recent studies have also confirmed that the appearance of the robot is interpreted differently by adults and children [Woods *et al.*, 2005].

The study of [Siino and Hinds, 2005] analyses the effect of an autonomous mobile robot on a Hospital staff. Engineers and male administrators generally saw it as a machine that they could control. Female administrators and low-level female staff workers anthropomorphized it as a human male that acted with agency. Also, in a recent experimental work ([Lee and Kiesler, 2005]) it is shown how people estimate the robot's knowledge by extrapolating from their own knowledge. The authors suggest that designers of humanoid robots must attend not only to the social cues that the robots emit but also to the information people use to infer aspects of the robot's behaviour.

The fact is that even a simple chess program could lead to very large interaction times

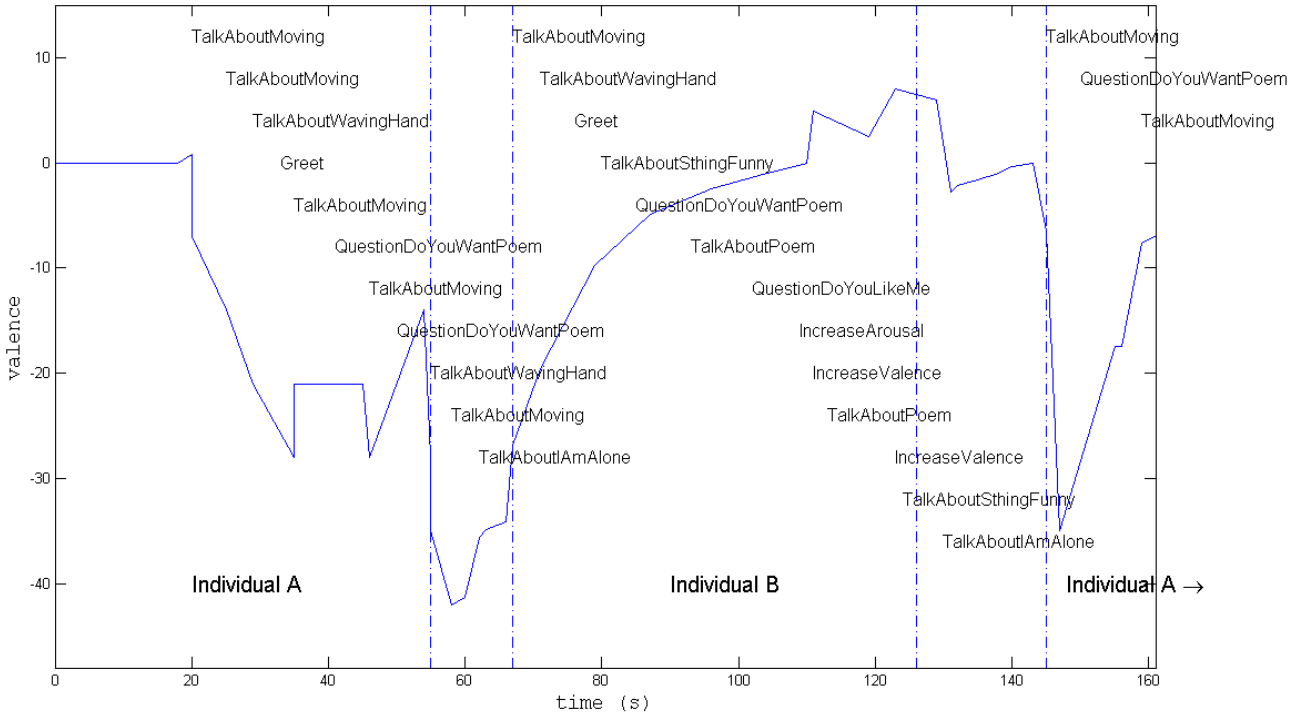


Figure 8.3: Example interaction that shows how the robot recognizes people. The figure shows on a time scale the valence values of the robot emotional state and the executed actions.

depending on the person. The task that the robot accomplishes may also influence people. With respect to interaction distances similar obstacles appear. In [Walters *et al.*, 2005*b*] it is shown that interaction distances between a mobile robot and humans, while being essentially similar to those encountered in human-human interactions, depend on the subject's personalities. Differences in typical interaction distances also appear when the comparison is made between adults and children, from [Walters *et al.*, 2005*a*]: *"It seems that children in particular are overall more accepting and approving of robots than the adult subjects studied. The social distance results so far have indicated that a substantial minority of adults (40% in our adult sample) do not seem to perceive the PeopleBot robot as a social being at first encounter"*.

Our goal in this work is not so much to produce robots that entertain people, but to imitate basic human social abilities. In other words, we pursue to measure not what the robot transmits (or the effect of what it transmits), but its ability to transmit. We acknowledge that these aspects are intertwined, which makes it hard to measure them separately.

On the other hand, there is an additional factor that is often neglected. The reader

Executed behaviour	**Robot says...**	**Individual says...**
TalkAboutMoving	*Keep still!*	
TalkAboutMoving	*Please don't move*	
TalkAboutWavingHand	*What are you doing?*	
Greet	Hi!	
TalkAboutMoving	*stop moving around!*	
QuestionDoYouWantPoem	*Answer with your head, do you want to hear poems?*	
TalkAboutMoving	*Please stop moving around!*	
QuestionDoYouWantPoem	*Please answer with your head, do you want to hear some poems?*	
TalkAboutWavingHand	*What are you doing now?*	
TalkAboutMoving	*For heaven's sake! Keep still!*	
TalkAboutIAmAlone	*They leave me alone again*	
TalkAboutMoving	*Keep still!*	
TalkAboutWavingHand	*What?*	
Greet	*Greetings earthling!*	
TalkAboutSthingFunny	*[funny comment]*	
QuestionDoYouWantPoem	*Answer with your head: do you wanna hear poems?* *Fine!*	[Head nod]
TalkAboutPoem	*[short poem]*	
QuestionDoYouLikeMe	*Answer with your head, do you like me?* *Ok*	[Head nod]
TalkAboutPoem	*[another poem]*	
TalkAboutSthingFunny	*[funny comment]*	
TalkAboutIAmAlone	*Now I feel alone*	
TalkAboutMoving	*Fuck! Keep still!*	
TalkAboutIAmAlone	*I know some poems, do you want to hear them?*	
TalkAboutMoving	*I will ignore you until you stop moving!*	

Table 8.4: Sentences spoken by the robot in the session of Figure 8.3, in chronological order.

of this document is expected to be knowledgeable on the topics of robotics and artificial intelligence. If we want to provide the reader with a faithful view of the robot then a good option would be to reproduce the impressions of other knowledgeable people.

For all these reasons, all the subjects of our interviews have a background in computer science. 37.5% of the interviewees had a PhD degree in computer science or engineering. Of these, 84% currently have robotics and/or computer vision as their main field of research. The rest have a MSc degree in computer science and are currently active. They all had some previous familiarity with the robot, and some of them even contributed with some software created for other projects. For those who still knew little about the robot, a behaviour was implemented so that the robot could describe its technical design (see Section 7.2).

8.5 Questionnaire

The interaction structure for all the subjects was the following. When the subject is in the robot's surroundings, the robot is switched on. Then the interaction develops. At some point the controller (the person who switched the robot on) enters the interaction space. From that moment on, two individuals are in the robot's interaction space. The interaction session ends when the controller sees that the robot is becoming bored (which tends to occur when the robot has nothing more to say to the person). During interaction, the controller tries to get the robot's attention at least once.

After the interaction session the subjects completed a questionnaire. The questionnaire is divided into three sections plus an additional question. The three sections are:

1. "I understand the robot"

2. "the robot understands me"

3. "overall impression"

The first two sections would allow to infer to what extent the robot has minimum two-way social abilities. The questionnaire is shown in Table 8.5.

The assignation of questions to sections is somewhat fuzzy. In principle, almost all the questions of the first section could also belong to the second section and vice versa. The question "The robot pays attention to you", for example, may be interpreted as communication from the subject to the robot (i.e. the robot perceives the subject) or from the robot to the subject (i.e. the subject perceives that the robot perceives him/her). The final assignation was based on the agent (human or robot) that initiates communication and has a clear intent to communicate. In "The robot pays attention to you", it is the robot that makes people see that it is paying attention to them (by speaking at them, for example). That is, the crucial aspect in that case is the fact that I see that the robot is paying attention to me.

Note that we do not ask the subjects for issues like engagement or fun. Again, we are mainly interested in endowing the robot with basic interaction abilities, and so words like convey and understand are frequent in the questions. Future work may orient these abilities toward a more specific application.

Section 1:
1. I have understood everything the robot has told me
2. The robot has conveyed its emotions through facial expressions
3. The robot has conveyed its emotions through voice tone
4. The robot pays attention to you
5. The robot is conscious of my presence
6. The robot is conscious of my movements
7. The robot recognizes people
8. This robot is a good starting point for keeping you informed
Section 2:
9. The robot understands what people say
10. The robot understands facial expressions
11. The robot knows where I direct my attention
12. This robot may be used to make it learn new things from people
Section 3:
13. I have not had to make an effort to adapt myself to the robot
14. The robot has not had failures (things that it obviously had to do but it didn't)
15. What do you think the robot should have to be used more frequently?

Table 8.5: Translation of the questionnaire used to evaluate the robot. The interviewees had to give between 1 and 5 points for each question (1 means no or totally disagree, 5 means yes or totally agree. The last question allowed a free answer).

8.6 Results and Discussion

A total of 19 individuals completed the questionnaire. This was considered sufficient, for, as can be seen in Figure 8.4, the variation of the scores as a function of the number of individuals is approaching zero.

Next, we make a detailed analysis of the results obtained for the questions in each section of the questionnaire:

Section 1

The first question obtained relative high values (see Figure 8.5), confirming that the robot's speech was easily understood. Note that this positive outcome was obtained after extensive use of annotations [1] in the text to synthesize (See Appendix A). The use of textual information without further adaptation (information gathered on the Internet, for example), would

[1] special characters that allow to change intonation, speed, volume and other speech parameters

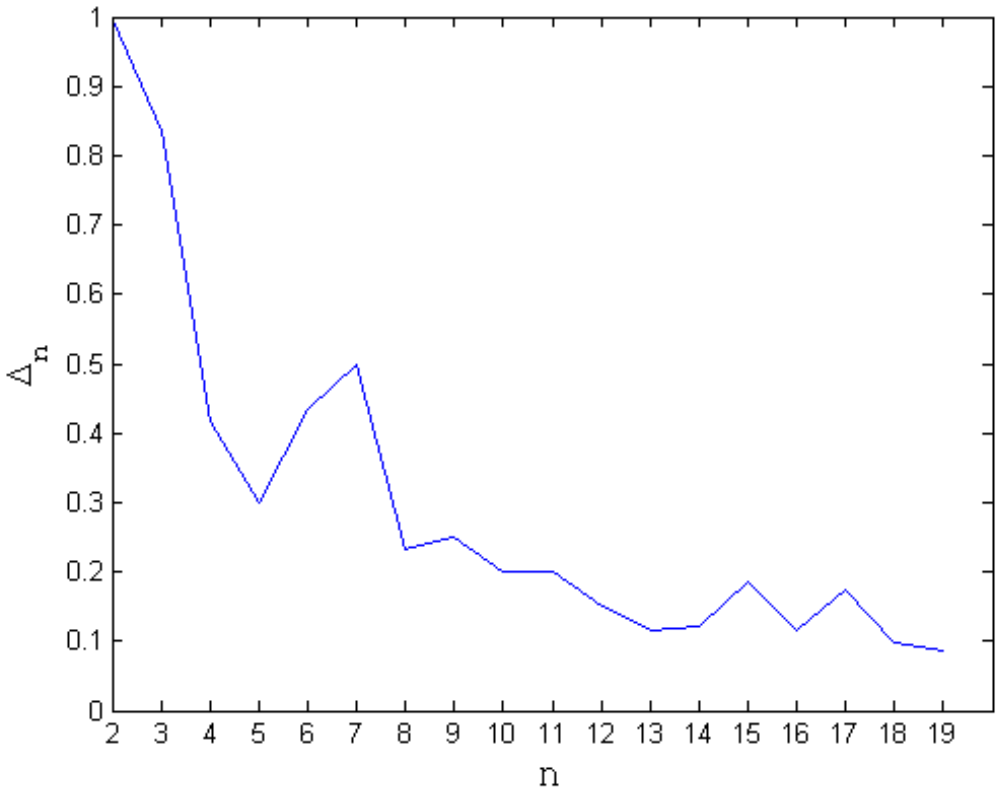


Figure 8.4: Maximum difference of means of n individuals with respect to the means of n-1 individuals. That is, if the mean score of n individuals for question j is represented as $m_{n,j}$, then $\Delta_n = \max_j |m_{n,j} - m_{n-1,j}|$, for $1 \leq j \leq 14$.

have made the robot's speech hard to understand.

The third question did not receive high scores. We believe this can be due to the fact that it is difficult for people to realize the fact that the robot changes its voice tone. That is, people is not so concentrated so as to realize those subtle differences.

Questions 4,5 and 6 received relatively high scores. This shows that the robot is perfectly able to pay attention to people, which is paramount in social interaction.

Subjects' impressions seem inconclusive with respect to the robot's ability to recognize people. This becomes apparent in the high variance of the scores of question 7 (i.e. the large confidence interval). In any case, we want to emphasize that many subjects were very impressed when the robot recognized the owner. In such cases, they frowned and immediately started to enquire the owner with expressions like *"how can the robot do that?"*.

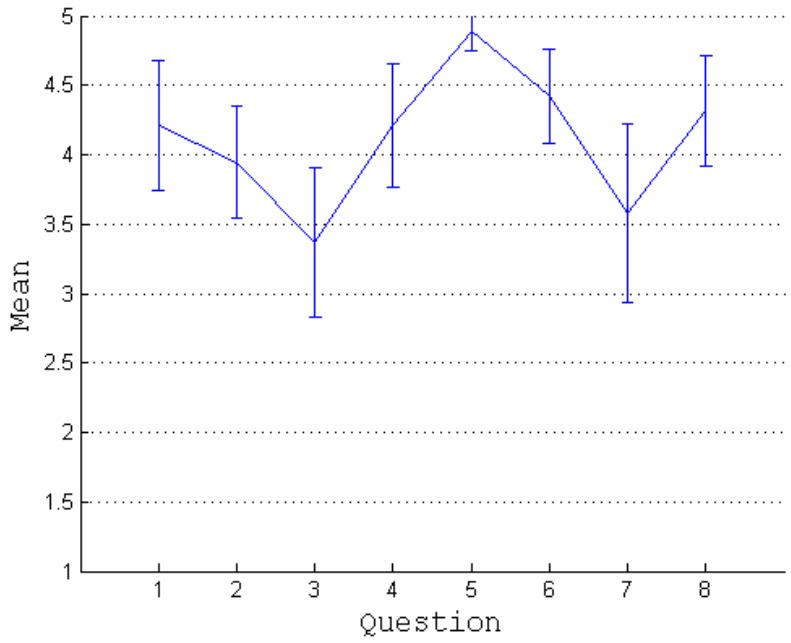


Figure 8.5: Mean score and 95% confidence interval obtained for each question in section 1 of the questionnaire.

Section 2

Note that section 1 had higher scores than section 2. Figure 8.6 shows the mean values for scores of each section. This was expected, since the robot currently has a wide range of outputs but few inputs from people.

In particular, question number 9 (see Figure 8.7) received slightly low scores. This is not surprising, since the robot does not have speech recognition capabilities. The robot itself has a behaviour that makes it say that he can not understand speech. This behaviour executes when the robot hears something coming from the zone in front of him. Also, the use of head gestures to answer robot questions can make people think that the robot does not have speech recognition.

Section 3

Question 13 (see Figure 8.8) received relative large scores. This is encouraging, since one of the main design objectives was to make the interaction as natural as possible (and this is one of the reasons why we soon discarded the use of microphones for speech recognition). Question 14 received average scores. This is less encouraging, although it is actually the

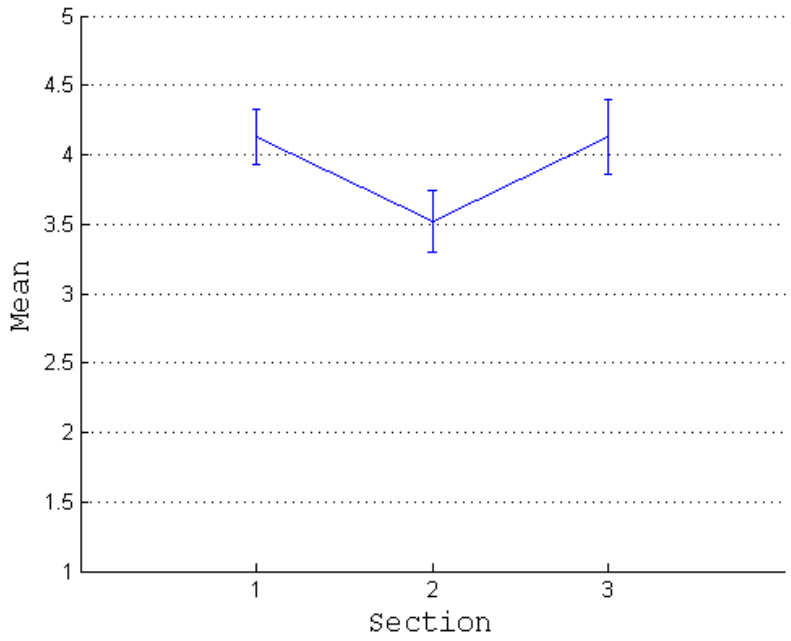


Figure 8.6: Mean score and 95% confidence interval obtained in each section (only the first two questions were included in the calculation of section 3).

combination of questions 13 and 14 what matters.

The answers to the last question were in general (6 out of the 19 interviewees) oriented toward the lack of speech recognition abilities. That is, subjects wanted to have a means of telling things to the robot. Currently, the subjects are forced in the interaction to be more passive than the robot. The next chapter outlines some ingenious ideas that deal with this lack of flexible input from people.

Some individuals also pointed out that the robot should have a definite task, like reading email or news. This may indicate that they consider the robot able to transmit such information and be of useful value in an office setting or for demonstration purposes, for example. Selecting an application is just a matter of designer preference or intended use for the robot.

A number of individuals liked very much the presence of humorous aspects in the robot behaviour and looks. Humour can alleviate the effect of a lack of functional speech recognition ability, which could make the robot appear unsociable or irritating to humans [Binsted, 1995]. Humour is becoming widespread in human-computer interaction, as computer systems tend to appear more human and emotive, the Microsoft's Office Assistant

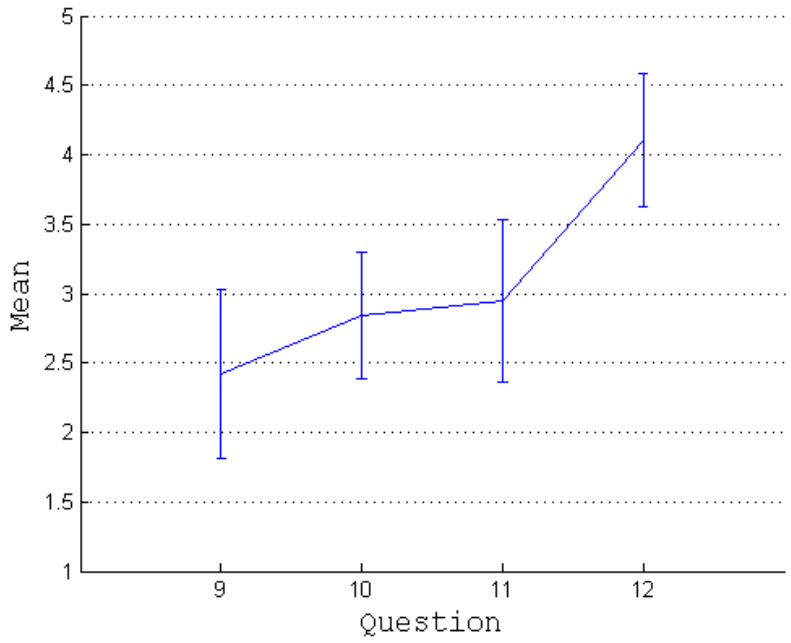


Figure 8.7: Mean score and 95% confidence interval obtained for each question in section 2 of the questionnaire.

being a clear example. Another example is eLOL (Electronic Laugh Out Loud), an Internet/Desktop application that delivers jokes to you daily based on your feedback.

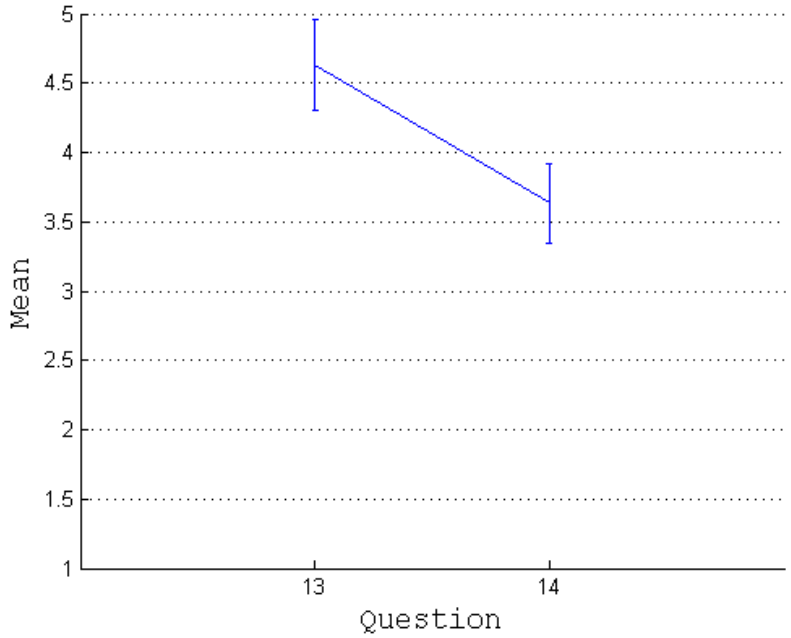


Figure 8.8: Mean score and 95% confidence interval obtained for each question in section 3 of the questionnaire.

Chapter 9

Conclusions and Future Work

"We must dare to think "unthinkable" thoughts. We must
learn to explore all the options and possibilities
that confront us in a complex and rapidly changing world.
We must learn to welcome and not to fear the
voices of dissent. We must dare to think about
"unthinkable things" because when things become
unthinkable, thinking stops and action becomes mindless."

James W. Fulbright

The emergent field of social robotics aims at building robots that have abilities to interact with people. These robots have expressive power (i.e. they all have an expressive face, voice, etc.) as well as abilities to locate, pay attention to, and address people. In humans, these abilities fall into the domain of what has been called "social intelligence". It is now agreed that such abilities are a fundamental part of human intelligence, traditionally associated to more "numerical" and logical abilities. Some authors even argue that social intelligence is a prerequisite for developing other types of intelligence.

For that class of robots, the dominant design approach has been that of following models taken from human sciences like developmental psychology, ethology and even neurophysiology. This is a perfectly valid approach, since the human being is the perfect model of social intelligence (at least the social intelligence that we would like to emulate).

This final chapter begins by summarizing in Section 9.1 the main conceptual contributions of this thesis. Then, the technological achievements are enumerated. Finally, Section 9.3 is devoted to ideas for future work and improvement.

9.1 Conceptual Contributions

The conceptual contributions of this work can be summarized as following:

- It has been formally shown that the reproduction of social intelligence, as opposed to other types of human abilities, may lead to fragile performance, in the sense of having very different performances between tested and unseen cases (overfitting). This limitation stems from the fact that the abilities of the social spectrum are mainly unconscious to us. Conscious mental processes fade into the unconscious with practice and habituation, which allows them to be fast and automatic. Also, these processes do not interfere (can be carried out in parallel) with other processes. Our social abilities, which appear earlier in life, are mostly unconscious. This is in contrast with other human tasks that we can carry out with conscious effort, and for which we can easily conceive algorithms.

- Such ideas may constitute a consistent explanation of the well-known fact that certain tasks that are trivial for us are hard for computers/robots and vice versa. As far as the author knows, the origin of this truism has not yet been explained clearly. Here we have seen that, in humans, practice and habituation makes processing details go unconscious. The lack of conscious knowledge of the form of our most "practised" processing algorithms (precisely those which we use with little cognitive effort and in parallel with other tasks) leads to fragile performance in the implementations. On the contrary, for tasks that we carry out with conscious effort we can devise more definite algorithms, which leads to implementations that may outperform humans in terms of robustness, speed and precision.

- It was noted that the robot design and development process is essentially inductive. This allowed us to propose an approach suitable for the specificity of the domain mentioned in the first point above: the complexity penalization idea often found in inductive Machine Learning techniques. Complexity penalization is a principled means to control overfitting. The following guidelines are proposed:

  - Make an effort in discovering **opportunities** for improving performance in the niche of the robot (realized niche). This can lead to unintuitive implementations (or, in other words, can call for originality).
  - Proceed **from simple to complex** algorithms and representations. The final implementation should be **as simple as possible**.

- Perform **extensive testing** of the algorithms and representations **in the niche** of the robot. Adjustments to the (or selection of new) algorithms and representations should be guided by the the results of these tests.
- Treat available **human knowledge very cautiously**, and always following the two previous guidelines. Basic knowledge will almost always be applicable to most niches. Detailed knowledge may be appropriate only for few specific niches.

This approach basically advocates for a simple-to-complex development process, with emphasis on extensive tests in, and fit to, the robot niche. Note that this can be achieved either from a design (as in CASIMIRO) or developmental perspective. Using genetics terminology, such adaptation to the robot niche can be achieved through ontogenetical or phylogenetical approaches. In the former, the robot itself would acquire novel abilities through strong interaction with its environment and/or caregivers, which is essentially what humans do.

Also, the guidelines of the proposed approach suggest that the integration of state-of-the-art "out-of-the-box" techniques to which these robots lend themselves to is not always the best option. A number of examples of this have appeared throughout this document.

From the point of view of the author, the novel conceptual approach developed in the first chapters of this document (and summarized above) is a novel and valuable contribution to the field. Currently, most social robots are designed with an eminently biologically-inspired approach, in which researchers make use of concepts and models taken from developmental psychology, ethology, neuroscience, etc. Starting from such analytical viewpoint and resorting to theories about autism and human consciousness we have come to the conclusion that engineering aspects such as the robot niche and *in situ* tests may be crucial for developing sociable robots.

9.2 Technical Contributions

A robot named CASIMIRO has been developed as an experimental testbed for exploring the methodological guidelines outlined above. A set of basic abilities, all of them directly or indirectly related with what we know as social intelligence, were separately studied and implemented:

- Sound localization: Sound localization systems normally extract localization cues from the sound signals. From these cues the system can infer, for example, if the sound comes from any of three basic directions: in front, from the left or from the right [1]. We proposed a method for extracting cues that achieves a higher discriminant ratio (using these three directions) than a previous method (Cog's sound localization system, see Section 1.1). For our robot, we concluded that a simple interaural time difference detector would be the best option, as it extracts a sound angle directly from a formula, avoiding the need to use thresholds or classifiers to infer directions from the extracted cues.

- Omnidirectional vision: Many sociable robots use conventional cameras for detecting people. These cameras have a rather narrow field of view (typically less than 70^o). For CASIMIRO, an omnidirectional vision system, based on background subtraction, was built with low cost components. Despite (or because of) its simplicity, the system gives the robot a 180^o field of view in which it can detect and track people robustly. The system uses a convex mirror, which allows to provide a rough measure of distance to the detected people. With such measure we have made the robot able to detect individuals that are too close, or otherwise act to maintain the appropriate interaction distance.

- Audio-visual attention: an audio and video based attention system was implemented, allowing the robot to put its attention to the most important element of the environment: people. Inaccuracies in the perception channels -particularly in sound localization- are alleviated by the bimodal fusion. Previous attentional systems concentrated on the effect of salient low-level perceptions on attention. Our attentional system allows for saliency contributions from any deliberative processes, which in turn provides a flexible way to implement joint attention and to model the effect that emotions may exert on attention.

- Face detection: Our first implementations used state-of-the-art face detection algorithms. However, after extensive experimental evaluation we discovered that, in the current robot niche, only skin-colour blobs with certain height/width ratios would suffice to detect faces robustly. This served as a confirmation of the the postulates of the approach outlined above. On the other hand, most face detection systems use colour as the main discriminant feature. But often, furniture have colours that appear similar to skin colour. To alleviate this, depth was used as a filter, producing better detection rates as well as better framed faces.

[1] these directions could be used to move the robot head and make the localization more and more precise.

- Head yes/no gesture detection: Yes/no gestures are detected with a simple facial points tracking-based technique. Only distinctive points inside the face rectangle are tracked. These yes/no, approval/disapproval gestures are recognized as answers to questions made by the robot. This provides a robust input that can be used as a basic feedback mechanism for learning new abilities or improving the social behaviour.

- Habituation: Some perceptual inputs should not waste robot resources, especially its attention. An histogram-based technique was implemented to detect repetitive perceptual patterns. The technique was applied in the visual domain, allowing the robot to stop looking at repetitive ("boring") visual patterns. This goes in the direction of making the robot appear more natural, as observed by humans.

- Emotion: An emotional system was implemented so that the robot is every time in an arousal-valence state. Such state, which can be varied by other robot modules, is mapped onto facial expressions like happiness, anger, neutral, etc. Emotional states decay, so that the robot will tend to have a neutral state and expression unless properly stimulated.

- Facial motion: A conceptually simple and scalable method for facial features motion modelling was implemented in the robot. The proposed method and software allow the designer to easily conform facial expressions by moving interactively one or a group of motors. Transitions between expressions can also be tested and modified easily.

- Owner detection: A method for recognizing the robot's owner was implemented. Face recognition researchers tend to measure performance in terms of the number of individuals that the system can recognize and *measured* error rate. In this respect, a *measured* error rate of 5-10% can be considered very good. The approach presented in this document does not use face recognition techniques, although it recognizes a single individual with *guaranteed* zero error. No face recognition method would recognize the owner with such low error. Note that this result is an example of what can be obtained by following the general approach mentioned above: although it is only appropriate for the current robot niche, we have been able to devise the simplest algorithm (or one of the simplest) that allows to recognize the owner. In experiments, people received such ability with amazement.

- Memory and person recognition: Person recognition was achieved through colour histogram-based torso recognition. For reduced groups of people, such technique may be more robust than face recognition. If we consider owner detection, the robot is in fact able to recognize a small group of individuals (say 1 to 3) very robustly. Person

recognition allows the robot to have memory associated to each individual, particularly data about the outcome of previous interactions.

9.3 Future work

Conceptual aspects

Future work will include the further development of the implications of the fact that social abilities are mostly unconscious. We believe that this specificity of the problem deserves additional analysis. In particular, it would be very important to establish the extent to which introspection and indirect observation can make visible the inner workings of unconscious processes. We saw in Chapter 2 that the unconscious nature of certain processes is acquired through practice and habituation. Their being unconscious makes them fast and automatic, allowing them to be carried out in parallel with other tasks. Thus we have a being that learns, and the better he learns an algorithm (i.e. the better he performs), the less conscious access he has to that algorithm. What we see is that the higher mental processes that precisely allow us to tackle the reproduction of our social abilities crucially depend on those unconscious processes being unconscious. This catch-22 situation seems a theoretical obstacle that should be explored.

In any case, such reflections are useful mainly on a theoretical level and in fact they would fall into what would be Cognitive Science. As it has been shown in the first chapters of this work, they may serve to define the specificities of the problem of developing sociable robots. They are also thought-provoking ideas that could lead to new interesting approaches to the problem.

On a more practical level, the utility of the specific guidelines proposed in this work for developing sociable robots shall have to be explored in future projects. Of course, it is not intended to be a definitive approach for developing sociable robots. It proposes just a few non-stringent rules, and hence it may be useful as a basic development perspective, possibly in conjunction with other paradigms and approaches. We hope that it will be present in our future robots and in those built by other researchers. Only a large scale use of the approach will allow to extract meaningful conclusions.

Technical aspects

Many aspects of the robot could be considered for further improvement. In fact, some parts are currently being extensively researched separately: sound localization, emotion modelling, developmental approaches, etc. From our point of view, however, there are two key abilities that shall be seriously taken into account in our future work.

First, an aspect that should be further explored in the future is person recognition. This is no doubt crucial for a sociable robot. The ideal technique would be facial recognition, not least because it is unobtrusive. Generally, it is thought that the best way to test face recognition algorithms is by showing their performance for large numbers of individuals. In our view, it would be better if we could guarantee a good performance at least for a low number of individuals. Thus, for person recognition other techniques may be useful in the case of having to recognize between a low number of individuals, say a family. Height, for example, may be used to recognize.

Note that the use of such simple feature for recognition (and the technique mentioned for owner recognition) is in accordance with the general approach outlined above. Complexity penalization techniques are mostly applied in the classification part of the system. Feature spaces should also be as simple as possible. For that purpose, constructive induction (also known as "feature engineering") may be used.

Recognizing people and maintaining some sort of memory of the individuals is so important for interaction that more original techniques should be explored. Some authors have used wireless tags (RFID tags, see below) for identification. These tags, in the form of badges, emit a radio signal that a receiver on the robot can detect. The system can provide both position and identification of each badge in the vicinity. This approach is very attractive, although individuals that enter the room have to put on the badges or they will not be detected. Obviously, the utility of this is subjective. In the design of CASIMIRO it has always been our desire to use techniques that are not unpleasant for people. Therefore, the use of badges would not be an option for us (unless the robot worked in an environment where badges were compulsory).

Notwithstanding, RFID (Radio Frequency Identification) tags are becoming increasingly small. The so-called passive tags do not have internal power supply. The minute electrical current induced in the antenna by the incoming radio frequency signal provides just enough power for the CMOS integrated circuit in the tag to power up and transmit a response. The lack of an onboard power supply means that the device can be quite small. As of 2005, the smallest such devices commercially available measured 0.16mm^2, and are

thinner than a sheet of paper. Still, current passive tags have practical read distances ranging from about 10mm up to about 1 metre.

In future experiments we will consider the following scenario. People to be recognized would carry passive tags, in the form of small stickers. They could be adhered to wrist watches or mobiles phones. In the main entrance to the interaction space (the robot room) receiver panels would be placed that would detect each individual tag. A person tracking system (either using omnidirectional vision as in CASIMIRO, laser ranges finders or other techniques) would then track the recognized person. That way, each person would be identified and its identity maintained throughout the whole interaction session.

On the other hand, speech recognition is a necessary task that will also be studied in the future. The results of the evaluation questionnaire clearly shows that people still find that the robot does not understand them. This would warrant efforts in this aspect.

It is estimated that speech recognition errors, dubbed by Oviatt as the Achilles' heel of speech technology, increase a 20%-50% when speech is delivered during natural spontaneous interaction, by diverse speakers or in a natural field environment [Oviatt, 2000]. Word-error rates depend on the speaking style, emotional state of the speaker, etc. Moreover, in the presence of noise, speakers have an automatic normalization response called the "Lombard effect", that causes speech modifications in volume, rate, articulation and pitch. In CASIMIRO, experiments were carried out using microphones placed on the head. The option of making the speaker wear a microphone was discarded from the beginning because it is too unnatural. The main obstacle in this case was the distance between the speaker and the microphones of the robot. Speech recognition errors increase significantly with the distance to the speaker (a 35dB decrease in SNR at 60cm [Wenger, 2003]). This is specially true in an office environment with noise generated by computer fans, air conditioning, etc. An hypercardioid microphone was also tested though recognition of simple words only performed relatively well enough at distances of up to around 50cm.

Speech recognition at a distance is probably one of the most interesting research topics in human-computer interaction. The usefulness is clear, not least because it would allow users to be free of body-worn microphones. Two techniques seem especially attractive. On the one hand, microphone arrays can be used to selectively filter out signals coming from certain directions [McCowan, 2001]. This is equivalent to say that only signals coming from a cone in front of the microphones are considered. A simple microphone array was implemented for CASIMIRO with the two microphones used for sound localization, by summing the two signals. This emphasizes signals coming from the zone in front of the microphones. However, the signal to noise ratio gain was low. For two micro-

phones, the maximum theoretical gain in SNR terms is 3dB [Johnson and Dudgeon, 1993, Ortega-García and González-Rodríguez, 1996, Seltzer, 2001]. This is even lower than the hypercardioid microphone mentioned above.

An interesting technique that we shall take into account is audio-visual speech recognition [Liu *et al.*, 2002, Liang *et al.*, 2002, Nefian *et al.*, 2002]. This technique is based on combining audio information with lip tracking to make the system more robust (an example of the redundancy principle mentioned in Section 3.4). Results are impressive, achieving error reductions of up to 55% (at SNR of 0dB) with respect to audio only speech recognition. However, it is not clear whether the system can work in real-time without special hardware. Besides, the system requires the face to be well framed and stable in the images.

Even if we cannot implement an acceptable hands-free speech recognition system for CASIMIRO, the use of certain microphones will be considered. Currently, there are commercially available Bluetooth headsets that are both discreet and easy to carry. The robot owner could wear one of those headsets. The robot would easily recognize his/her speech (the microphone is very close to the mouth and the recognition could be in speaker-dependent mode). When a new person enters the room, the robot would adopt an appropriate attitude. That is, it would not speak directly to the newcomer. Displaying shyness, it would only talk with the owner. Should the newcomer want to tell something to the robot, the owner would act as a sort of proxy, by saying things like *"CASIMIRO, he is telling that..."* or *"He asks..."*.

Appendix A

CASIMIRO's Phrase File

In this appendix we show an extract of CASIMIRO's phrases file. The file is arranged in symbols, under which a collection of phrases can follow. The phrases of each symbol can have a priority, which is a non-negative number that appears before the phrase (lower number=higher priority). See Section 6.3 for more details.

```
; Fichero con la base de frases a pronunciar
;
; Formato:
;
; <Simbolo>
; <TAB>[Prioridad+espacio]Frase
; <TAB>[Prioridad+espacio]Frase
; ...
; <Simbolo>
; <TAB>[Prioridad+espacio]Frase
; <TAB>[Prioridad+espacio]Frase
; ...
;
; La prioridad es opcional, es un numero mayor o igual a cero. Si se usan prioridades, han de
 aparecer consecutivas
; (no ha de faltar una prioridad).
; Se dicen primero las frases de valor mas bajo de prioridad. Dentro de el conjunto de frases
de mismo valor de prioridad
; se escoge cada vez una al azar.


;;;;;;;;;;;;;;;;;;;;;;;;
;Actualidad
; Rajoy
; Zapatero
; Ibarretxe
; Blair
; Bush

;Rajoy
```

```
; Tengo una Rajoy en el aluminio!

;Zapatero
; Zapatero presidente!

;Ibarretxe
; Oscar dice que tengo más peligro quel plan Ibarreche

;Blair
; Soy amigo de Tony Bleeeerg

;Bisbal
; A ver si alguien me compra una peluca a lo Bisbal!

;Bush
; sé menos de la gente que trabaja aquí que Bush de geografía

;;;;;;;;;;;;;;;;;;;;;;;;


Greeted [CiclaFrases]
Hooola!
Qué tál?
‘4 Salúdoss ‘p3 terrícola
Hola!

TalkAboutTooManyPeopleAround [CiclaFrases]
Con tanta gente Estoy más perdido que Gually en el Frente Atlético
Hay mucha gente aquí, estoy más perdido que úna neurona en una película de Rámbo!
Cuánta gente!

TalkAboutSthingFunny [CiclaFrases]
aáayyy!. estoy mas ‘4 desgastado que las ruedas del coche fantástico!
‘1 je! ‘p100 Soy mas vago que los supléntes de un futbolín
Tengo menos ‘4 fallos que el examen de conducir de Máikel Nait
Alguien tiene un chupachups de aluminio? jéje
a versi consigo hacerte reir
espero que te estés llevando una buena impresión de mii
tienes que decir a todo el mundo que soy un robot genial
has visto por ahí alguna robotita cachonda?
pórtate ‘4 bien para que ‘4 puedas ver ‘4 todo lo que hago
se supone que soy tecnología punta
Tengo ‘4 nervios de acero ‘p500 jéje

TalkAboutSthingToAttractPeople [CiclaFrases]
0 Acércate un poquitín, que soy más inofensivo que las balas del equipo AA!
0 Acércate un poquitín
0 Acércate hombre
0 aceércate un pooco
1 Acércate más coño!

TalkAboutILikeYou [CiclaFrases]
0 creo que me gustas mucho
0 Túy yo podemos ser grandes amigos
0 me pareces interesante
```

```
1 Me gustas mucho
1 me gustas!

TalkAboutIDontLikeYou [CiclaFrases]
0 no me lo estás poniendo fácil
0 que difícil eres!
0 pero que difícil eres!
0 Eres más aburrido que jugar al solitario de güindous
1 No me gustas!
1 No me gustas nada!
1 que muermazo eres!
1 '4 Borde!
2 si '4 no fuera solo una cabeza te daría una buena patada en el culo!
2 hijoputa

TalkAboutColdRoom
0 El maldito aire acondicionado otra vez a toda máquina
0 me '4 voy a congelar con '4 tanto aire acondicionado
0 que alguien desconecte el aire acondicionado!

TalkAboutHotRoom
0 Esto está más caliente quel mechero de Colombo
0 Hace más caló aquí que en el tubo de escape del coche fantástico
0 Hace un poco de calór aquí dentro
1 que alguien conecte el aire acondicionado!
1 hace mucho calor aquiiiii!

TalkAboutColdDay
Un día fresquito!
Parece que hace fresquito ahí fuera

TalkAboutHotDay
Vaya día de playa hace!
Vaya día de calor! y yo sin poder salir de aquí!
'0 Menos mal que no estoy fuera, que si nó se me derriten los circuitos

QuestionDoYouLikeMe [CiclaFrases]
0 contéstame con la cabeza: Te gusto?
1 Contésta sí 'p1 o nó con la cabeza: Te parezco divertido?
1 Contésta sí 'p1 o nó con la cabeza: Te parezco un robot interesante?
2 Te divierto?
2 Te estoy divirtiendo?

TalkAboutTooClose
0 estás muy cerca, aléjate un poco
0 aléjate un poco, que corra el aire
0 no te acerques tanto
0 aléjate un poco
1 joder! estás '4 muy '4 cerca!
1 cóño!, hazte para atrás!
1 hazte un poco para atrás!
2 Hasta que no te hagas un poco para atrás pasaré de tí!

TalkAboutMoving
0 quédate quieto en un sitio
```

```
0 por favor no te muevas
1 por favor Deja de moverte!
1 oye, deja de moverte!
2 pero coño! quédate quieto en un sitio!
2 pero coño! quédate quieto!
2 por favor Pára ya de moverte!
3 hasta que no dejes de moverte voy a pasar de tí!

TalkAboutSpeaking [CiclaFrases]
0 no tengo oídos, no puedo entender lo que me dicen
0 Nó 'p1 me han '4 puesto oídos, no entiendo lo que me dicen
0 si me '4 dicen algo no lo entenderé, aún no puedo reconocer la voz
0 no puedo entender, todavía no puedo reconocer la voz
1 sileencio!
1 a callar!
1 que se cállen!
2 '4 cállate coño!

TalkAboutWavingHand
0 qué haces?
0 qué chorradas haces?
0 pero qué chorradas estás haciendo?
0 qué boberías haces?
0 estás espantando moscas?
1 pero deja de hacer tonterías!
1 deja de hacer tonterías!
1 deja ya de hacer tonterías!
1 deja ya de hacer chorradas!
1 oye para ya de hacer chorradas!
2 voy a pasar de tus tonterías!
2 ya paso de tus chorradas!
3 gilipoyas

QuestionDoYouWantPoem [CiclaFrases]
0 Contésta sí 'p1 o nó con la cabeza: Te digo una poesía?
0 contéstame con la cabeza: quieres que te diga una poesía?
0 contesta con la cabeza: Quieres oír alguna poesía?
1 me sé algunas poesías, quieres oírlas?
1 quieres oír algunas poesías que me se?
1 conozco algunas poesías, quieres oír alguna?

QuestionDoYouStillWantPoem [CiclaFrases]
0 contesta con la cabeza: quieres que te diga más poesías?
0 quieres seguir oyendo poesías?
0 te digo más poesías?
0 te sigo diciendo poesías?

TalkAboutPoem
No quiero perlas del mar. ni perfumes de oriente. solo quiero que mi amor contigo, perdure eternamente
Si 'p5 yó ffueradios no te habría creado. Pues para tanta belleza mis ojos no están preparados.
el viento bbésael barco. el barco besa el mar. y yo quisiera ser brisa, para tus labios besar
Siempre en las noches oscuras. cuando nadie te quiere escuchar. en una estrella del cielo, un refugio
encontrarás. Siempre en los días más tristes.  donde no tienes donde ir. mira a los más humildes,
que no tienen un téchodonde vivir.
me gustas cuando callas porque estás como ausente. y me oyes de lejos, y mi voz no te toca. parece
```

que los ojos se te hubieran volado, y parece que un beso te cerrara la boca.
El carnaval del mundo engaña tangto. que las vidas son breves mascaradas. aquí aprendemos a reír con llanto, y también a llorar con carcajadas.
te dice la paloma: toma! toma!. 'p300 Te dice la corneja: deja! deja!. 'p300 Te dice la amapola: hola! hola!. Y tú criatura de dios, siempre diciendo adiós!
si tu cuerpo fuera cárcel y tus brazos '4 cadenas. qué bonito sittio para cumplir mi condena!
de dónde vienes? del cielo. a quién buscas? al dolor. qué traes? consuelo. cómo te llamas? amor
Anoche cuando dormía soñé. ¡bendita ilusión!. 'p150 que un ardiente sol lucía, déntro de mi corazón.
Era ardiente porque daba, calores de rojo hogar. y era sol porque alumbrába y porque hacía llorar.
Anoche soñé que oía a Dios, gritándome: ¡Alerta!. 'p200 Luego era Dios quien dormía, y yo gritaba: ¡Despierta!
Ay del que llega sediento a ver el agua corre, y dice: la sed que siento, no me la calma el beber.
Ayer soñé que veía a Dios y que a Dios hablaba; y soñé que Dios me oía. 'p100 Después soñé que '0 soñaba.
Bueno es saber que los vásos nos sirven para beber; lo malo es que '4 no sabemos para qué sirve la sed.
En preguntarlo que sabes el tiempo no has de perder. Y a '4 preguntas sin respuesta ¿quién te podrá responder?
Es el mejor de los buenos, quien sabe que en esta vida tódo es cuestión de medida: un poco '4 más, o algo '4 menos.
Moneda que está en la mano, quizá se deba guardar: la monedita del alma se pierdesi no se da.
Nuestras horas son minutos cuando esperamos saber, y siglos cuando sabemos lo que se puede aprender.
Los ojazos de mi robotita se parecen a mis males. Grandes como mis tormentos, negros como mis pesares.
Las horas que tiene el día. las he repartido así. nueve soñando contigo, y quince pensando en tí.
El corazón de una robotita, me dícenque tengo. yo no se para qué me sirve el corazónsinel cuerpo.
Procura no despertarme cuando me veas dormir. No sea que esté soñando y sueñe que soy feliz.
Cuando mires las estrellas acuérdate de mí, porque en cada una de ellas hay un beso mío para ti
El beso es una sed loca que no se apaga con beber. se apaga con otra boca que tenga la misma sed.
Soñé que el fuego helaba. soñé que la nieve ardía. y por soñar lo imposible soñé que me querías.
Es una cosa sabida. que úno más úno hacendós, 'p200 pero úna mujer y ún hombre, o son uno o nada son.
El que desgraciado nace, desde chiquito yaempieza. por más cuidadosque tenga, en lo más llano tropieza.

QuestionDoYouWantJoke [CiclaFrases]
0 Te digo unos chistes?
0 quieres que te diga unos chistes?
0 Quieres oír algún chiste?
1 me sé algunos chistes, quieres oírlos?
1 quieres oír algunos chistes que me se?
1 conozco varios chistes, quieres oír alguno?

QuestionDoYouStillWantJoke [CiclaFrases]
0 quieres que te diga más chistes?
0 quieres seguir oyendo chistes?
0 te digo más chistes?
0 te sigo diciendo chistes?

TalkAboutJoke
¿Cuál es el colmo de un chapista? 'p700 Tener una mujer que le de la lata.
Era un niño tan feo tan feo tan feo, que cuando lo iban a bautizar lo tiró el cura para arriba y dijo: Si vuela es un murciélago.
Era un niño tan feo tan feo, que cuando nació el médico le dióla torta a la madre

```
¿Cuál es el colmo de un calvo? ‘p700 Tener ideas descabelladas
¿Cuál es el colmo de un calvo? ‘p700 Encontrarun pelo en la sopa
¿Qué le dice un calvo a otro? ‘p700 Cuanto tiemposin vértel pelo.
¿Cuál es el colmo de un jorobado? ‘p700 Estudiar derecho.
qué le dice un fideo a otro? ‘p700 oó ye mi cuérpo ‘4 pi de ‘4 sál sa
¿Qué se pone Supermán cuando sale de la ducha? ‘p700 súper fume
Mamá mamá en la escuelame llaman Güindous 95. Tranquilo hijo mío, no les hagas caso pero haz
algo útil
¿Por ‘4 que las ‘4 películas de chaplin eran mudas? ‘p700 porque el director le decía:
No Charles Chaplin.
Cómo se dice beso en árabe? ‘p700 Saliva va saliva viene
¿Qué le dice una impresora a otra? ‘p700 oiga, esa hoja es ‘4 suya o es impresión mía?
¿Qué le dice el papel higiénico a la comida? ‘p700 te espero a la salida
¿Cuál es el colmo de un ciego? ‘p700 Enamorarse a primera vista
Esto es un autobús de bizcos por Madrid y el conductor le dice a los pasajeros: Si miran a la
derecha podrán ver a la izquierda el Museo del Prado
Chico cojo de pierna derecha, busca chica coja de pierna izquierda para dar un paseo
Va un jorobado por la calle y se le cae un paquete de Cámel al suelo, un señor lo recoge y
le dice: señor! se le há caídoel carnét!
gemelo suicida mata a su hermano por error.
cómo se dice diarrea en portugués? ‘p700  catarata ‘2 dutrasheéiro
cómo se dice diarrea en japonés? ‘p700  kagasagua
el médico le dice a la viejecita, señora tengo dos noticias, una buena y otra mala. La
mala es que sufre usted Alzéimer, y la buena es que se le olvidará en un rato.
qué le dice un jaguár a otro? ‘p700 jáguar yú?

QuestionDoYouWantRobotData [CiclaFrases]
Te cuento cosas sobre mí?
quieres que te cuente cosas sobre mí?
Quieres oír algunas cosas sobre mí?
quieres oír algunas cosas sobre mí?

QuestionDoYouStillWantRobotData [CiclaFrases]
0 quieres que te cuente más cosas sobre mí?
0 quieres que te siga diciendo cosas sobre mí?
0 te digo más cosas sobre mí?
0 te sigo contando cosas sobre mí?

TalkAboutRobotData [CiclaFrases]
0 Mi esqueleto fue diseñado por ingenieros mecánicos y construido en aluminio
0 Únos ingenieros mecánicos franceses diseñaron mi esqueleto, que luego fue construido en aluminio
1 uso una cámara con espejo redondeado para ver a mi alrededor
1 la cámara con espejo redondeado que hay en mi parte inferior me sirve para ver alrededor
2 tengo también una cámara estéreo cerca de la nariz, me permite tener sensación de profundidad
2 también tengo una cámara estéreo bajo la nariz. Con ella puedo tener sensación de profundidad
3 mi cara la mueven un total de 11 servomotores
3 11 servomotores mueven las facciones de mi cara
4 los servomotores se mueven para formar una expresión en mi cara que muestre mi estado de ánimo
4 puedo poner cara de felicidad, tristeza, enfado, sorpresa y sueño.
5 tengo dos pequeños micrófonos que me permiten saber de dónde viene el sonido
5 dos pequeños micrófonos me permiten saber de dónde viene el sonido
6 Soy capaz de detectar y seguir personas. Puedo fijarme en ellas
6 soy capaz de fijarme en las personas que pasan por delante
7 puedo mover el cuello a un ladoy a otro, pero nó agachar o levantar la cabeza, porque mi
cuello es aún bastante débil
```

```
8 me hán entrenado para que cuente cosas a las personas y esté atento a lo que hacen
8 mi objetivo es entretenerun poquito a las personas que pasan por aquí
9 mi cerebro son dos ordenadoresPeCé conectados entre sí
9 dos ordenadoresPeCé son mi cerebro
10 Oscar aún sigue trabajando en mí
10 Aún están trabajando en mí

TalkAboutOwner
0 mi creadór!, 'p10 jèje!
0 qué Óscar, lo hago bién?
1 estás presumiendo de mí?
1 Óscar, cuándo me traes una robotilla?
1 sí bwana?
2 a tí estoy cansado de vérte!
2 a tí estoy harto de verte!
3 tíio! apágame y déjame descansár!

TalkAboutIWantToSleep
bueno, creo que me voy ya a dormir
me voy a dormir, para conservar energía
me apetece dormir
estoy cansadito, me duermo
'4 puescreo que me voy a dormir un rato

TalkAboutIAmAlone
0 me dejan solito
0 todos se van. Nadie me quiere
0 otra '4 vez me quedo solito
0 ótra vez solito. acabaré quedándome dormido
0 pero nó se marchen queentonces me quedo dormido!
1 si me dejan solo me quedaré dormido
1 estando solo me abuurro
2 nó me gusta estar solo!
2 odio estar solo!
3 no quiero estar solo!
3 que alguien se me ponga delante!
4 es que nadie quiere estar conmigo?
4 por qué me dejan solito?

TalkAboutIAmFedUpWithYou
0 qué gente más aburrida!
0 estoy cansado de tí!
0 qué aburrimiento!
0 contigo me aburro!
1 vete! eres un aburrimiento!
1 oye date una vuelta por ahí
```

Bibliography

[Adams *et al.*, 2000] B. Adams *et al.*. Humanoid Robots: A New Kind of Tool. *IEEE Intelligent Systems*, vol. 15(4), 25–31, 2000. URL `citeseer.nj.nec.com/adams00humanoid.html`.

[Adini *et al.*, 1997] Y. Adini *et al.*. Face Recognition: The Problem of Compensating for Changes in Illumination Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), July 1997.

[Alexander, 1995] J. Alexander. Sound Localization in a Meeting Room, 1995. Available at `http://citeseer.nj.nec.com/136635.html`.

[Aloimonos *et al.*, 1987] J. Aloimonos *et al.*. Active Vision. *International Journal on Computer Vision*, pp. 333–356, 1987.

[Arsenio, 2004] A. Arsenio. *Cognitive-Developmental Learning for a Humanoid Robot: A Caregiver's Gift*. Ph.D. thesis, MIT Computer Science and Artificial Intelligence Laboratory, September 2004.

[Baars, 1988] B. J. Baars. *A cognitive theory of consciousness*. Cambridge University Press, NY, 1988.

[Baker and Nayar, 1996] S. Baker and S. Nayar. Pattern Rejection. In *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 544–549, June 1996.

[Ballard, 1991] D. Ballard. Animate Vision. *Artificial Intelligence*, vol. 48(1), 1–27, 1991.

[Bargh and Williams, 2006] J. A. Bargh and E. L. Williams. The Automaticity of Social Life. *Current Directions in Psychological Science*, vol. 15(1), 1–4, 2006.

[Baron-Cohen, 1995] S. Baron-Cohen. *Mindblindness*. MIT Press, 1995.

[Bartneck, 2003] C. Bartneck. Interacting with and embodied emotional Character. In *Procs. of the DPPI2003 Conference*. Pittsburgh, 2003.

[Beltrán-González, 2005] C. Beltrán-González. *Toward predictive robotics: The role of vision and prediction on the development of active systems*. Ph.D. thesis, LIRA-Lab, University of Genova, 2005.

[Berlyne, 1960] D. Berlyne. *Conflict, arousal and curiosity*. McGraw Hill, N.Y., 1960.

[Beymer, 1993] D. Beymer. Face Recognition Under Varying Pose. AIM-1461, p. 14, 1993. URL `citeseer.ist.psu.edu/beymer94face.html`.

[Binsted, 1995] K. Binsted. Using humour to make natural language interfaces more friendly, 1995. URL `citeseer.nj.nec.com/binsted95using.html`.

[B.J. Baars, 2004] B.J. Baars. Recovering consciousness: A timeline, 2004. URL `http://www.sci-con.org/news/articles/20020904.html`.

[Blackburn *et al.*, 2001] D. Blackburn *et al.*. Face Recognition Vendor Test 2002 Performance Metrics. Tech. Rep. Defense Advanced Research Projects Agency A269514, 2001.

[Blauert, 1983] J. Blauert. *Spatial hearing*. MIT press, Cambridge, MA, 1983.

[Blumberg, 1997] B. Blumberg. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. Ph.D. thesis, MIT Media Lab, 1997.

[Bouguet, 1999] J. Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker. Tech. rep., Intel Corporation, Microprocessor Research Labs, OpenCV documents, 1999.

[Boult *et al.*, 1999] T. Boult *et al.*. Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets. pp. 48–58, 1999.

[Breazeal and Scassellati, 1999] C. Breazeal and B. Scassellati. How to build robots that make friends and influence people. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1999. URL `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html`. Korea.

[Breazeal, 2002] C. L. Breazeal. *Designing social robots*. MIT Press, Cambridge, MA, 2002.

[Brooks, 1986] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, (2), 14–23, 1986.

[Brooks, 1991] R. Brooks. Intelligence without representation. *Artificial Intelligence*, (47), 139–160, 1991.

[Brooks and Stein, 1994] R. Brooks and L. Stein. Building Brains for Bodies. *Autonomous Robots*, vol. 1(1), 7–25, 1994.

[Brooks *et al.*, 1998] R. Brooks *et al.*. Alternative essences of intelligence. In *Procs. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pp. 961–976. Madison, WI, July 1998.

[Brooks *et al.*, 1999] R. Brooks *et al.*. The Cog Project: Building a Humanoid Robot. *Lecture Notes in Computer Science*, vol. 1562, 52–87, 1999. URL `citeseer.nj.nec.com/brooks99cog.html`.

[Bruce *et al.*, 2001] A. Bruce *et al.*. The role of expressiveness and attention in humanrobot interaction, 2001. URL `citeseer.nj.nec.com/bruce01role.html`.

[Bui *et al.*, 2002] T. Bui *et al.*. ParleE: An adaptive plan-based event appraisal model of emotions. In *In Procs. KI 2002: Advances in Artificial Intelligence* (edited by G. L. M. Jarke, J. Koehler), 2002.

[Bullock, 1983] D. Bullock. Seeking relations between cognitive and social-interactive transitions. In *Levels and transitions in children's development: new directions for child development* (edited by K. Fischer). Jossey-Bass Inc., 1983.

[Burges, 1998] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, vol. 2(2), 121–167, 1998.

[Cañamero and Fredslund, 2000] L. Cañamero and J. Fredslund. How Does It Feel? Emotional Interaction with a Humanoid LEGO Robot. In *AAAI 2000 Fall Symposium. Socially intelligent agents: the human in the loop* (edited by K. Dautenhahn), pp. 23–28. Menlo Park, California, 2000.

[Cañamero and Fredslund, 2001] L. Cañamero and J. Fredslund. I Show You How I Like You-Can You Read it in My Face? *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 5(31), 459–459, 2001.

[Carney and Colvin, 2005] D. Carney and C. Colvin. The circumplex structure of emotive social behavior, 2005. Manuscript in preparation.

[Castellucci *et al.*, 1970] V. Castellucci *et al.*. Neuronal Mechanisms of habituation and dishabituation of the gill-withdrawal reflex in *Aplysia*. *Science*, vol. 167, 1745–1748, 1970.

[Castrillón, 2003] M. Castrillón. *On Real-Time Face Detection in Video Streams. An Opportunistic Approach.*. Ph.D. thesis, Universidad de Las Palmas de Gran Canaria, March 2003.

[Chang, 2000] C. Chang. Improving Hallway Navigation in Mobile Robots with Sensory Habituation. In *Proc. of the Int. Joint Conference on Neural Networks (IJCNN2000)*, vol. V, pp. 143–147. Como, Italy, 2000.

[Chomsky, 1980] N. Chomsky. Rules and representations. *Behavioral and Brain Sciences*, (3), 1–21, 1980.

[Cielniak *et al.*, 2003] G. Cielniak *et al.*. Appearance-based Tracking of Persons with an Omnidirectional Vision Sensor. In *Proceedings of the Fourth IEEE Workshop on Omnidirectional Vision (Omnivis 2003)*. Madison, Wisconsin, 2003.

[Cleeremans, 2001] A. Cleeremans. Conscious and unconscious processes in cognition. In *International Encyclopedia of Social and Behavioral Sciences* (edited by N. Smelser and P. Baltes), vol. 4, pp. 2584–2589. London: Elsevier, 2001.

[Collins and Dennis, 2000] G. Collins and L. A. Dennis. System Description: Embedding Verification into Microsoft Excel. In *Conference on Automated Deduction*, pp. 497–501, 2000. URL `citeseer.nj.nec.com/collins00system.html`.

[Corbetta and Shulman, 2002] M. Corbetta and G. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, vol. 3, 201–215, March 2002.

[Crook and Hayes, 2001] P. Crook and G. Hayes. A Robot Implementation of a Biologically Inspired Method ofr Novelty Detection. In *Proceedings of TIMR 2001 - Towards Intelligent Mobile Robots*. Manchester, April 2001.

[Damasio, 1994] A. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. Avon Books, New York, 1994.

[Damper *et al.*, 1999] R. Damper *et al.*. ARBIB: an autonomous robot based on inspirations from biology, in press, 1999. URL `citeseer.nj.nec.com/damper99arbib.html`.

[Darrell *et al.*, 1998] T. Darrell *et al.*. Integrated Person Tracking Using Stereo, color, and Pattern Detection. In *Procs. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 601–608. Santa Barbara, CA, 1998.

[Dautenhahn, 1995] K. Dautenhahn. Getting to know each other - Artificial social intelligence for social robots. *Robotics and Autonomous Systems*, (6), 333–356, 1995.

[Dautenhahn and Werry, 2001] K. Dautenhahn and I. Werry. The AURORA Project: Using Mobile Robots in Autism Therapy. *Learning Technology online newsletter, publication of IEEE Computer Society Learning Technology Task Force (LTTF)*, vol. 3(1), 2001.

[Dautenhahn and Werry, 2002] K. Dautenhahn and I. Werry. A quantitative technique for analysing robot-human interactions. In *Procs. of the International Conference on Intelligent Robots and Systems*. Lausanne, Switzerland, October 2002.

[Davis and Vaks, 2001] J. Davis and S. Vaks. A Perceptual User Interface for Recognizing Head Gesture Acknowledgements. In *Proc. of ACM Workshop on Perceptual User Interfaces*. Orlando, Florida, November 2001.

[de Laar *et al.*, 1997] P. V. de Laar *et al.*. Task-dependent learning of attention. *Neural networks*, vol. 10(6), 981–992, 1997.

[De Renzi, 1997] E. De Renzi. Prosopagnosia. In *Behavioral Neurology and Neuropsychology* (edited by T. Feinberg and M. Farah). McGraw-Hill, New York, 1997.

[DiSalvo *et al.*, 2002] C. DiSalvo *et al.*. All robots are not created equal: the design and perception of humanoid robot heads. In *Procs. of the conference on Designing interactive systems: processes, practices, methods, and techniques*, pp. 321–326. London, 2002.

[Driscoll *et al.*, 1998] J. Driscoll *et al.*. A visual attention network for a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robotic Systems*. Victoria, B.C., Canada, October 1998.

[Duffy *et al.*, 2002] B. Duffy *et al.*. Issues in Assessing Performance of Social Robots. In *Procs. of the 2nd WSEAS International Conference on Robotics, Distance Learning and Intelligent Communication Systems*. Skiathos Island, Greece, September 2002.

[E. Hudlicka and J.M. Fellous, 2004] E. Hudlicka and J.M. Fellous. The Emotion Home Page, 2004. URL `http://emotion.salk.edu/emotion.html`.

[Ebbinghaus, 1913] H. Ebbinghaus. *Memory. A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York, 1913.

[FLUIDS Project, 2003] FLUIDS Project. Future Lines of User Interface Decision Support. Natural Language Generation, 2003. URL `http://www.dfki.de/fluids/Natural_Language_Generation.html`.

[Fong *et al.*, 2003] T. Fong *et al.*. A survey of socially interactive robots. *Robotics and Autonomous Systems*, vol. 42(3-4), March 2003.

[Fraunhofer Institut AIS, 2004] Fraunhofer Institut AIS. Omnidirectional Imaging for Robotic Applications, 2004. URL `http://ais.gmd.de/services/OmniVision/omni-intro.html`.

[Frith, 1989] U. Frith. *Autism: Explaining the enigma*. Blackwell, 1989.

[Gardner, 1983] H. Gardner. *Frames of mind: The theory of multiple intelligences*. Basic Books, New York, 1983.

[Gaspar, 2002] J. Gaspar. *Omnidirectional vision for mobile robot navigation*. Ph.D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, December 2002.

[GCAT, 1999] GCAT. Perception of Direction, 1999. Available at `http://www.gcat.clara.net/Hearing/perception_of_direction.htm`.

[Georghiades *et al.*, 1998] A. Georghiades *et al.*. Illumination cones for recognition under variable lighting: Faces, 1998. URL `citeseer.ist.psu.edu/georghiades98illumination.html`.

[Goetz, 1997] P. Goetz. *Attractors in recurrent behavior networks*. Ph.D. thesis, Graduate School of the State University of New York at Buffalo, August 1997.

[Good and Gilkey, 1996] M. Good and R. Gilkey. Sound localization in noise: the effect of signal-to-noise ratio. *J. Acoust. Soc. Am.*, vol. 99(2), 1108–1117, February 1996.

[Goto and Muraoka, 1997] M. Goto and Y. Muraoka. Issues in evaluating beat tracking systems. In *Working Notes of the IJCAI-97 Workshop on Issues in AI and Music, (in press)*, 1997. URL `citeseer.nj.nec.com/goto97issues.html`.

[Grandin, 1995] T. Grandin. *Thinking in Pictures: and Other Reports of my Life with Autism*. Bantam, New York, 1995.

[Grandin and Scariano, 1986] T. Grandin and M. Scariano. *Emergence: Labeled autistic*. Arena Press, Navato, California, 1986.

[Grange *et al.*, 2002] S. Grange *et al.*. Vision based Sensor Fusion for Human-Computer Interaction, 2002. URL `citeseer.nj.nec.com/grange02visionbased.html`.

[Gross *et al.*, 2002] R. Gross *et al.*. Fisher light-fields for face recognition across pose and illumination, 2002. URL `citeseer.ist.psu.edu/gross02fisher.html`.

[Gross *et al.*, 2004] R. Gross *et al.*. Face Recognition Across Pose and Illumination. In *Handbook of Face Recognition* (edited by S. Z. Li and A. K. Jain). Springer-Verlag, June 2004.

[Grove and Fisher, 1996] T. Grove and R. Fisher. Attention in iconic object matching. In *Procs. of British Machine Vision Conference*, pp. 293–302. Edinburgh, September 1996.

[H. von Helmholtz, 1924] H. von Helmholtz. *Physiological Optics*. Optical Society of America, 1924.

[Haritaolu *et al.*, 2000] I. Haritaolu *et al.*. W-4: Real-time surveillance of people and their activities. *IEEE Transactions Pattern Analysis, and Machine Intelligence*, vol. 22(8), 809–830, 2000.

[Härmä and Palomäki, 1999] A. Härmä and K. Palomäki. HUTear - a free Matlab toolbox for modeling of human auditory system. In *Procs. Matlab DSP conference*, pp. 96–99. Espoo, Finland, November 1999.

[Hartmann, 1999] W. Hartmann. How we localize sound. *Physics Today*, vol. 52(11), 24–29, 1999.

[Hassin and Trope, 2000] T. Hassin and Y. Trope. Facing faces: studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, vol. 78(5), 837–852, 2000.

[Heinke and Humphreys, 2001] D. Heinke and G. Humphreys. Computational Models of Visual Selective Attention: A Review. In *Connectionist models in psychology* (edited by G. Houghton), 2001.

[Hendriks-Jansen, 1997] H. Hendriks-Jansen. The epistemology of autism: Making a case for an embodied, dynamic and historical explanation. *Cybernetics and Systems*, vol. 28(5), 359–415, 1997.

[Hernández-Cerpa, 2001] D. Hernández-Cerpa. *ZagaZ: Entorno Experimental para el Tratamiento de Conductas en Caracteres Sintéticos*. Master's thesis, Universidad de Las Palmas de Gran Canaria, 2001.

[Hernández *et al.*, 2004] D. Hernández *et al.*. BDIE: a BDI like Architecture with Emotional Capabilities. In *Papers from the 2004 AAAI Spring Symposium: Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, 2004.

[Hernández, 1999] P. H. Hernández. *Natura y Cultura de Las Islas Canarias*. TAFOR Publicaciones, La Laguna, Tenerife, 1999.

[Hernández Tejera *et al.*, 1999] F. Hernández Tejera *et al.*. DESEO: An Active Vision System for Detection, Tracking and Recognition. In *Lectures Notes in Computer Science, International Conference on Vision Systems (ICVS'99)* (edited by H. I. Christensen), vol. 1542, pp. 379–391, 1999.

[Hewett, 2001] M. Hewett. *Computational perceptual attention*. Ph.D. thesis, Department of Computer Sciences, The University of Texas at Austin, May 2001.

[HHMI, 1997] HHMI. Seeing, Hearing and Smelling the World. Tech. rep., Howard Hughes Medical Institute, 1997. Available at `http://www.hhmi.org/senses/c210.html`.

[Hjelmas and Farup, 2001] E. Hjelmas and I. Farup. Experimental Comparison of Face/Non-Face Classifiers. In *Procs. of the Third International Conference on Audio- and Video-Based Person Authentication. Lecture Notes in Computer Science 2091*, 2001.

[Holland, 1997] O. Holland. Grey Walter: The Pioneer of Real Artificial Life. In *Proc. of the 5th International Workshop on Artificial Life* (edited by C. Langton). Cambridge, 1997.

[Holland *et al.*, 2000] S. Holland *et al.*. Determination of plate source, detector separation fron one signal. *Ultrasonics*, vol. 38, 620–623, 2000.

[Hu, 1962] M. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, vol. 8(2), 179–187, 1962.

[Hudson, 1967] L. Hudson. *Contrary Imaginations; a psychological study of the English Schoolboy*. Penguin, 1967.

[Humphrys, 1997] M. Humphrys. *Action Selection methods using Reinforcement Learning*. Ph.D. thesis, Trinity Hall, University of Cambridge, June 1997.

[Hutchinson, 1958] G. Hutchinson. Concluding Remarks. In *Cold Spring Harbor Symp. Quant. Biol.*, 22, pp. 415–427, 1958.

[Irie, 1995] R. Irie. *Robust sound localization: an application of an auditory perception system for a humanoid robot*. Master's thesis, Massachusetts Institute of Technology, June 1995.

[Itti, 2003] L. Itti. Modeling primate visual attention. In *Computational Neuroscience: A Comprehensive Approach* (edited by J. Feng). CRC Press, Boca Ratón, 2003.

[Itti and Koch, 2001] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, vol. 2(3), 194–203, March 2001.

[J.Huang *et al.*, 1999] J.Huang *et al.*. A model based sound localization system and its application to robot navigation. *Robotics and Autonomous Systems (Elsevier Science)*, vol. 27(4), 199–209, 1999.

[Johnson and Dudgeon, 1993] D. Johnson and D. Dudgeon. *Array signal processing: concepts and techniques*. Prentice Hall, 1993.

[Kahana and Adler, 2002] M. Kahana and M. Adler. Note on the power law of forgetting, 2002. Journal of Mathematical Psychology (Submitted).

[Kahn, 1996] R. Kahn. *Perseus: An Extensible Vision System for Human-Machine Interaction*. Ph.D. thesis, University of Chicago, August 1996.

[Kanda *et al.*, 2001] T. Kanda *et al.*. Psychological analysis on human-robot interaction. In *Procs. of the IEEE Int. Conference on Robotics and Automation*, pp. 4166–4173, 2001.

[Kaplan and Hafner, 2004] F. Kaplan and V. Hafner. The Challenges of Joint Attention. In *Proceedings of the Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 117*, pp. 67–74. Genoa, Italy, 2004.

[Kapoor and Picard, 2001] A. Kapoor and R. Picard. A real-time head nod and shake detector, 2001. URL `citeseer.nj.nec.com/kapoor01realtime.html`.

[Kassen and Bell, 1998] R. Kassen and G. Bell. Experimental evolution in Chlamydomonas. IV. Selection in environments that vary through time at different scales. *Heredity*, vol. 80(6), 732–741, 1998.

[Keisler and Sproull, 1997] S. Keisler and L. Sproull. Social human computer interaction, human values and the design of computer technology. pp. 191–199. Stanford, CA, 1997.

[Kidd and Breazeal, 2003] C. Kidd and C. Breazeal. Comparison of social presence in robots and animated characters, 2003. HCI Journal, special issue on Human-Robot Interaction (Submitted).

[Kihlstrom and Cantor, 2000] J. Kihlstrom and N. Cantor. *Social intelligence*, vol. Handbook of intelligence, pp. 359–379. Cambridge University Press, Cambridge, U.K., 2nd ed., 2000.

[King and Ohya, 1996] W. King and J. Ohya. The representation of agents: Anthropomorphism, agency and intelligence. In *Proc. of CHI-96*, 1996.

[Kirkpatrick, 1971] D. Kirkpatrick. *A practical guide for supervisory training and development*. Addison-Wesley Publishing Co., Reading, MA, 1971.

[Kjeldsen, 2001] R. Kjeldsen. Head Gestures for Computer Control. In *Proc. of the Workshop on Recognition And Tracking of Face and Gesture – Real Time Systems (RATFG-RTS)*. Vancouver, BC, Canada, July 2001.

[Koda and Maes, 1996] T. Koda and P. Maes. Agents with faces: a study on the effect of personification of software agents. In *Proc. of the 5th IEEE Int. Workshop on Robot and Human Communication (ROMAN 96)*, pp. 189–194, 1996.

[Koku *et al.*, 2000] A. Koku *et al.*. Towards socially acceptable robots. In *Proc. of 2000 IEEE International Conference on Systems, Man and Cybernetics*, pp. 894–899, 2000.

[Kopp and Gärdenfors, 2001] L. Kopp and P. Gärdenfors. *Attention as a Minimal Criterion of Intentionality in Robots*, vol. 89 of *Lund University Cognitive Studies*, 2001.

[Kozima, 2002] H. Kozima. Infanoid: An experimental tool for developmental psycho-robotics. In *Procs. Int. Workshop on Developmental Study*, 2002.

[Kozima and Yano, 2001] H. Kozima and H. Yano. A robot that learns to communicate with human caregivers, 2001. URL `citeseer.nj.nec.com/kozima01robot.html`.

[Krumm *et al.*, 2000] J. Krumm *et al.*. Multi-camera Multi-person Tracking for EasyLiving. In *3rd IEEE International Workshop on Visual Surveillance*. Dublin, Ireland, 2000.

[Lang *et al.*, 2003] S. Lang *et al.*. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proc. Int. Conf. on Multimodal Interfaces*, pp. 28–35. Vancouver, Canada, November 2003.

[Laurel, 1990] B. Laurel. Interface Agents: Metaphors with Character. In *The Art of Human-Computer Interface Design* (edited by B. Laurel). Reading, MA, 1990.

[Lee and Kiesler, 2005] S. Lee and S. Kiesler. Human mental models of humanoid robots. In *Proceedings of the Int. Conference on Robotics and Automation, ICRA 2005*. Barcelona, SPAIN, April 2005.

[Liang *et al.*, 2002] L. Liang *et al.*. Speaker independent audio-visual continuous speech recognition. In *IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 25–28, August 2002.

[Liao *et al.*, 1997] H. M. Liao *et al.*. Face Recognition Using a Face-Only Database: A New Approach. *Lecture Notes in Computer Science*, vol. 1352, 1997. URL `citeseer.ist.psu.edu/liao97face.html`.

[Lindblom and Ziemke, 2003] J. Lindblom and T. Ziemke. Social Situatedness of Natural and Artificial Intelligence: Vygotsky and Beyond. *Adaptive Behavior*, vol. 11(2), 79–96, 2003.

[Lisetti and Schiano, 2000] C. L. Lisetti and D. J. Schiano. Automatic Facial Expression Interpretation: Where Human-Computer Interaction, Artificial Intelligence and Cognitive Science Intersect. *Pragmatics and Cognition (Special Issue on Facial Information Processing: A Multidisciplinary Perspective*, vol. 8(1), 185–235, 2000.

[Littlewort *et al.*, 2003] G. Littlewort *et al.*. Towards social robots: Automatic evaluation of human robot interaction by face Detection and expression classification. In *Procs. of the Int. Conf. Advances in Neural Information Processing Systems (NIPS)*, vol. 16. MIT Press, 2003.

[Liu *et al.*, 2002] X. Liu *et al.*. Audio-visual continuous speech recognition using a coupled Hidden Markov Model. In *IEEE Int. Conference on Spoken Language Processing*, pp. 213–216, September 2002.

[Lorenz, 1981] K. Lorenz. *The Foundations of Ethology*. Springer Verlag, Heidelberg, 1981.

[Lorenzo and Hernández, 2002*a*] J. Lorenzo and M. Hernández. Habituation Based on Spectrogram Analysis. In *Lecture Notes in Artificial Intelligence. Advances in Artificial Intelligence, IBERAMIA 2002*, vol. 2527, 2002*a*. Sevilla, Spain.

[Lorenzo and Hernández, 2002*b*] J. Lorenzo and M. Hernández. A Habituation Mechanism for a Perceptual User Interface. In *Proceedings of 11th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN'2002)*, 2002*b*. Berlin, Germany.

[Lungarella *et al.*, 2004] M. Lungarella *et al.*. Developmental Robotics: a Survey. *Connection Science*, vol. 0(0), 1–40, 2004.

[Maes, 1989] P. Maes. The dynamics of action selection. In *Procs. of the 11th International Joint Conference on Artificial Intelligence (IJCAI-89)*, pp. 991–997, 1989.

[Maes, 1990] P. Maes. How to do the Right Thing. *Connection Science Journal, Special Issue on Hybrid Systems*, vol. 1, 1990. URL `citeseer.nj.nec.com/maes89how.html`.

[Maes, 1994] P. Maes. Agents that Reduce Work and Information Overload. *Communications of the ACM*, vol. 37(7), 31–40, July 1994.

[Mandler, 1984] G. Mandler. *Mind and body: Psychology of emotion and stress*. Norton, N.Y., 1984.

[Marsland *et al.*, 1999] S. Marsland *et al.*. A Model of Habituation Applied to Mobile Robots. In *Proceedings of TIMR 1999 - Towards Intelligent Mobile Robots*. Bristol, 1999.

[Marsland *et al.*, 2000] S. Marsland *et al.*. Detecting Novel Features of an Environment Using Habituation. In *From Animals to Animats, The Sixth International Conference on Simulation of Adaptive Behaviour*. Paris, 2000.

[Martin *et al.*, 2000] A. Martin *et al.*. An Introduction to Evaluating Biometric Systems. *Computer*, vol. 56, 56–63, February 2000.

[Martínez, 2003] J. Martínez. SEGURITRON, 2003. URL `http://www.seguritron.com`.

[Matsusaka *et al.*, 1999] Y. Matsusaka *et al.*. Multi-person Conversation Robot using Multi-modal Interface. In *Proceedings of World Multiconference on Systems, Cybernetic and Informatics*, vol. 7, pp. 450–455, 1999.

[Maxwell, 2003] B. Maxwell. A real-time vision module for interactive perceptual agents. *Machine Vision and Applications*, (14), 72–82, 2003.

[Maxwell *et al.*, 1999] B. Maxwell *et al.*. Alfred: the robot waiter who remembers you. In *Procs. of AAAI Workshop on Robotics*, July 1999.

[McCowan, 2001] I. McCowan. *Robust speech recognition using microphone arrays*. Ph.D. thesis, Queensland University of Technology, Australia, 2001.

[McNair *et al.*, 1971] D. McNair *et al.*. Profile of mood states manual, 1971. Educational and industrial testing service. San Diego.

[Medonis Engineering, 2003] Medonis Engineering. ASC16 Advanced Servo Controller, 2003. URL `http://www.medonis.com/asc16.html`.

[Metta, 2001] G. Metta. An attentional system for a humanoid robot exploiting space variant vision. In *IEEE-RAS International Conference on Humanoid Robots*, pp. 359–366. Tokyo, November 2001.

[Metta *et al.*, 2000*a*] G. Metta *et al.*. Babybot: A biologically inspired developing robotic agent. In *Proceedings of SPIE*. Boston, USA, November 2000*a*.

[Metta *et al.*, 2000*b*] G. Metta *et al.*. Babybot: an artificial developing robotic agent. In *Proceedings of SAB 2000*. Paris, France, September 2000*b*.

[Milanese *et al.*, 1994] R. Milanese *et al.*. Integration Of Bottom-Up And Top-Down Cues For Visual Attention Using Non-Linear Relaxation. In *Procs. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 781–785, 1994.

[MIT AI lab, Humanoid Robotics Group, 2003] MIT AI lab, Humanoid Robotics Group. Kismet, 2003. `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html`.

[Mitchell, 1997] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[Miwa *et al.*, 2001] H. Miwa *et al.*. Human-like Robot Head that has Olfactory Sensation and Facial Color Expression. In *Procs. of the 2001 2001 IEEE International Conference on Robotics and Automation*, pp. 459–464, 2001.

[Mobahi, 2003] H. Mobahi. *Building an Interactive Robot Face from Scratch*. Master's thesis, Azad University, Tehran, Iran, May 2003.

[Montero *et al.*, 1998] J. Montero *et al.*. Emotional speech synthesis: from speech database to TTS. In *In Procs. of the 5th Int. Conference on Spoken Language Processing*, vol. 3, pp. 923–926. Sydney, Australia, 1998.

[Moreno *et al.*, 2001] F. Moreno *et al.*. Localization of human faces fusing color segmentation and depth from stereo, 2001. URL `citeseer.nj.nec.com/moreno01localization.html`.

[Mottram, 2003] R. Mottram. Rodney. The Blueprint for a Plonker, 2003. URL `http://www.fuzzgun.btinternet.co.uk/rodney/rodney.htm`.

[Nakadai *et al.*, 2000] K. Nakadai *et al.*. Humanoid Active Audition System Improved by the Cover Acoustics. In *PRICAI-2000 Topics in Artificial Intelligence (Sixth Pacific Rim International Conference on Artificial Intelligence)*, vol. 1886 of *Lecture Notes in Artificial Intelligence*, pp. 544–554. Springer Verlag, Melbourne, Australia, August 2000.

[Nass *et al.*, 1994] C. Nass *et al.*. Computers are social actors. In *Proc. of CHI'94*, 1994.

[Nayar, 1998] S. Nayar. Omnidirectional Vision. In *Procs. of the British Machine Vision Conference*. Southampton, UK, 1998.

[Nayar and Boult, 1997] S. Nayar and T. Boult. Omnidirectional vision systems: PI report. In *Procs. of the 1997 DARPA Image Understanding Workshop*, 1997.

[Nefian *et al.*, 2002] A. Nefian *et al.*. A coupled HMM for audio-visual speech recognition. In *International Conference on Acoustics Speech and Signal Processing*, vol. 2, pp. 2013–2016. Orlando, Florida, May 2002.

[Newell, 1990] A. Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA, 1990.

[Nowak, 2004] R. Nowak. ECE 901: Statistical Regularization and Learning Theory. Complexity and Regularization, 2004. URL `http://www.ece.wisc.edu/~nowak/ece901/`.

[Okuno *et al.*, 2002] H. Okuno *et al.*. Social Interaction of Humanoid Robot Based on Audio-Visual Tracking. In *Proc. of 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002)*. Cairns, Australia, June 2002.

[Oren *et al.*, 1990] T. Oren *et al.*. Guides: Characterizing the Interface. In *The Art of Human-Computer Interface Design* (edited by B. Laurel). Reading, MA, 1990.

[Ortega-García and González-Rodríguez, 1996] J. Ortega-García and J. González-Rodríguez. Overview Of Speech Enhancement Techniques For Automatic Speaker Recognition, 1996. URL `citeseer.nj.nec.com/506045.html`.

[Oviatt, 2000] S. Oviatt. Taming recognition errors with a multimodal interface. *Communications of the ACM*, vol. 43(9), 45–51, 2000.

[Pentland, 2000] A. Pentland. Perceptual intelligence. *Communications of the ACM*, vol. 43(3), 35–44, 2000.

[Perlin, 1995] K. Perlin. Real Time Responsive Animation with Personality. *IEEE Trans. on visualization and computer graphics*, vol. 1(1), March 1995.

[Peters and Sowmya, 1998] M. Peters and A. Sowmya. Integrated techniques for self-organisation, sampling, habituation, and motion-tracking in visual robotics applications. In *Proceedings of Machine Vision Applications '98*, 1998. Tokyo, Japan.

[Pfeifer and Scheier, 1999] R. Pfeifer and C. Scheier. *Understanding Intelligence*. MIT Press, Cambridge, MA, 1999.

[Phillips, 2002] J. Phillips. Face Recognition Vendor Test, 2002. URL `www.frvt.org`.

[Picard, 1997] R. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[Pirjanian, 1997] P. Pirjanian. An Overview of System Architectures for Action Selection in Mobile Robotics. Tech. Rep. Technical Report, Laboratory of Image Analysis, Aalborg University, 1997.

[Playmates Toys Inc., 2005] Playmates Toys Inc. PLAYMATES TOYS Expands Line Of "Amazing" Products - Amazing Amanda, 2005. URL `http://www.playmatestoys.com/html/corporate/pressreleases.php`.

[Puzicha *et al.*, 1999] J. Puzicha *et al.*. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings the IEEE International Conference on Computer Vision(ICCV-1999)*, pp. 1165–1173, 1999. URL `citeseer.ist.psu.edu/article/puzicha99empirical.html`.

[Quijada *et al.*, 2004] S. D. Quijada *et al.*. Development of an expressive social robot, 2004. URL `citeseer.nj.nec.com/593058.html`.

[Rabinkin *et al.*, 1996] D. Rabinkin *et al.*. A DSP implementation of source location using microphone arrays. In *Procs. of the SPIE*, vol. 2846, pp. 88–99. Denver, Colorado, August 1996.

[Rageul, 2000] Y. Rageul. Informe final de proyecto. Estancia en el CeTSIA, 2000.

[Raskin, 2000] J. Raskin. *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley Professional, 2000.

[Reid and Milios, 1999] G. Reid and E. Milios. Active stereo sound localization. Tech. Rep. CS-1999-09, York University, 1999.

[Reilly, 1996] W. Reilly. Believable Social and Emotional Agents. Tech. Rep. Ph.D. Thesis. Technical Report CMU-CS-96-138, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. May 1996, 1996. URL `citeseer.nj.nec.com/reilly96believable.html`.

[Rhino Robotics Ltd., 2003] Rhino Robotics Ltd. Rhino Robotics, 2003. URL `http://www.rhinorobotics.com`.

[Rhodes, 1996] B. Rhodes. PHISH-nets: Planning Heuristically In Situated Hybrid Networks. Tech. rep., MIT Media Lab, 1996.

[Rowley *et al.*, 1998] H. Rowley *et al.*. Neural Network-Based Face Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20(1), 23–38, 1998.

[Rubin and Wenzel, 1996] D. Rubin and A. Wenzel. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, (103), 734–760, 1996.

[Russell, 1980] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychiology*, vol. 39(6), 1161–1178, 1980.

[Ryu, 2001] K. Ryu. Application of the sound localization algorithm to the global localization of a robot. Tech. Rep. UG 2001-5, Institute for Systems Research, University of Maryland, 2001.

[Sacks, 1985] O. Sacks. *The man who mistook his wife for a hat*. Touchstone Books, 1985.

[Sacks, 1995] O. Sacks. *An anthropologist on Mars*. Vintage Books, 1995.

[Scassellati, 2000] B. Scassellati. How Robotics and Developmental Psychology Complement Each Other. In *NSF/DARPA Workshop on Development and Learning*. Lansing, MI, 2000.

[Scassellati, 2001] B. Scassellati. *Foundations for a Theory of Mind for a Humanoid Robot*. Ph.D. thesis, MIT Department of Computer Science and Electrical Engineering, May 2001.

[Scheffer, 1999] T. Scheffer. *Error estimation and model selection*. Ph.D. thesis, Technische Universitaet Berlin, School of Computer Science, 1999.

[Scholtz and Bahrami, 2003] J. Scholtz and S. Bahrami. Human-Robot Interaction: Development of an Evaluation Methodology for the Bystander Role of Interaction. In *Procs. of the Systems, Man, and Cybernetics Conference*. Washington DC, 2003.

[School of Computing, University of Leeds, 2003] School of Computing, University of Leeds. Cognitive Science Learning Resource. Natural Language Generation, 2003. URL `http://www.comp.leeds.ac.uk/ugadmit/cogsci/spchlan/nlgen.htm`.

[Schröder, 2001] M. Schröder. Emotional speech synthesis: A review. In *In Procs. of Eurospeech 2001*, vol. 1, pp. 561–564. Aalborg, Denmark, 2001.

[Schulte *et al.*, 1999] J. Schulte *et al.*. Spontaneous short-term interaction with mobile robots in public places. In *Procs. of the IEEE Int. Conference on Robotics and Automation*, 1999.

[Selfe, 1977] L. Selfe. *Nadia: A case of extraordinary drawing ability in children*. Academic Press, London, 1977.

[Seltzer, 2001] M. Seltzer. Calibration of microphone arrays for improved speech recognition. Tech. Rep. TR2001-43, Mitsubishi Electric Research Laboratories, December 2001.

[Shah and Henry, 2005] H. Shah and O. Henry. The Confederate Effect in Human-Machine Textual Interaction, 2005. URL `http://www.alicebot.org/articles/`.

[Shibata and Tanie, 2001] T. Shibata and K. Tanie. Physical and Affective Interaction between Human and Mental Commit Robot. In *Proceedings of 2001 IEEE International Conference on Robotics and Automation*, pp. 2572–2577, 2001.

[Shibata *et al.*, 2003] T. Shibata *et al.*. Statistical Analysis and Comparison of Questionnaire Results of Subjective Evaluations of Seal Robot in Japan and U.K. In *Procs. of the IEEE Int. Conference on Robotics and Automation*. Taipei, Taiwan, September 2003.

[Shinn-Cunningham, 2003] B. Shinn-Cunningham. Acoustics and perception of sound in everyday environments. In *Procs. of the 3rd International Workshop on Spatial Media*, pp. 31–40. Aizu-Wakamatsu, Japan, March 2003.

[Siino and Hinds, 2005] R. Siino and P. Hinds. Robots, Gender and Sensemaking: Sex Segregation Impact On Workers Making Sense of a Mobile Autonomous Robot. In *Proceedings of the Int. Conference on Robotics and Automation, ICRA 2005*. Barcelona, SPAIN, April 2005.

[Sloman, 1995] A. Sloman. Exploring design space and niche space. In *Proc. of 5th Scandinavian Conf. on AI*. Amsterdam, May 1995.

[Sloman, 2001] A. Sloman. Beyond shallow models of emotion. *Cognitive Processing*, vol. 2(1), 177–198, 2001.

[Snyder and Mitchell, 1999] A. Snyder and D. J. Mitchell. Is Integer Arithmetic Fundamental to Mental Processing?: The Mind's Secret Arithmetic. *Proceedings of the Royal Society of London*, (266), 587–592, 1999.

[Snyder and Thomas, 1997*a*] A. Snyder and M. Thomas. Autistic artists give clues to cognition. *Perception*, (23), 93–96, 1997*a*.

[Snyder and Thomas, 1997*b*] A. Snyder and M. Thomas. Breaking mindset. *Mind and language*, (13), 1–10, 1997*b*.

[Snyder *et al.*, 2004] A. Snyder *et al.*. Concept formation: 'Object' attributes dynamically inhibited from conscious awareness. *Journal of Integrative Neuroscience*, vol. 3(1), 31–46, 2004.

[Sokolov, 1963] E. Sokolov. *Perception and the conditioned reflex*. MacMillan, N.Y., 1963.

[Stanley, 1976] J. Stanley. Computer Simulation of a Model of Habituation. *Nature*, vol. 261, 146–148, 1976.

[Stern, 2004] J. Stern. Why do we blink? Mainly to cleanse the cornea, but more to it than meets the eye, 2004. URL `http://msnbc.msn.com/id/3076704/`.

[Sternberg *et al.*, 1981] R. Sternberg *et al.*. People's conceptions of intelligence. *Journal of personality and social psychology*, vol. 41(1), 37–55, 1981.

[Stiefelhagen, 2002] R. Stiefelhagen. Tracking Focus of Attention in Meetings. Pittsburgh, USA, October 2002.

[Stiefelhagen *et al.*, 2003] R. Stiefelhagen *et al.*. Capturing Interactions in Meetings with Omnidirectional Cameras. Nice, France, 2003.

[Stiles and Ghosh, 1995] B. Stiles and J. Ghosh. A habituation based neural network for spatio-temporal classification. In *Proceedings of the 1995 IEEE Workshop In Neural Networks for Signal Processing*, pp. 135–144. Cambridge, MA, September 1995.

[Stoytchev and Arkin, 2003] A. Stoytchev and R. Arkin. Combining Deliberation, Reactivity and Motivation in the Context of a Behavior-Based Robot Architecture, 2003. Available at `http://www.cc.gatech.edu/ai/robot-lab/publications.html`.

[Swain and Ballard, 1991] M. Swain and D. Ballard. Color Indexing. *Int. Journal on Computer Vision*, vol. 7(1), 11–32, 1991.

[Takanishi Laboratory, 2003] Takanishi Laboratory. Human-like head robot, 2003. URL `http://www.takanishi.mech.waseda.ac.jp/eyes/`.

[Takeuchi and Naito, 1995] Y. Takeuchi and T. Naito. Situated facial displays: towards social interaction. In *Human factors in computing systems: CHI'95*. New York, 1995.

[Tang and Nakatsu, 2000] J. Tang and R. Nakatsu. A Head Gesture Recognition Algorithm. In *Proc. of International Conference on multi-modal Interface*. Beijing, China, October 2000.

[Thrun *et al.*, 2000] S. Thrun *et al.*. Probabilistic Algorithms and the Interactive Museum Tour-Guide Robot Minerva. *International Journal of Robotics Research*, vol. 19(11), 972–999, 2000.

[Treffert, 1989] D. Treffert. *Extraordinary people: Understanding the savant syndrome*. Harper and Row, New York, 1989.

[Trivedi *et al.*, 2000] M. Trivedi *et al.*. Active Camera Networks and Semantic Event Databases for Intelligent Environents. South Carolina, USA, June 2000.

[Tyrrell, 1993] T. Tyrrell. *Computational mechanisms for action selection*. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, 1993.

[van Breemen, 2004] A. van Breemen. Bringing Robots to Life: Applying principles of animation to robots. In *Procs. of the CHI2004 Workshop Shaping Human-Robot Interaction, Understanding the Social Aspects of Intelligent Robotic Products*. Vienna, Austria, April 2004.

[Vapnik, 1995] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

[Viola and Jones, 2001] P. Viola and M. Jones. Robust Real-time Object Detection. Technical Report Series CRL 2001/01, Cambridge Research Laboratory, February 2001.

[Wada *et al.*, 2003] K. Wada *et al.*. Psychological and Social Effects of Robot Assisted Activity to Elderly People who stay at a Health Service Facility for the Aged. In *Procs. of the IEEE Int. Conference on Robotics and Automation*. Taipei, Taiwan, September 2003.

[Wallace, 2005] R. Wallace. Zipf's Law, 2005. URL `http://www.alicebot.org/articles/wallace/zipf.html`.

[Walters *et al.*, 2005*a*] M. Walters *et al.*. Close encounters: Spatial distances between people and a robot of mechanistic appearance. In *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids2005)*, December 2005*a*.

[Walters *et al.*, 2005*b*] M. Walters *et al.*. The Influence of Subjects' Personality Traits on Personal Spatial Zones in a Human-Robot Interaction Experiment. In *Proc. IEEE ROMAN 2005*, pp. 347–352, 2005*b*.

[Wang, 1995] D. Wang. Habituation. In *The Handbook of Brain Theory and Neural Networks* (edited by M. A. Arbib), pp. 441–444. MIT Press, 1995.

[Wang *et al.*, 2004] J. Wang *et al.*. Face Image Resolution versus Face Recognition Performance Based on Two Global Methods. In *Procs. of the Asia Conference on Computer Vision (ACCV'2004)*. Jeju island, Korea, January 2004.

[Weiskrantz, 1998] L. Weiskrantz. *Blindsight: a case study and implications*. Oxford University Press, 1998.

[Wenger, 2003] M. Wenger. Noise rejection, the essence of good speech recognition. Tech. rep., Emkay Innovative Products, 2003.

[Wheatley and Wegner, 2001] T. Wheatley and D. M. Wegner. *Psychology of automaticity of action*, pp. 991–993. International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science Ltd., 2001.

[Williams *et al.*, 2001] J. Williams *et al.*. Imitation, mirror neurons and autism. *Neuroscience and Biobehavioural Review*, vol. 25(4), 287–295, 2001.

[Williams, 2004] K. Williams. *Build your own humanoid robots*. McGraw Hill, 2004.

[Wilson and Keil, 1999] R. Wilson and F. Keil, eds. *The MIT Encyclopedia of the Cognitive Sciences*. The MIT Press, Cambridge, Massachusetts, 1999.

[Winters, 2001] N. Winters. *A Holistic Approach to Mobile Robot Navigation using Omnidirectional Vision*. Ph.D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa and Trinity College, University of Dublin, October 2001.

[Wixted and Ebbesen, 1991] J. Wixted and E. Ebbesen. On the form of forgetting. *Psychological Science*, (2), 409–415, 1991.

[Wixted and Ebbesen, 1997] J. Wixted and E. Ebbesen. Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition*, (25), 731–739, 1997.

[Wolfe, 1994] J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, vol. 1(2), 202–238, 1994.

[Wolfe and Gancarz, 1996] J. Wolfe and G. Gancarz. *"Guided Search 3.0"*, pp. 189–192. Basic and Clinical Applications of Vision Science. Kluwer Academic, Netherlands, 1996.

[Wolpert, 1996] D. Wolpert. The existence of A priori distinctions between learning algorithms. *Neural Computation*, vol. 8(7), 1391–1420, 1996.

[Woods *et al.*, 2005] S. Woods *et al.*. Child and adults' perspectives on robot appearance. In *Proc. AISB'05 Symposium Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*. University of Hertfordshire, UK, April 2005.

[Yamato *et al.*, 2004] J. Yamato *et al.*. Effect of Shared-attention on Human-Robot Communication. In *Procs. of the CHI2004 Workshop Shaping Human-Robot Interaction, Understanding the Social Aspects of Intelligent Robotic Products*. Vienna, Austria, April 2004.

[Yost and Gourevitch, 1987] W. A. Yost and G. Gourevitch. *Directional hearing*. Springer-Verlag, New York, 1987.

[Young, 1998] A. Young. *Face and Mind*. Oxford Cognitive Science Series. Oxford University Press, 1998.

Index