Expressive Robotic Face for Interaction

Oscar Déniz, Luis Antón, Modesto Castrillón, Mario Hernández

Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería

Universidad de Las Palmas de Gran Canaria

Campus de Tafira

Las Palmas

odeniz@dis.ulpgc.es

Abstract

This paper describes the current development status of a robot head with basic interactional abilities. On a theoretical level we propose an explanation for the lack of robustness implicit in the so-called social robots. The fact is that our social abilities are mainly unconscious to us. This lack of knowledge about the form of the solution to these abilities leads to a fragile behaviour. Therefore, the engineering point of view must be seriously taken into account, and not only insights taken from human disciplines like developmental psychology or ethology. Our robot, built upon this idea, does not have a definite task, except to interact with people. Its perceptual abilities include sound localization, omnidirectional vision, face detection, an attention module, memory and habituation. The robot has facial features that can display basic emotional expressions, and it can speak canned text through a TTS. The robot's behavior is controlled by an action selection module, reflexes and a basic emotional module.

1 Introduction

In recent years there has been a surge in interest in a topic called social robotics. As used here, social robotics does not relate to groups of robots that try to complete tasks together. For a group of robots, communication is simple, they can use whatever complex binary protocol to "socialize" with their partners. For us, the adjective social refers to humans. In principle, the implications of this are much wider than the case of groups of robots. Socializing with humans is definitely much harder, not least because robots and humans do not share a common language nor perceive the world (and hence each other) in the same way. Many researchers working on this topic use other names like human-robot interaction or perceptual user interfaces.

This document describes CASIMIRO (The name is an Spanish acronym of "expressive face and basic visual processing for an interactive robot), a robot with basic social abilities. CASIMIRO was not designed for performing a certain precise task, like is the case of many traditional robots. If any, its task would be to interact with humans (note the vagueness of that). CASIMIRO is still under development and its capabilities will be expanded in the future. This paper is organized as follows. Section 2 gives an overview of other influential social robots. Section 3 outlines the approach taken in the building of our robot. Then we briefly describe the implemented perception and action abilities, Sections 5 and 6, and behavior control in Section 7. Finally, we summarize the conclusions and outline future work.

2 Previous Work

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. Kismet [3] 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 [11].

Inspiration and theories from human sciences was from the beginning involved in the design of these robots, mainly from psychology, ethology and infant social development studies. 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 [7]. The relationship between autism and social robotics has been twofold. On the one 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 [13]). On the other hand, social robots have been used as a therapeutic tool in autism.

Infanoid [9] 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 in autism. Attention sharing, the activity of paying attention to someone else's attentional target, plays an indispensable role in mindreading and learning, and it has been shown to be important for human-robot communication. 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.

Careful analysis of the related work leads to the question of whether these and other robots that try to accomplish social tasks have a robust behaviour. Particularly, face recognition (the social ability par excellence) is extremely sensitive to illumination, hair, eyeglasses, expression, pose, image resolution, 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. As for the other factors, acceptable performances can be achieved only under certain circumstances [10].

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 [12]. Above all, environmental noise is considered to be the worst obstacle [15]. The mouth-microphone distance is in this respect crucial. The typical achievable recognition rate (2003) 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.

In summary, there is the impression (especially among the robot builders themselves) that performance would degrade up to unacceptable levels when conditions are different from those used to train or 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 "under control". This leads us to the important question: is building a social robot in any sense different than building other kinds of robots? An answer will be given later on.

3 Is Building a "Social Robot" in any Sense Different than Building a Other Kinds of Robots?

We argue that the answer to the question that entitles this section is yes. Our account shall be brief for space reasons. The activities and processes that social robots try to replicate are generally of unconscious nature in humans, face recognition being the best example. Nowadays, the existence of unconscious processes in our brain seems to be beyond doubt. Freud's work already acknowledged that unconscious ideas and processes are critical in explaining the behaviour of people in all circumstances. Helmholtz, 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. In linguistics something similar has also been observed. There is evidence that speakers unconsciously assign a structure in constituents to sequences of words. Note that this is in contrast with other mental processes for which we are somehow able to articulate a vague form of the solution.

Some authors content that the reason why some mental processes fade into the unconscious is repetition and practice [1]. If this is the case, our social abilities should be more unconscious as they appear earlier in life. The reason of their well performing may well be the fact that they are unconscious, although we do not delve further on that aspect.

Therefore, and if we speak in terms of machine learning, our algorithms designed to tackle those problems will, in general, produce less robust results. This is due to the fact that less knowledge on the form of the solution (independent of the number and quality of the samples used to train) leads to an overfitting-like behaviour. Taking this into account, CASIMIRO has extensively used simple techniques (i.e. the engineering point of view, instead of using only human models as the main guideline), as will be shown in the next sections.

4 Robot Overview

This section describes the hardware that constitutes CASIMIRO. Details will be in general left out as the information is mainly technical data. It is important to introduce the hardware at this point because that helps focus the work described in the following sections. CASIMIRO is a robotic face: a set of (9) motors move a number of facial features placed on an aluminium skeleton. It also has a neck that moves the head. The neck has the pan and tilt movements, although they are not completely independent. The global aspect of the robot is shown in Figure 1.

5 Perception Abilities

This section gives an overview of the perceptual abilities implemented in CASIMIRO yet. Due to space constraints details will be omitted.

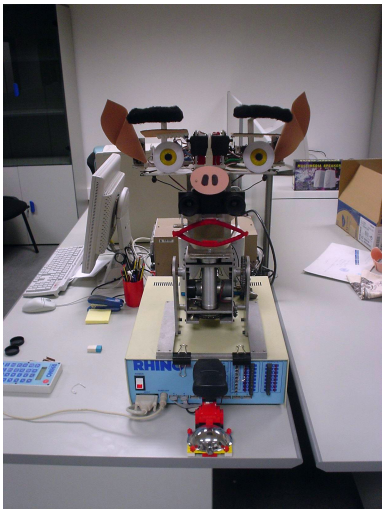

Figure 1: Global aspect of CASIMIRO.

Omnidirectional vision

As can be seen in Figure 1 in front of the robot there is an omnidirectional camera. The camera was built using a webcam plus a curved reflective surface (a ladle, note the engineering point of view). It allows the robot to have a 180^o field of view, similar to that of humans. Through adaptive background subtraction, the robot is able to localize people in the surroundings, and pan the neck toward them. The curvature of the mirror allows to extract a rough measure of the distance to the robot.

Sound localization

The robot has two omnidirectional microphones placed on both sides of the head. The signals gathered by them are amplified and filtered. The direction of the sound source is then estimated by calculating the ITD (*Interaural Phase Delay*) through cross-correlation. The sound localization module only works when the robot's facial motors are not working.

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 center the robot's perceptual resources and as a mean of letting the observer know that the robot has intentionality. In CASIMIRO, a simple but flexible and functional attentional model is described. The model fuses both visual and auditive information extracted from the robot's environment, and can incorporate knowledge-based influences on attention.

Basically, the attention mechanism gathers detections of the omnidirectional vision and sound localization modules and decides on a focus of attention (FOA). Although this can be changed, the current implementation sets the FOA to the visual detection nearest to the sound angle. In other cases the FOA is set to the visual detection nearest to the previous FOA, which is a simple tracking mechanism.

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 was integrated in CASIMIRO, it uses color images taken by a color stereo camera placed near the robot's nose. The face detection application is ENCARA [5], which can also detect smiles. As color is its primary source of detection, we had to use the depth map provided by the cameras to filter out distant skin-color blobs that corresponded to furniture, doors, etc. (see Figure 2).

Head nod and shake detection

Voice recognition was not implemented in CASIMIRO. It is estimated that voice recognition errors, dubbed by Oviatt as the Achilles' hell of speech technology, increase a 20%-50% when speech is delivered during natural spontaneous interaction, by diverse speakers or in a natural field environment [12]. The option of making the speaker wear a microphone was discarded from the beginning because it is too unnatural. 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. Head nods and shakes are very simple in the sense that they only provide yes/no, understanding/disbelief, approval/disapproval meanings. However, their importance must not be underestimated because of the following reasons: the meaning of head nods and shakes is almost universal, they can be detected in a relatively simple and robust way and they can be used as the minimum feedback for learning new capabilities.

The major problem of observing the evolution of simple characteristics like intereye position or the rectangle that fits the skin-color 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. The implemented algorithm uses the pyramidal Lucas-Kanade tracking algorithm described in [2]. 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-color blob, observing their evolution and deciding based on what dimension (horizontal or vertical) shows a larger displacement. Figure 2 shows an example of the system.

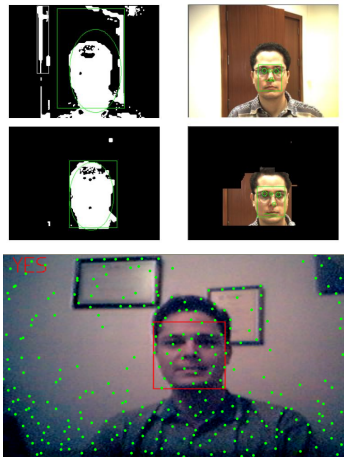

Figure 2: **Top:** Top row: skin color detection. Bottom row: skin color detection using depth information. . **Bottom:** Head nod/shake detector.

Memory and forgetting

In [14] 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 has a memory of the individuals that it sees. Color histograms of (part of) the person's body are used as a recognition technique. Color 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), though that was considered sufficient. The region of the person's body from which histograms are calculated depends on the box that contains the face detected by ENCARA. Intersection was used to compare a stored pair of histograms with the histograms of the current image. Memory will be represented in a list of histogram pairs, with data associated to each entry. Each entry in the list is associated to an individual. Currently, the data associated to the individuals are Boolean predicates like "Content", "Greeted", etc.

Memory is of utmost importance for avoiding predictable behaviors. However, memorizing facts indefinitely leads to predictable behaviors too. Behavioral 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. The first controlled studies of forgetting mechanisms were carried out by Ebbinghaus [6]. 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 [16]. In our system the power law of 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: $f(t) = max(0, 1 - t \cdot \exp(-k))$, where k is a constant. We apply the f function to the set of Boolean predicates that the robot retains in memory. When a predicate is to be forgotten, it takes the value it had at the beginning, when the system was switched on.

Habituation

An habituation mechanism developed by the authors was implemented in CASIMIRO, for signals in the visual domain only, i.e. images taken by the stereo camera. 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 center 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 center of mass. 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. The neck pan and tilt movements produce changes in the images, though it was observed that they are not periodic, and so habituation does not grow.

6 Action Abilities

Facial expression

A three-level hierarchy 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.

The designer 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.

As for the third level in the mentioned hierarchy, facial expressions refer to poses of the different groups, each with a certain degree. Currently, CASIMIRO has the following expressions: Neutral, Surprise, Anger, Happiness, Sadness, Fear and Sleep.

Voice generation

CASIMIRO uses canned text for language generation. 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 was used). 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 TTS, reads the text file when it starts. 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 starting again. The Talk module pronounces phrases with an intonation that depends on the current facial expression. This is done by changing the intonation parameters of the TTS.

7 Behavior

Action selection

CASIMIRO's action selection module is based on ZagaZ [8]. ZagaZ is an implementation 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. The action selection loop in ZagaZ has a period of a few milliseconds for relatively simple networks. It was necessary to introduce a delay of 500 ms on each cycle for the whole system to work well. Behaviors implemented has Boolean inputs like "Frontal Face Detected" which may also correspond to memorized values. The repertory of actions is currently limited to changes in the emotional state (which in turn modifies the displayed facial expression) and commands for talking about something.

Emotions

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 expression. 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.

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 [4]. In our implementation each emotion can have a decay factor associated, by default set at 2.

The emotions that the robot has experienced while interacting with an individual are stored in the memory associated to that individual. 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 center 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 stops working.

8 Conclusions and Future Work

This paper has described the current development status of a robot head with basic interactional abilities. The implementation of social abilities in robots necessarily leads to unrobust behaviour, for those abilities are mainly unconscious to us, as opposed to other mental abilities. The approach taken has been to use the engineering point of view as a main guideline. Our robot does not have a definite task, except to interact with people. Its perceptual abilities include sound localization, omnidirectional vision, face detection, an attention module, memory (treated here as perception because it serves as additional input to the action selection module) and habituation. The robot has facial features that can display basic emotional expressions, and it can speak canned text through a TTS. The robot's behaviour is controlled by an action selection module, reflexes and a basic emotional module.

Future work will include research into the possibility of integrating hands-free speech recognition. This 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.

Acknowledgments

This work was partially funded by research projects *PI2003/165* and *PI2003/160* of Gobierno de Canarias, Consejería de Educación, Cultura y Deportes-Dirección General de Universidades, *TIN2004-07087* of Ministerio de Educación y Ciencia-FEDER.

References

[1] Bernard J. Baars. *A cognitive theory of consciousness*. Cambridge University Press, NY, 1988.

[2] J. Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker. Technical report, Intel Corporation, Microprocessor Research Labs, OpenCV documents, 1999.

[3] Cynthia L. Breazeal. *Designing social robots*. MIT Press, Cambridge, MA, 2002.

[4] T.D. Bui, D. Heylen, M. Poel, and A. Nijholt. Parlee: An adaptive plan-based event appraisal model of emotions. In G. Lakemeyer M. Jarke, J. Koehler, editor, *In Procs. KI 2002: Advances in Artificial Intelligence*, 2002.

[5] M. Castrillon. *On Real-Time Face Detection in Video Streams. An Opportunistic Approach.* PhD thesis, Universidad de Las Palmas de Gran Canaria, March 2003.

[6] H. Ebbinghaus. *Memory. A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York, 1913.

[7] U. Frith. *Autism: Explaining the enigma*. Blackwell, 1989.

[8] D.J. Hernández-Cerpa. Zagaz: Entorno experimental para el tratamiento de conductas en caracteres sintéticos. Master's thesis, Universidad de Las Palmas de Gran Canaria, 2001.

[9] H. Kozima and H. Yano. A robot that learns to communicate with human caregivers, 2001.

[10] A. Martin, P.J. Phillips, M. Przybocki, and C.I. Wilson. An introduction to evaluating biometric systems. *Computer*, 56:56–63, February 2000.

[11] MIT AI lab, Humanoid Robotics Group. Kismet, 2003. `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.ht%ml`.

[12] S. Oviatt. Taming recognition errors with a multimodal interface. *Communications of the ACM*, 43(9):45–51, 2000.

[13] B. Scassellati. *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, MIT Department of Computer Science and Electrical Engineering, May 2001.

[14] J. Schulte, C. Rosenberg, and S. Thrun. Spontaneous short-term interaction with mobile robots in public places. In *Procs. of the IEEE Int. Conference on Robotics and Automation*, 1999.

[15] M.P. Wenger. Noise rejection, the essence of good speech recognition. Technical report, Emkay Innovative Products, 2003.

[16] J.T. Wixted and E.B. Ebbesen. Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition*, (25):731–739, 1997.