CASIMIRO: A Robot Head for Human-Computer Interaction

O. Déniz, M. Castrillón, J. Lorenzo, C. Guerra, D. Hernández, M. Hernández *
Universidad de Las Palmas de Gran Canaria
Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería
Edificio Central del Parque Tecnológico - Campus de Tafira
35017 Las Palmas - Spain
E-mail: {odeniz,mcastrillon,jlorenzo,cguerra,dhernandez,mhernandez}@dis.ulpgc.es

Abstract

The physical appearance and behavior of a robot is an important asset in terms of Human-Computer Interaction. Multimodality is also fundamental, as we humans usually expect to interact in a natural way with voice, gestures, etc. People approach complex interaction devices with stances similar to those used in their interaction with other people. In this paper we describe a robot head, currently under development, that aims to be a multimodal (vision, voice, gestures,...) perceptual user interface. Modules are described for face detection, tracking, facial movement, action selection and sound localization. Preliminary results indicate that the robot head can potentially achieve the goals we are interested in, namely human interaction and assistance.

1 Introduction

A characteristic of our society is the introduction of the computer in daily life, but with devices that are not natural for human beings to interact with [15]. Users normally need a training period to make use of these devices, so in some cases it can appear a rejection to the use of computers due to the unnatural design of the communication devices. This is due to the fact that users must adapt to the computers instead of the opposite. Human beings are sociable by nature and use their sensorial and motor capabilities to communicate with their environment; we communicate not only with words but with sounds and gestures. Therefore, if the man-machine interaction was more similar to the interaction among humans, the access to artificial devices would be higher and they would play a role as assistants.

Perceptual User Interfaces (PUI) [23] is the paradigm that explores the techniques used by the human beings to interact among them and with their environment. These techniques take into account the human capabilities to interact with the technology in order to model the man-machine interaction. This interaction must be multimodal because it is the most natural manner to interact with computers. Raisamo [17] gives a intuitive approach defining a multimodal user interface when "a system accepts many different inputs that are combined in a meaningful way". Thus, in a multimodal system the user interacts with several modalities like voice, gestures, sight, etc. So, multimodal interaction models the study of mechanisms that integrate modalities to improve the man-machine interaction.

In this work we present the architecture and initial development of an experimental multimodal interface. The paper is organized as follows. In Section 2, the architecture of the whole system is described. The modules that are being developed are described in Section 3. In Section 4 some preliminary results are shown. Finally, the main conclusions and future directions of this work will be presented.

2 CASIMIRO architecture

In this section we describe CASIMIRO, an architecture of a Perceptual User Interface which will make easier the interaction between people and computers. This architecture is based on the scheme of semantic fusion and the different modes that compound the system are considered independent. To achieve this goal the interface has human-like behaviors, which are based on a humanoid head (Fig. 4) with facial movements that allows to add gestures as a mean of interaction. The perceptual side of the interface will make use of an active vision approach already used in the DESEO system [10]. Sounds and voice are also elements of the perceptual capabilities of the architecture.

Casimiro is made up of five major modules (Figure

*Work partially funded by DGUI-Gobierno de Canarias *PI2000/042* research project. The first author is supported by graduate grant *D260/54066308-R* of Universidad de Las Palmas de Gran Canaria.

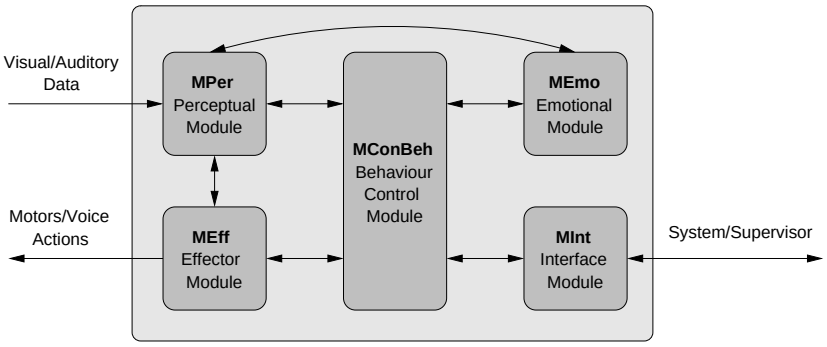

Figure 1: CASIMIRO architecture

1): MPer, Meff, MconBeh, MInt and MEmo. Below we give a brief description of these modules.

- MPer: This module integrates all the activities related to perceptions: visual and auditory. The submodule devoted to the visual perception is conceived according to the paradigm of active perception [4] and realizes visual attention tasks and sensorial abstraction. It takes as input the visual data as well as outputs from MConBeh modules like control commands, and as outputs produces visual abstraction that are sent to the MConBeh module. The visual attention tasks are devoted to detect and track objects and the sensorial abstraction includes activities for face, posture or gesture recognition. Auditory perceptions are processed by a submodule which is included into MPer. The auditory submodule, similar to the visual submodule, includes attention and recognition tasks. This submodule can detect the orientation of a sound sources to direct the visual attention. It should also be able to separate different sound sources to provide the input to voice recognition tasks.

- MEff: The effector activities are associated with the MEff modules, both motor and sound/voice actions. Motor actions control the gestures and the reactive response of the system to activities that come from the MPer module, specifically the actions related with the head and the eyes. Gestures generated by the MConBeh module must be decoded into motor actions to get the desired face expression.

- MConBeh: The core of the perceptual-effector system, it is in charge of the generation of the conductual sequences. This module will be based on Behavior Activation Networks [14], which allow to define relations between behaviors and their activation and activated variables as STRIPS-like rules. It also resolves the behavior transition problem and gives a simple and complete solution to the action selection problem.

- MInt: This module serves as interface with the extern system or with an external supervisor. If CASIMIRO is embedded into another system, like a service robot, this module will be the interface with it.

- Memo: The seminal work of Damasio [7] introduces the idea that the human intelligence relies on emotions in aspects like those related with the decision making activity in dynamic and complex environments. The inclusion of a Emotional Module, whose aim would be to act on the preconditions of the MConBeh module to "tune" them [16, 6], would yield as result a human behavior rather than a pure rational behavior guided only by perceptions from the environment.

3 Modules under development

In this section the major modules that are being developed in the framework of the proposed architecture are briefly described.

3.1 Face detection module: ENCARA

The face detection module is one of the perceptual modules of the whole system. Detecting any possible facial pose at any size is an extremely hard problem and certainly not trivial. Most face detection techniques in the literature [12, 24], perform an exhaustive search for a set of restricted poses and sizes on the image. This task requires a great computational effort, affecting seriously the performance of the system as a whole. None of those systems was conceived as a face detector for real-time video streams, but for still images.

As pointed out in [19], some information is available for improving performance. The authors refer to color information, among other cues, as a tool for optimizing the algorithm which helps to restrict search area, and also providing the advantage of its orientation invariance, its robustness against scale changes, partial occlusion and its fast calculation making it suitable for real time systems. However, color perception can vary substantially for different environments (indoor, outdoor) mainly due to varying lighting conditions [22].

The ENCARA system aims to achieve robustness as a consequence of a combination of weak processes in an opportunistic way. We focused the problem making use of simple techniques applied in a cascade an opportunistic approach, in order to confirm/reject the initial frontal face hypothesis. Those techniques are combined and coordinated with temporal information extracted from the video stream to improve performance. Indeed the process tries to detect first the potential eyes, and once they have been located proceeds performing some confirmation tests. These tests are based on contextual knowledge about face geometry, appearance and temporal coherence in order to validate or refuse the hypothesis that eye positions recovered are coherent for a frontal view. The procedure is briefly described as follows, (a more detailed explanation is available in [20]).

Appearance Temporal Coherence: If a face was detected recently a temporal coherence test is performed. Previous eyes detected are searched in current frame, checking their appearance and position. Thus color stage is avoided, providing a faster test.

Color Blob Detection and Ellipse Aproximation: Normalized red and green color space is used for face detection. Blobs classified as skin coloured are fitted to a general ellipse using the technique described in [21]. Some ellipses are rejected using geometric filters.

Face Orientation: Ellipse fitting also provides an orientation for the face. The orientation obtained is employed for rotating the source image in order to get a face image where both eyes lie on a horizontal line.

Neck Elimination: Face geometric knowledge and heuristics are used to eliminate those blob pixels that are not part of the face, as for example neck. Finally a new ellipse is approximated and the image rotated using the same procedures explained in previous steps.

Eyes Detection: ENCARA searches for eyes in a coherent geometric manner where eyes should be for a frontal face using previous frames information. In this step two eye sets of candidates are obtained, one by using gray levels minima, and another based on correlation, using the patterns provided by the last couple of good eyes detected.

Geometric tests: Some geometric tests (intereye distance, eyes should lie on an horizontal line) are applied to both sets, accepting the best set for a frontal face configuration.

Normalization: A candidate set that verifies all the previous requirements is then scaled and translated to fit a standard size.

Appearance tests: The normalized image and the area around the eyes are projected using PCA. Their reconstruction provides an error that is used as an appearance measure.

The face is considered frontal: For faces that have reached this point, we update eye patterns to use them for detecting eyes with correlation.

3.2 Tracking module

The tracking module implemented keeps a moving target in the centre of the image, supporting the work of other perceptual modules such as the face detection task. Depending on the distance of the person to the camera it can track the whole head or just a part of it, as the eye or the mouth. Although it has been mainly used for faces, the tracking module has been designed to adapt automatically to any kind of object.

The module can use the fusion of several cues to perform a robust tracking, although template matching is the main one. Basically, the module is composed of a variable number of tracking methods, which make use of different cues (template matching, color, Hausdorff distance, etc.). The results of all of them are sent to a combination method that performs a fusion according to a certain algorithm.

Each tracking method is composed of two parts: the descriptors and the tracking method algorithm. The descriptors are the data that the tracking method algorithm uses for searching the object of interest. For example, in the case of a tracking method based in template matching, the descriptors are the templates that the algorithm uses.

3.3 Sound localization

A perceptual module that uses sound is the sound localization system. Our sound localization abilities

stem from the fact that we have two ears. 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 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. Both ILD and ITD cues are dependent on the sound frequency. ITD cues are reliable for relatively low frequencies, while ILD cues are better for higher frequencies (see [8] for an explanation of this). A more detailed description of sound localization mechanisms can be found in [5, 25, 9].

A sound localization system using both ITD and ILD cues was implemented for the robotic head. Tests of the system were carried out with a custom-made plastic head, see Figure 2. The hardware setup included two *Philips Lavalier* omnidirectional microphones (placed 28cm apart), pre-amplifier circuits and a professional sound card (*Terratec EWS88*).

Figure 2: Plastic head used in the experiments, next to the sound card external rack and preamplifiers.

A *DirectX* application was developed to integrate all the necessary processing stages: low-pass filtering, sound source detection, feature extraction, data recording and playing (for off-line analysis), and classifying (see Figure 3).

A number of improvements were introduced in the feature extraction stage, as compared with the baseline system, described in [13]. First, extracted ILD and ITD cues are normalized to take into account the intensity of the input signals. Also, some cues are discarded when there is a possibility that they are wrongly extracted. Other minor modifications were also introduced. In order to study the reliability of the extracted cues, the ratio between inter-class to intra-class variances was used as a goodness measure. A lower overlap ratio was obtained for all the sounds

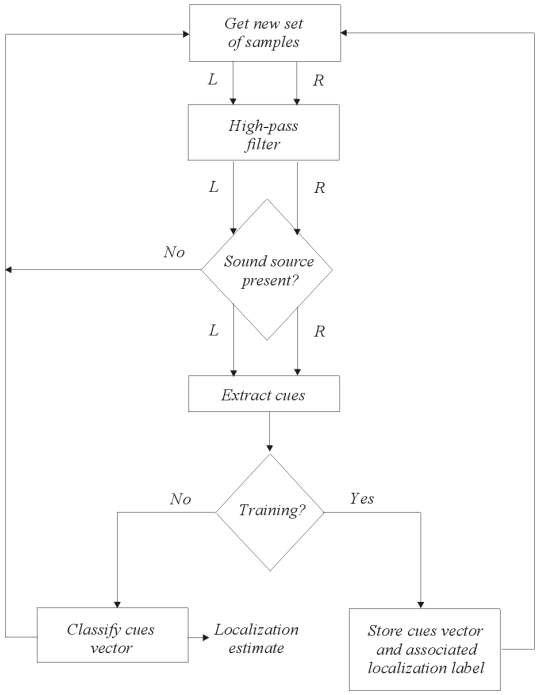

Figure 3: Steps performed by the developed sound localization module.

tested (maraca, mobile phone, hand claps and whistle).

3.4 Facial movements

In this section we address the topic of facial configurations and facial gestures. The goal is to control the 9 servomotors attached to the different facial parts so that they can adopt various recognizable poses, see Figure 4. The facial system adopts a three-layer hierarchy:

Figure 4: Mechanical structure of the robot head.

1. Motor groups: left eyebrow, right eyebrow, mouth, etc.
2. Poses for facial characteristics like eyebrows, mouth, ears, etc.

3. Poses for the whole face, that refer to poses of facial characteristics.

Level 2 and 3 always include a neutral pose. While the system is working it can adopt a pose with a certain intensity degree, for example "smile 50%". The intensity degree refers to a separation from the neutral pose. This framework is very similar to that of other robotic heads, see [2, 1, 3].

An application called Pose Editor was developed for the control of the head. It provides the user with a flexible and easy means of controlling all the motors, create new poses, test them, save the configuration, etc. If the transition between two poses is not visually convincing (motion is linear in motor coordinates), the user can introduce intermediate points, as well as intermediate velocities. A snapshot of the main application window is shown in Figure 5.

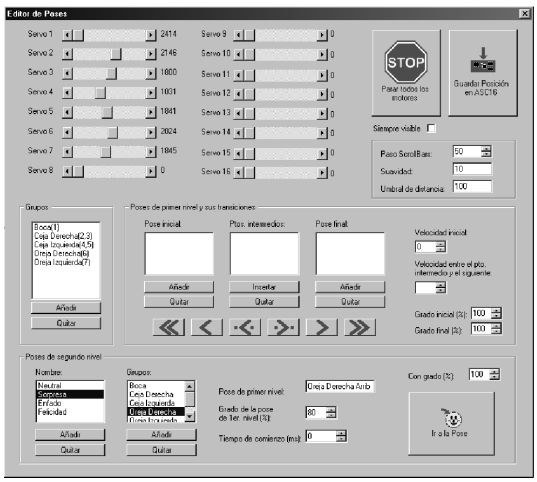

Figure 5: Pose Editor.

3.5 Action selection module

A software tool called ZagaZ [11] was developed for designing, testing and running PHISH-Nets [18], an improved version of Maes' Behavior Activation Networks [14]. It provides an easy and scalable way to create behavior networks that can accept external symbolic inputs and generate simple output signals. Also, a debugger was included that allows the implementer to test the designed network. Network modules (normally simple behaviors) can also be other networks, allowing hierarchical behavior designs. Parameters that "tune" the network can also be easily controlled. This tool was designed to be completely independent of the inputs, outputs, or goals of the robotic system.

Figure 6: CASIMIRO working in front of a user.

4 Preliminary results

As a first experiment, tracking was integrated with the mechanical structure. A simple automaton was used to model transitions among neutral, surprise and angry states, according to the motion dynamics of the object being tracked, see Figure 6. Voice synthesis from a commercial package was also used. This system was able to work satisfactorily in one 733MHz PC.

Sound localization was also tested separately using the plastic head attached to a pan-tilt unit acting as a neck. Although the precision of the direction estimate is still rather limited (only front, right or left), the system is able to produce good decisions in terms of the three basic directions. These experiments were carried out in a cluttered environment (a room with many pieces of furniture) and using the four sound types mentioned before. Classification errors obtained are shown in Table 1.

Classifier	Mean error	Std. dev.
nearest-neighbor	4.14 %	0.15 %
tree classifier using purity (gini value)	1.16 %	0.15 %
Levenberg-Marquardt neural net, 8 hidden units, 500 iterations	3.27 %	0.45 %

Table 1: Classification errors using 3-fold cross validation, 10 runs.

5 Conclusions and future work

In this paper we have described a robot head, currently under development, that aims to be a multimodal (vision, voice, gestures,...) perceptual user interface. The proposed final architecture has been

described, and modules have been presented for face detection, tracking, action selection, facial movement and sound localization.

Future work will include voice recognition, development of the emotional module and the integration of visual and auditory cues for person localization. All the modules will be eventually integrated in two PCs connected through a high-speed local network.

Acknowledgments

Thanks are due to mechanical students Y. Rageul, C. Eymard and O. Mincato, from *IFMA*.

References

[1] An anthropomorphic head robot WE-3RIV. Available at `http://www.takanishi.mech.waseda.ac.jp/eyes/`.

[2] Kismet. Available at `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.ht%ml`.

[3] Minerva: Carnegie mellon's robotic tourguide project. Available at `http://www-2.cs.cmu.edu/~minerva/`.

[4] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, Aug. 1988.

[5] Jens Blauert. *Spatial hearing*. MIT press, Cambridge, MA, 1983.

[6] D. Canamero. Modelling motivations and emotions as a basis for intelligent behavior. In *Procs. Of Agents'97*. ACM, 1997.

[7] A. R. Damasio. *Descartes' Error: Emotion, Reason and and the Human Brain*. Picador, 1994.

[8] GCAT. Perception of direction, 1999. Available at `http://www.gcat.clara.net/Hearing/perception_of_direction.htm`.

[9] William M. Hartmann. How we localize sound. *Physics Today*, 52(11):24–29, 1999.

[10] M. Hernández, J. Cabrera, M. Castrillón, A .C Domínguez, C. Guerra, D. Hernández, and J. Isern. DESEO: An active vision system for detection, tracking and recognition. *Lecture Notes on Computer Science, 1542*, 1999. Springer-Verlag, ICVS'99, Gran Canaria.

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

[12] E. Hjelmas and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3), 2001. Erik Hjelmas and Boon Kee Low.

[13] Robert 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.

[14] P. Maes. How to do the right thing. *Connection Science Journal*, 1(3):291–323, 1989.

[15] Nicholas Negroponte. *Being digital*. Vintage Books, 1995.

[16] R. Picard. *Affective Computing*. MIT Press, 1997.

[17] R. Raisamo. *Multimodal Human-Computer Interaction: a constructive and empirical study*. PhD thesis, Departament of Computer Science, Univ. of Tampere, 1999.

[18] B. Rhodes. PHISH-nets: Planning heuristically in situated hybrid networks. Technical report, MIT Media Lab, 1996.

[19] Henry A. Rowley. *Neural Network-Based Face Detection*. PhD thesis, Carnegie Mellon University, May 1999.

[20] M. Castrillón Santana, J. Lorenzo Navarro, J. Cabrera Gámez, F.M. Hernández Tejera, and J. Méndez Rodríguez. Detection of frontal faces in video streams. In *Post-ECCV Workshop on Biometric Authentication*, 2002. Copenhagen, Denmark.

[21] Karin Sobottka and Ioannis Pitas. A novel method for automatic face segmentation, face feature extraction and tracking. *Signal Processing: Image Communication*, 12(3), 1998.

[22] Moritz Strring, Hans J. Andersen, and Erik Granum. Skin colour detection under changing lighting conditions. In *7th Symposium on Intelligent Robotics Systems*, July 1999.

[23] M. Turk. Proceedings of the workshop on perceptual user interfaces. San Francisco, November 1998.

[24] Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[25] W. A. Yost and G. Gourevitch. *Directional hearing*. Springer-Verlag, New York, 1987.