Multimodal Cognitive System For Immersive User Interaction

Alessandro Calbi
Technoaware srl
Corso Buenos Aires 8/11
16126 Genova - ITALY
tel. +390105539239
alessandrocalbi@technoaware.com

Alessio Dore
DIBE – Università di Genova
Via dell'Opera Pia 11/A
16145 Genova - ITALY
tel. +390103532060
dore@dibe.unige.it

Lucio Marcenaro
Technoaware srl
Corso Buenos Aires 8/11
16126 Genova - ITALY
tel. +390105539239
lucio.marcenaro@technoaware.com

Carlo S. Regazzoni
DIBE – Università di Genova
Via dell'Opera Pia 11/A
16145 Genova - ITALY
tel. +390103532792
carlo@dibe.unige.it

ABSTRACT
In the recent years many efforts have been made to provide machines with the capability of effectively interact with its users. Here a system is introduced that supplies a virtual guide service to users by means of a mobile device (e.g. Palm, tablet PC, etc.). The architecture takes inspiration from a biological model of the cognitive processes, the Cognitive Cycle, performed by the brain while interacting with other entities. A variety of multimodal interfaces communicates messages to interact adaptively with the user. Results show the effectiveness of the different message modalities in real situation where an user is moving towards a target guided by the system.

Keywords
Ambient intelligence, cognitive cycle, immersive.

1. INTRODUCTION
In this paper the architecture of an Ambient Intelligence system is dealt with, which was designed with the aim of providing a virtual guide to users moving in an environment. Many challenges are to be faced to accomplish this task. First of all the person must be localized and tracked in the area where the service is supplied. Then a strategy to select the guidance messages to send is to be defined. Moreover an important issue regards the modality by which the message is provided to the user. This problem is connected to the establishment of a proactive, immersive and pervasive communication between a system and an user who services are directed to. Humans are able to interact one to the other using conjunctively the voice, the facial expression, the body attitude, etc. To simulate this capability in an artificial system a bio-inspired model of how intelligent beings exchange information is necessary. Moreover the suitability of the interfaces that communicate messages to an user in a pervasive and immersive way is another crucial challange.

Mc Ara - Mc William in [3] outlines how human/machine *interaction* is a central problem in Ambient Intelligence. In [4], Anania highlights that the new discoveries in cognitive neuroscience can help to design understanding-based systems for interactive and immersive applications. Jebara and Pentland [5] report an approach to inference interactions between two humans talking to each other by tracking their heads and hands. They also address to the human-machine intercommunication as an emulation of the human behavior.

The development of Ambient Intelligence systems is a very active line of research where both companies and academia are being involved. For example Philips Research [6] is carrying on a long-term project aiming at creating an environment aware of the user location, identity and intentions. The prototypes of the services are tested in a real living space, the HomeLab, where temporary residents stay and interact with the supplied technologies. The MIT's Oxygen project [7] focuses the attention in human-centered computation, that is, designing configurable devices, either handheld or embedded in the environment, that support human activities bringing computation, whenever the user needs it and wherever he/she might be. The key features of these applications are that they have to be reached and used everywhere (*pervasive*), they must be *embedded* in our world sensing and affecting it, *adaptable* to user requirements and operating conditions, *eternal*, i.e. the system must never shut down or reboot and it must provide services regardless upgrades, errors, etc. Another important work aiming to define and accomplish a complete and complex test bed is

IMMERSCOM 2007, October 10-12, Verona, Italy

DOI 10.4108/ICST.IMMERSCOM2007.2259

"Virtual Immersive COMmunication" (VICOM) [8]. This Italian three-year project involves many academic and private partners, and is developing strategies for context-data extraction and management as well as multiple sensor systems and ad hoc networking communication technologies.
The remainder of the paper is organized as follows: in Section 2 a model of the processes performed by the brain while interacting with other entities is described. In Section 3 the bio-inspired proposed system architecture is presented. In Section the interactive capabilities of the 4 system are demonstrated. Finally, in section 5, we conclude.

2. THE COGNITIVE CYCLE

An Ambient Intelligence system should be designed to interact in an effective and unobtrusive way with a human user. The system should then be modeled according to this aim with an user-centered [1] approach in order to focus on the effectiveness and usability of the applications. A first necessary consideration regards the distinction between the Internal and the External World (with respect to the system). For example in this scenario the Internal World are the devices composing the system whereas the External World includes the users and the other entities within the scope of the system. To establish a system able to interact with the user and react to external events two basics capabilities are to be introduced, that is, the *context awareness* and a "*conscious*" reasoning. A system can be provided with these skills only by defining what consciousness is.

With this aim, basing on the work of the neurophysiologist A. Damasio [2], we consider two key players, the intelligent *organism* and the perceived *object* and the *relationships* those players hold in the course of their natural interactions. In this context, the organism can be addressed as the AmI system whereas the object is any entity that gets to be known by the system; the temporal stream of causal relationships between the organism and the object are the contents of the knowledge we call *consciousness*. From this point of view, consciousness lies in the construction of knowledge resulting from two occurrences: 1) the organism is involved in the relation to some objects; 2) the object in the relation causes a change in the organism. The brain devices that are devoted to supervise and manage the internal state of the organism and the relation with an external entities are respectively the Proto Self and the Core Self. The Proto Self computes the information regarding the organism's internal state gathered by the Proto Sensor. Similarly the Core Sensors analyse the External World and provide data to the Core Self.

The Cognitive Cycle is a possible model to represent the behavior of any living being while interacting with the External World. Four phases (see Figure 1) can be identified in this process: 1) sensing, 2) analysis, 3) decision, 4) action. The Sensing stage performs the acquisition of the observations regarding the internal state and the surrounding environment. The internal state comprises all the system parameters that can vary during the functioning (e.g. available bandwidth in communication links, work load on a PC, PC login, open/close status of a door, etc.), while the external data are generated by the object in the system scope. The Analysis stage compute the acquired information to provide to the system an higher level synthetic representation of the context, more suitable for the subsequent steps. In the Decision stage the Cognitive system has to decide which is the most proper action to the received external stimulus; the choice is based on the embedded internal knowledge, the past experience and the current context. Finally, in the Action/Communication stage the modality of interaction with the entity are established and actuated.

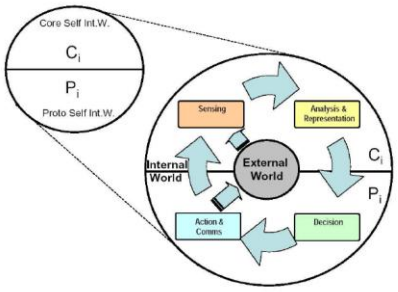

Figure 1 - Representation of the Cognitive Cycle

3. SYSTEM ARCHITECTURE

The proposed system aims at providing a service of virtual guidance to an user, equipped with a mobile device (e.g. PDA, tablet PC, etc.), moving in and environment guarded by cameras. The architecture has been designed with the target of establishing an effective and immersive interaction with the user in order to provide the most suitable information and to

react to eventual user mistakes in following the indications. Then the system interacts with the user by means of guidance messages sent to the mobile device; the user, on the other hand, influences the system decisions with his/her trajectory according to the fact that it is coherent with the path towards the target or not. This process can be represented as a symmetric interaction between the two Cognitive Cycle of the entities (see Figure 2).

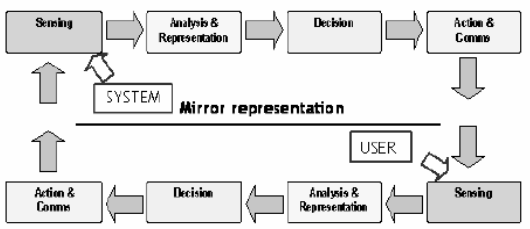

Figure 2: Symmetric representation: user and system modeled by the same four main capabilities.

Different communication modalities can be provided to the user to increase the capability of interaction and the pervasiveness. More precisely, a variety of interfaces (i.e. avatar, 2D guide, 3D guide, augmented reality) are developed to transmit the information in an adaptive, personalized and (if the application allows that) emotional way, resembling the human communication capacities. The avatar guide, for example, by changing its expression and vocal tone can attract the user attention when he/she is moving to the wrong direction, increasing the perception of the message. On the other hand, if the person is going towards the target a continuous stream of messages can be perceived as useless and invasive.

In the following the system architecture is described. It should be noticed that it operates as modeled by the Cognitive Cycle. In fact external data are acquired by video and radio sensors to localize the user. This information is processed and sent to the decision module that, using a priori knowledge of the environment, individuate the direction to follow to reach the target. Finally the message is elaborated in one of the possible modalities to effectively interact with the user.

3.1 PHYSICAL ARCHITECTURE

At the physical level, the architecture of the system includes three types of components interconnected through an ensemble of networks and interfaces: sensors (cameras, gyroscope, etc.), client devices (typically mobile devices as PDA and tablet PC) and a control server.

3.1.1 Network Infrastructure

In Figure 3, the physical components, the network and the communication interfaces are shown.

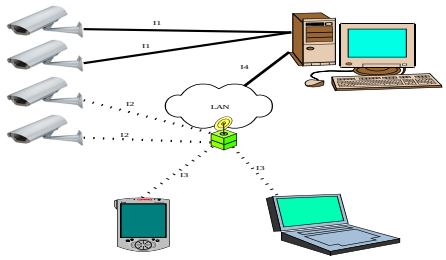

Figure 3 - Network infrastructure of the system

The different components mainly communicate through wireless and wired TCP/IP networks. In particular, a WLAN network composed of four access point IEEE 802.11b/g has been installed and configured. The interfaces can be divided in:

- I1 (RG59): image acquisition from coaxial cable;
- I2 (wireless): image acquisition from IP cameras;
- I3 (wired and/or wireless): communication interface between server and the client.
- I4 (wired): Ethernet interface between server and the access point network; it develops the role of distribution network.
- I5 (wireless): it allows the communication between the mobile clients and it furnishes the support for the WLAN based location.

3.1.2 Sensors

The physical architecture makes use of two fixed cameras and two wireless cameras.

3.1.3 Client

Several heterogeneous client devices (desktop PC, laptop or tablet, PDA) furnish the different functionalities or system services to users.

On every device it is also present a control application that allows the user to select one of the services offered by the system that will be discussed more extensively later in the paper

3.1.4 Server machine

The server machine manages the infrastructural part of the system. In fact, it contains the software modules that analyze the video signal of the cameras,

the control module, an "a priori" ambient information database and a list of the available services.

In the "loosely-coupled" software architecture these elements communicate through XML-PRC protocol allowing, in case, the allocation of these modules on distinct machine to maximize the system scalability.

In particular, the server contains the video tracking module, that processes acquired video images, localizing and identifying moving objects present in the scene. The system maintains a global map of the environment updated with tracked objects position. This information can be integrated with the WLAN localization to allow joint user identification.

The real time video processing is possible by using hardware with high performances (e.g. CPU 3Ghz, RAM 1GB) and a multi-channel frame grabber.

Finally, the server application is provided of a database with an "a-priori" knowledge of the monitored environment connected to the services provided to its clients; this information has an important role also in the real time configurability of the WLAN-based location services.

The server application also allows to start the identity verification software, supposing that the computer is endowed with a webcam pointing at the face of the user.

3.2 LOGICAL ARCHITECTURE

The logical architecture can be divided in four different logical functionalities, as we can see in Figure 4:

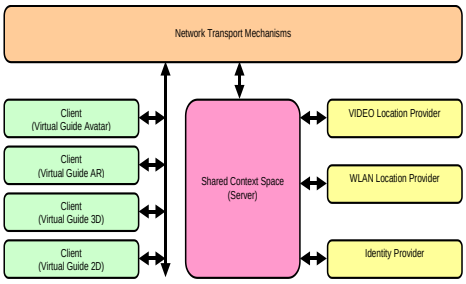

Figure 4 - Logical architecture of the system

1. *Location and identity provider*: it represents the location and identification modules based mainly on the audio, video and radio signals.
2. *Shared context space*: the shared context management logical modules are able to memorize and to organize the information of location and identification inside a common space, and the information related to the targets to reach and the relative paths, as well as the sensor topology.
3. *Network*: it represents the connectivity of the network. The wireless network ensures the coverage of all the selected zones by means of the same access points used by the radio location.
4. *User applications*: it represents all the applications that the user can exploit on his mobile device (PDA or tablet pc): 3D and 2D virtual guide, augmented reality, avatars speaking.

3.3 CONTEXT MANAGEMENT SERVER

The server application has an "a-priori" knowledge of the environment. A real system for this kind of applications would need a real database, such as GIS (Geographical Information System) able to elaborate in real time the guide requests. We have emulated these functionalities using a simple database that encapsulates the principal information related to the ambient structure and with guide functionalities. The modular and expandable structure allows the next integration with a real GIS.

The information of the environment structure are memorized into multilayer maps:

- 2D maps;
- 3D maps and models
- Radio maps.

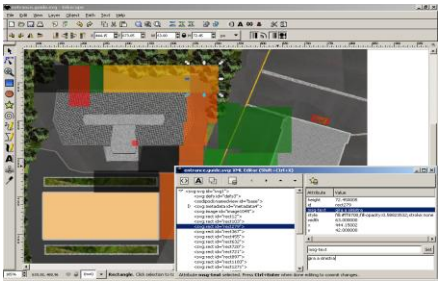

Figure 5 – SVG map of the environment.

In Figure 5 is shown a map realized defining a data format based on XML/SVG, that allows to view and to modify the guide information using standard graphical editors. This representation also allows to represent guide indications on macro-area inside the area represented by the map. In each area all the

attributes necessary to permit the realization of an augmented guide service and a virtual guide service assisted by an Avatar are defined; in particular:

- a bi-dimensional image represents the area where the user has to be guided;
- one or more target where the user has to be guided when he comes in the interest area;
- an high number of interest sub-areas;
- each sub-area allows the reference direction to follow (used by the visual guides and by the augmented reality guide);
- the definition of the position and of the characteristics of the objects.

4. RESULTS: MULTIMODAL SERVICES

Several applications able to guide the users into the environment has been developed. In the next paragraphs these services are described:

1. User Identity verification;
2. User localization (camera, WLAN);
3. Virtual guidance:
    - 3D guide;
    - Augmented Reality guide;
    - Avatar guide;
    - 2D multimodal guide.

4.1 User Identity Verification

This module (see Figure 6) allows the user identity verification using face recognition algorithms like PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) or using a combination of these methods [9].

Figure 6 - User identity verification module

4.2 User localization

This module performs the image acquisition and elaboration and allows the real-time automatic comprehension. The video processing is based on a modular architecture and integrates different functionalities to obtain a reliable analysis of the events, like: image acquisition, calibration, change detection, motion detection, corner and lines extraction, background updating, shadows removal, object classification, tracking, trajectory and behavior, etc. Finally, a multi-sensor data fusion (consisting in three steps: alignment, association and state estimation) fortifies the system in case of occlusions.

Figure 7 - Video localization

4.3 Virtual guidance

4.3.1 3D guide

A three-dimensional virtual ambient reproduces in detail the rooms and the interesting zones of the ambient intelligence system. The user can select his preferred viewer modality: the developed software (see Figure 8) allows the choice of the guide visualization from different points of view, both fixed (camera, top, home view) and mobile (first person and follow person). It is also possible to enable the optimal path and the view of all the moving objects into the scene.

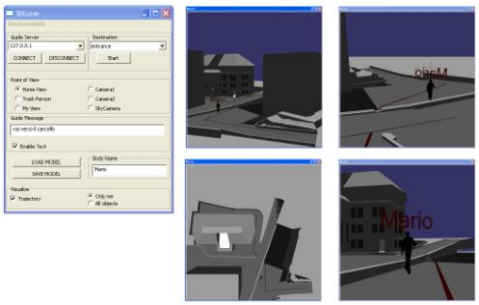

Figure 8 - 3D guide interface with different points of view (home, camera, top and user view). The red line indicates the path to follow to reach destination.

4.3.2 Augmented Reality guide

This module allows users, equipped with a webcam and a gyroscopic sensor, to receive a guide message

like an arrow that indicates the direction to follow to reach the selected destination (Figure 9-a). The arrow orientation is updated in real time and shows in a fluent and coherent representation the orientation changes.

(a) (b)

Figure 9 – a) Augmented Reality Guide interface; b) Avatar Guide in execution on HP iPaq 3970

4.3.3 Avatar guide

This guide modality exploits a semi-conventional interaction model where the user interacts with the application using mouse and keyboard and receives information in a natural language through voice, facial expressions, eyes movement.

A *Multimodal Synthesis Engine* generates the animations and the Avatar voice, and send them to the *Animation Player* that reproduces the guide messages in real time (see Figure 9-b).

4.3.4 2D multimodal guide.

A 2D multimodal guide allows the user to receive instructions from one or more modalities (map, text, voice, Avatar), also simultaneously.

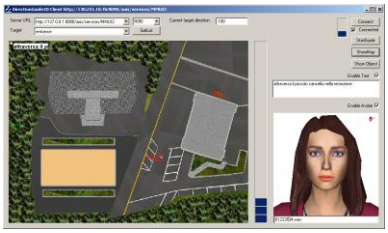

Figure 10 - Multimodal 2D guide with map, arrow, text and avatar choices.

In this manner the interface facilities the info comprehension permitting to the user the selection of the more appropriated communication modalities to his/her preferences and situations [10].

5. CONCLUSIONS

In this paper a system using advanced interaction modalities hiding to users the pervasive Ambient Intelligence structure has been presented. In this environment, innovative interfaces permit really multimodal interaction modalities like 3D guide, 2D guide, Avatar or Augmented Reality.

6. ACKNOWLEDGMENT

This work has been partially supported by the project Virtual Immersive COMmunication (VICOM) founded by the Italian Ministry of University and Scientific Research (FIRB Project). Special acknowledgments go to the University of Cagliari – DIEE and to the University of Genova - DIST which contributed to face verification module and to 2D and Avatar modules respectively.

7. REFERENCES

[1] ISO 13407:1999, "Human-centred design processes for interactive systems," Tech. Report, International Organization for Standardization, 2004.

[2] A. Damasio, "The Feeling of What Happens: Body and Emotion in the Making of Consciousness", William Heinemann:London, 1999.

[3] I. Mc Ara-Mc William, "Foreword" in P. Remagnino, G.L. Foresti T. Ellis eds., Ambient Intelligence, A Novel Paradigm, Springer, USA, ISBN 0-387-22990-6, 2005, pp. xi-xiii.

[4] L. Anania "Preface", in Riva, G., Davide, F, Vatalaro, F. and Alcañiz, M., Ambient Intelligence: The Evolution of Technology, Communication and Cognition Towards the Future of Human-Computer Interaction, Amsterdam, IOS Press, 2004.

[5] T. Jebara and A. Pentland. Action Reaction Learning: Automatic Visual Analysis and Synthesis of Interactive Behaviour. International Conference on Computer Vision Systems (ICVS), 1999

[6] E. H. L. Aarts, and B. Eggen (editors) (2002), Ambient Intelligence in HomeLab, Neroc, Eindhoven., The Netherlands.

[7] Eric S. Brown, "Project Oxygen's New Wind", *Technology Review*, December 2001.

[8] VICOM Project [Online] Available: http://www.vicom-project.it/

[9] G.L. Marcialis and F. Roli, "Fusion of PCA and LDA for Face Verification", Proc. of Post-ECCV Workshop on Biometric Authentication (BIOMET2002), M. Tistarelli, J. Bigun and A.K. Jain Eds., Copenhagen, Denmark, June 2002, LNCS 2359, pp. 30-37

[10] C. Bonamico, F. Lavagetto, C. Regazzoni, L. Marchesotti, "Video Processing and Understanding Tools for augmented Perception and mobile user interaction with in smart Spaces", International Journal on Image and Graphics (Vol. 5, No. 3) (679), July 2005