

INTERACTIVE AGENTS FOR MULTIMODAL EMOTIONAL USER INTERACTION

Eva Cerezo, Sandra Baldassarri, Francisco Seron
Advanced Computer Graphics Group (GIGA)
Computer Science Department
Engineering Research Institute of Aragon(I3A), University of Zaragoza, Spain
{ecerezo, sandra, seron}@unizar.es

ABSTRACT

In this paper, Maxine, a powerful engine to develop applications with embodied animated agents is presented. The engine, based on the use of opensource libraries, enables multimodal real-time interaction with the user: via text, voice, images and gestures. Maxine virtual agents can establish emotional communication with the user through their facial expressions, the modulation of the voice and expressing the answers of the agents according to the information gathered by the system: noise level in the room, observer's position, emotional state of the observer, etc. Moreover, the user's emotions are considered and captured through images. For the moment, Maxine virtual agents have been used as virtual presenters and, a specific application, MaxinePPT, has been developed to allow non-programmers to create 3D presentations easily from classical PowerPoint presentations. Nevertheless, other applications are also envisaged.

KEYWORDS

emotional interface, multimodal interaction, virtual agent, ambient intelligence, virtual worlds.

1. INTRODUCTION: NEW TOOLS FOR SUPPORTING NEW INTERACTION FORMS

Currently, most research on social interfaces is related to the design of embodied conversational agents - ECAs- (Casell et al 2000). ECAs are agents that are visible in the interface sometimes as an animated talking face, may be displaying facial expressions and, when using speech synthesis, with lip synchronization, and sometimes they have 3D graphical representation, with complex facial and body movements. These virtual characters are being used in a wide range of contexts (Mignonneau and Sommere 2005), including education and learning (Boff et al 2005, Graesser et al 2005), therapy (Marsella et al 2000), persuasion (Rosis et al 2003, Berry et al 2005), marketing and entertainment (El-Nasr et al 1999, Yuan and Chee 2005), among others.

These computational agents should show affective and expressive behaviors (Burlison et al 2004): affective expressions have been argued to be useful to help make agents "believable", expressive behaviors have additionally been associated with useful outcomes such as making agents likeable. Moreover, making agent expressions responsive to human expressions, contributes to make agents "relational," able to construct long-term social-emotional relationships with users. The general vision is that if a user's emotion could be recognized by a computer, human-computer interaction would become more natural, enjoyable, and productive (Prendinger and Ishizuka 2005). The computer could offer help and assistance to a confused user or try to cheer up a frustrated user, and hence react in ways that are more appropriate than simply ignoring the user's affective state as is the case with most current interfaces.

Our research concerns interfaces that employ embodied agents that support emotional and multimodal and interaction. By emulating multimodal human-human communication and displaying social cues including synthetic speech, communicative gestures, and the expression of emotion, our characters may implement the "computers as social actors" metaphor (Burlison et al 2000) and be useful in many different scenarios. We present their use as virtual presenters though other uses as domestic helpers or educational tutors are being explored.

In the system developed, called Maxine, the virtual agent is endowed with the following differentiating features:

- it supports interaction with the user through different channels: text, voice, peripherals (mouse, keyboard), which makes the use of the generated applications available to a wide range of users, in terms of communication ability, age, etc.
- it gathers additional information on the user and the environment: noise level in the room, position of the user to establish visual contact, image-based estimate of the user's emotional state, etc.
- it supports voice communication with the user in natural language and in Spanish.
- it has its own emotional state, which may vary depending on the relationship with the user and which modulates the presenter's facial expressions, the answers it gives and the modulation of its voice.

The paper is organized as follows. Maxine engine is described in Section 2, in which particular attention is paid to the system input/output management that makes multimodal interaction possible. Section 3 presents a specific application of the engine for the development of presentations to be made by virtual presenters. And finally, in Section 4, the conclusions are presented, together with a description of current and future work.

2. DESCRIBING THE MAXINE ARCHITECTURE

The overall architecture of our system is shown in Figure 1. Within this section, all the modules that conform Maxine are explained: the Sensory/Perception Modules, that process the inputs of the system, the Deliberative/Generative Modules, in charge of managing the appropriated reactions according to the inputs, and the Motor Module, that generates the final outputs of the system. In this work we are going to focus on the system inputs and outputs.

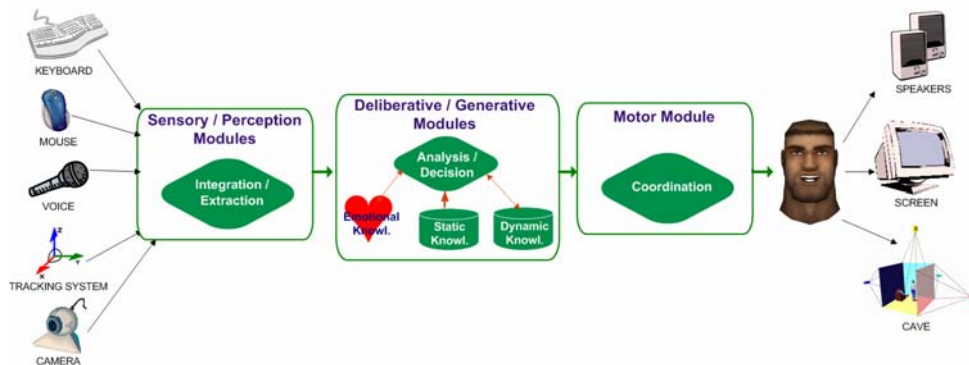


Figure 1. Maxine's Architecture

2.1 Inputs: Sensory/Perception Modules

During the development of the system, special attention was paid to creating multimodal user interaction, via text, voice, image and movement. This broadens the number of potential users of the system by making interaction with disabled users (for example hearing-impaired or paraplegics) and people of different ages and with different levels of education (people with or without a knowledge of computers) possible. While reasoning based on a user's directly input behaviors is important and useful, it is also limited. Therefore, an endeavour is also made to collect the largest possible amount of information on the user by means of body language or facial expression, without requiring him or her to enter data. The ultimate aim is to enhance interaction and establish emotional communication between the user and the virtual character. The sensory/perception modules integrate all the information coming from inputs to the system.

Interaction via console (keyboard) / mouse

Advanced users can fully control the scene thanks to the scripting language used (Lua). For non-programmer users, it is also possible to associate the execution of a command to the pressing of a certain key or clicking the mouse and, due to the power of some of the functions available for the elements and, in particular, of the of scripting language used, the options are very varied.

Voice Interaction

The user formulates an order, a question or any sentence in natural: the sound or audio generated by the user is picked up by the microphone and the sound card (Figure 2). The aim of this first stage is to obtain a text chain from the words said by the user in Spanish. To do this, a voice recognition engine has been constructed on the basis of Loquendo ASR (Audio Speech Recognition) software. The ASR is based on a dynamic library that enables a recognition device to be created and integrated ad hoc; however, the disadvantage is that it has to be developed largely from scratch; in particular it is necessary to pick up the audio and prepare it for processing by the recogniser, develop a grammar with the words that are going to be recognised, and process the results returned by the recognition function. One of the requisites of our system is that it must be able to "understand" and speak Spanish. This constraint prevented us from using existing opensource libraries, all of them in English.

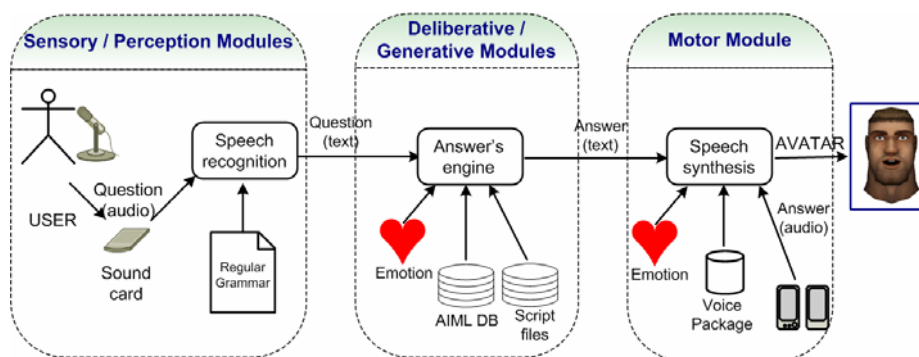


Figure 2. Stages of the user-avatar voice communication process

Image Interaction

A webcam takes pictures of the user's face. The aim of these pictures is to obtain additional information on the user and, in particular, on his or her emotional state. An overview of the stages in the image interaction process can be observed in Figure 3.

We work with Ekman's emotional classification (Ekman 1999), which distinguishes between six basic emotions: happiness, sadness, anger, fear, surprise, disgust, plus a neutral category. The emotional classification system developed is based on the detection of 10 points and five distances, as can be seen in Figure 4.

The face classification implemented is an effective method based on the theory of evidence and has been presented elsewhere (Cerezo and Hupont 2006). From the five distances extracted from the user image and from a series of thresholds defined from the analysis of a sufficiently broad image database, the classification results are acceptable, and range from a 63% success rate in recognising surprise to a rate of 100% in the case of happiness or disgust. For the moment it requires manual operation, but we are working on its total automation.

The information extracted is used for "mirroring": the virtual character is instructed to visually mirror the user's emotional state.

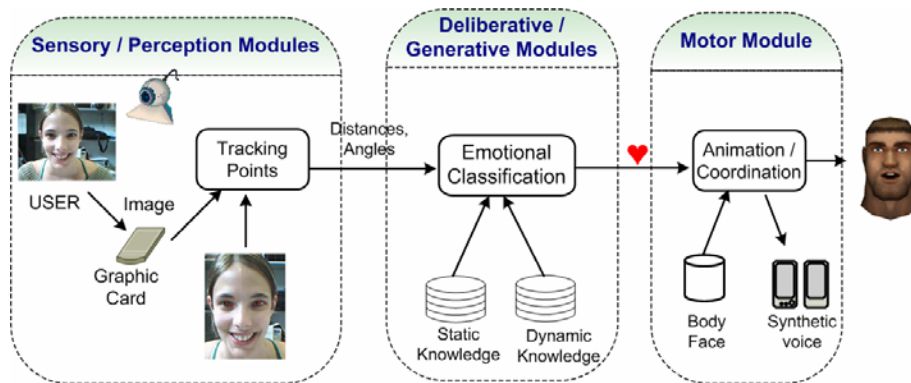


Figure 3. Stages of the user-avatar image interaction process

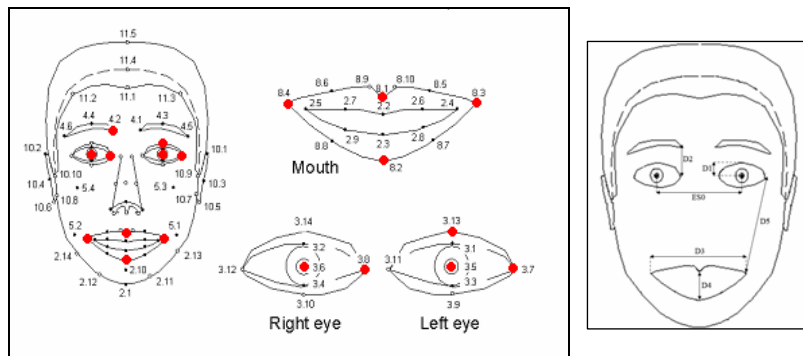


Figure 4. User Face Tracking. The points we work on correspond with the MPEG4 standard points (left). They are used to calculate five basic distances (right) and to obtain emotional information about the user.

2.2 Agent Reactions: the Deliberative and Generative Modules

After extracting input information, the system must manage the appropriated reactions. Two kinds of actions are distinguished:

- Purely reactive: for example, if the user keys in something, the virtual presenter interrupts the presentation; if the user's position changes, the presenter's look/orientation changes; if a lot of background noise is detected, it requests silence, etc. These reactions are managed in the generative module.
- Deliberative: the choice of the reaction of the virtual character calls for more complex analysis. This analysis is done in the deliberative module, which is in charge of obtaining an answer from the user in voice interaction.

The dialog's engine: Getting the right answers

It corresponds to the second stage shown in Figure 2 and it is basically in charge of generating the answer to the user's questions in text mode; its is based on the recognition of patterns, to which fixed answers are associated (static knowledge). These answers, however, vary depending on the virtual character's emotional state, or may undergo random variations so that the user does not get the impression of repetition if the conversation goes on for a long time (dynamic knowledge). The development of this part of the system was based on chatbot technology under GNU GPL licences: ALICE (Alice 2007) and CyN (CyN Project). However, CyN is only designed to hold conversations in English, so we had to modify the code to enable it to support dialogues in Spanish.






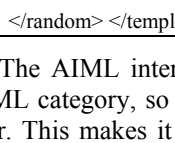
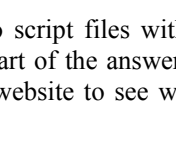
The knowledge of the virtual character is specified in AIML -Artificial Intelligence Markup Language- (AIML 2001). The power of AIML lies in three basic aspects:

- AIML syntax enables the semantic content of a question to be extracted easily so that the appropriate answer can be given quickly.

- The use of labels to combine answers lends greater variety to the answers and increases the number of questions to which an answer can be given.
- The use of recursion enables answers to be provided to inputs for which, in theory, there is no direct answer.

In the conversation emotions are taken into account not only in the voice synthesis but also in the generation of the answers, at two levels:

1. The answer depends on the avatar's emotional state. For this reason, the AIML <random> command has been redesigned to add this feature. As can be seen in the example below, there may be more than one answer with the same label; in this case, one of these answers would be given at random. There must always be an answer (applied to neutral emotional state) that does not have an associated label.
2. Besides, the emotional state of the virtual character may change during a conversation, depending on how the conversation develops. That is, if the conversation is on a topic that pleases the character, it gets happy; if it is given information it was not aware of, it is surprised; it is insulted, it gets angry; if it is threatened, it gets frightened, etc.

<pre> <category> <pattern> I BELIEVE THAT WE WOULD HAVE TO LEAVE THIS CONVERSATION </pattern> <template> <random> <sad/>Well, I suppose I am not what you expected </pre>		
	<pre> <angry/>But is the matter with you? Don't you like me? </pre>	
	<pre> <surprised/> Why? We were having such a good time! </pre>	
	<pre> <happy/>Come on, let's talk a little more Ok, we will continue another time </random> </template> </pre>	

The AIML interpreter has also been modified to include commands or calls to script files within the AIML category, so these commands are executed and their results are returned as part of the answer to the user. This makes it possible, for example, to consult the system time, log on to a website to see what the weather is like, etc.

2.3 Outputs

2.3.1 Facial and Body Animation

The skeletal animation technique is used for both facial and body animation. The animations with which the system works come from two sources: animations from motion capture (the group has got access to an optical system) and animations generated by means of commercial software. In both cases, these animations are loaded in the scene via script file. Regarding facial animation, we work with the six basic expressions (see Figure 5) defined by Ekman and the nomenclature followed is that of the VHML standard [22]. In all the cases, animation blending was achieved with the help of the Cal3D library (Cal3D 2006).

2.3.2 Speech Synthesis

The synthesis of the voice is made using SAPI, but the function uses packages of Spanish voice offered by Loquendo. SAPI allows to obtain the information about the visemes that take place pronouncing the phrase that is wanted to be synthesized, what allows to solve the problem of the labial synchronization. The voice

generated by text-voice converters usually sounds artificial, which is one of the reasons why avatars tend to be rejected by the public. To succeed in making the synthesiser appear “alive”, it is essential to generate voice “with emotion”. In our system we work, again, with the six universal emotion categories of Ekman: joy, sadness, anger, surprise, disgust and fear, plus the neutral one. SAPI enables tone, frequency scale, volume and speed to be modified, which is why we have used it as a basis. To represent each emotion, fixed values are assigned to the parameters that enable the relevant emotion to be evoked. The configuration of these emotional parameters is based on the studies of Boula et al 2002, Francisco et al 2005 and Iriondo et al 2000. The process carried out to find the values at which these parameters must be fixed in order to obtain each of the six emotions was voice assessment by users.

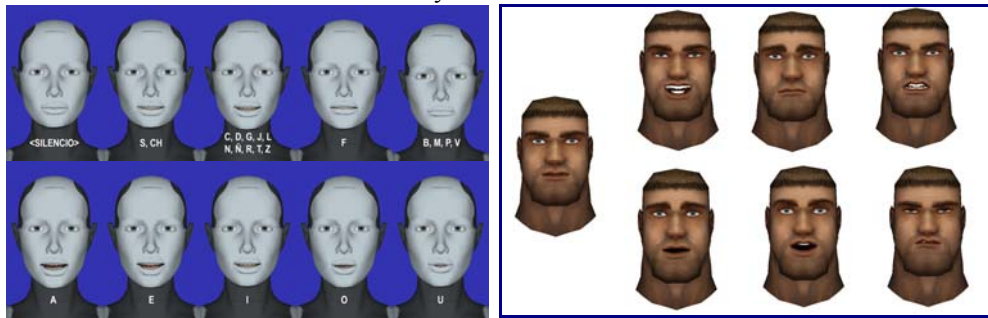


Figure 5. Visemes designed for Lip-Sync in Spanish (left).
Neutral face plus six facial emotions: happiness, sadness, anger, fear, surprise (right)

2.3.3 Expression/Speech Coordination: Lip-sync

A lip-sync module specially developed for Spanish language has been implemented. In Figure 5 the visemes used for synchronization in Spanish are shown, whereas in Figure 6 a user talking to an agent is shown.



Figure 6. User talking to one of Maxine’s agent

3. A MAXINE APPLICATION: MAXINE PPT

The system described in the previous sections has been used in a specific application that allows PowerPoint information to be presented through a like-life character on a graphic display. In this way, the traditional human presenter in Power Point presentations is replaced by a 3D virtual character. This kind of presenter has demonstrated to be especially useful when the same presentation has to be repeated several times or given in a different language (for example in English by a non-fluent English speaker).

The most important features of the MaxinePPT application are:

- It is capable of creating and performing a virtual presentation in a 3D virtual scenario enriched with virtual actors and additional information such as videos, images, etc. from a classical PowerPoint file.
- All the aspects of the virtual presentation are controlled by an XML-type language called PML (Presentation Markup Language). The PML instructions are added to the page notes of the PowerPoint slides in order to determine, for example, the text to be spoken by the avatar.

- Once the presentation has been created, the presentation is performed automatically and verbally, either in Spanish or English, by using a virtual character specially created by the user, or by the default avatar included in the system.

Figure 7 gives an overview of the process involved in creating a presentation.

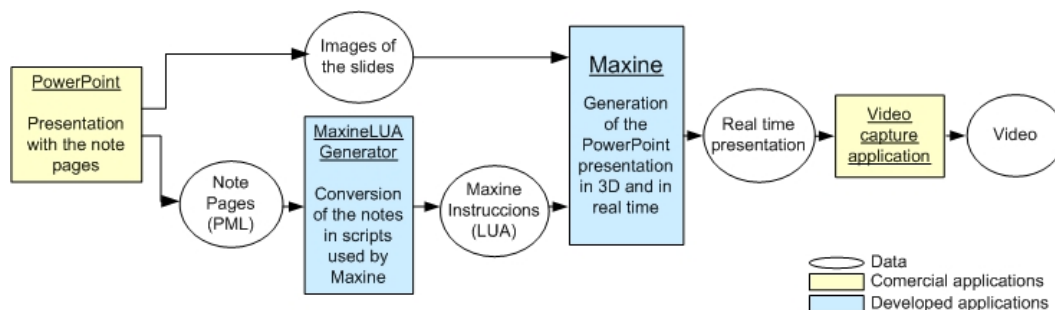


Figure 7. Overview of the Process for Generating a 3D Presentation

The system has been used for the moment in the academic domain to present the students the activities of the Advanced Computer Graphics Group and different applications in the Computer Graphics field. In Figure 8, some screenshots showing one of the presentations are shown. The user can stop the presentation, select specific parts and ask about certain topics.

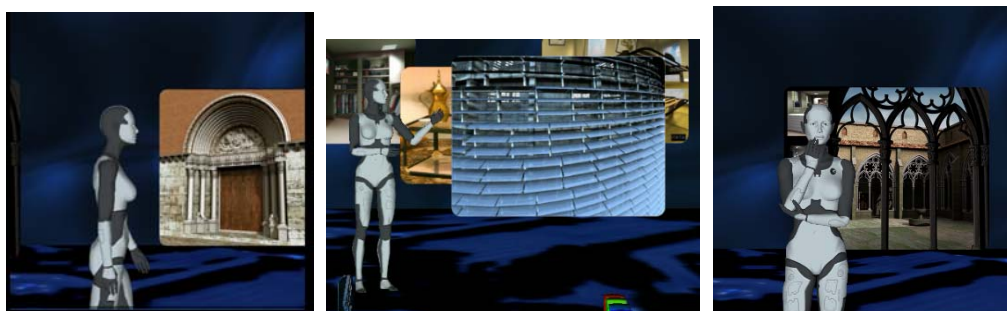


Figure 8. Some Screenshots from the Presentation performed by a virtual agent

4. CONCLUSIONS AND FUTURE WORK

Maxine, a powerful engine for managing virtual environments and agents was presented. The system allows the development of new applications where interaction is based on virtual agents supporting multimodal and emotional interaction. Special emphasis has been done in capturing the user's emotion through images, and in synthesizing the emotion of the virtual agent through its facial expressions, the modulation of its voice and its answers. A specific application to show some of its potential has also been presented (MaxinePPT). The authors are now working on a new applications for Maxine characters, in particular as domotic helpers.

Several other research lines also remain open, most of which focus on enriching interaction between the virtual character and the user:

- to consider not only emotion but personality models for the virtual character
- to give the system learning mechanisms, so that it can modify its display rules based on what appears to be working for a particular user, and improve its responses while interacting with that user
- proper validation of Maxine system and characters by their users.

ACKNOWLEDGMENT

This work has been partially financed by the Spanish "Dirección General de Investigación" (General Directorate of Research), contracts number N° TIN2004-07926 and TIN2004-07672-C03-03 and by the Regional Government of Aragon through the WALQA agreement (ref. 2004/04/86).

The authors wish to thank Sergio García Masip, David Anaya and Isabelle Hupont for their collaboration in this work.

REFERENCES

- AIML, 2001. Artificial Intelligence Markup Language Version 1.0.1, A.L.I.C.E. AI Foundation Working Draft <http://www.alicebot.org/TR/2001/WD-aiml/>
- ALICE, 2007 Artificial Intelligence Foundation <http://www.alicebot.org/>
- Berry, D. et al, 2005. Evaluating a realistic agent in an advice-giving task. *In International Journal in Human-Computer Studies*, N° 63, pp. 304-327.
- Boff, E. et al, 2005. An affective agent-based virtual character for learning environments. *Proceedings of the Workshop on Motivation and Affect in Educational Software, 12th International Conference on Artificial Intelligence in Education*. Amsterdam, Holland, pp 1-8.
- Boula, P. et al, 2002. Generation of Emotions by a Morphing Technique in English, French and Spanish. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, pp. 187-190.
- Burleson, W. et al, 2004. A Platform for Affective Agent Research. *Proceedings of the Workshop on Empathetic Agents, International Conference on Autonomous Agents and Multiagent Systems*, New York, USA,
- Cal3D, 2006, <https://gna.org/projects/cal3d/>
- Casell, J. et al (eds), 2000. *Embodied Conversational Agents*. MIT Press, Cambridge, USA.
- Cerezo E. and Hupont, I., 2006. Emotional Facial Expression Classification For Multimodal User Interfaces. *Lectures Notes In Computer Science* N° 4069, pp. 405-413.
- CyN Project, <http://www.daxtron.com/cyn.htm>
- Ekman, P., 1999. *Facial Expression, The Handbook of Cognition and Emotion*. John Wiley et Sons.
- El-Nasr, M. S. et al, 1999. A PET with Evolving Emotional Intelligence. *Proceedings of the 3rd Annual Conference on Autonomous Agents*. Seattle, USA, pp. 9 – 15.
- Francisco, V. et al, 2005. Análisis y síntesis de expresión emocional en cuentos leídos en voz alta. *Proceedings of UCAmI 2005*, Granada, Spain, pp.353-360.
- Graesser, A. et al, 2005. AutoTutor: An Intelligent tutoring system with mixed-initiative dialogue. *In IEEE Transactions on Education*, Vol. 48, N° 4, pp. 612-618.
- Iriondo, I. et al, 2000. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. *Proceedings of the ISCA Workshop on Speech and Emotion*. Northern Ireland, pp.161-166.
- Marsella S. C et al, 2000. Interactive Pedagogical Drama. *Proceedings of the 4th International Conference on Autonomous Agents*. Barcelona, Spain, pp. 301–308.
- Mignonneau, L. and Sommerer, C., 2005. Designing emotional, metaphoric, natural and intuitive interfaces for interactive art, edutainment and mobile communications. *In Computer & Graphics*, Vol. 29, pp. 837-851.
- Prendinger, H. and Ishizuka, M., 2005. The Empathic Companion: A Character-Based Interface that Addresses Users' Affective States. *In Applied Artificial Intelligence*, Vol.19, pp.267–285.
- Rosis, F. et al, 2003. From Greta's mind to her face: modelling the dynamics of affective status in a conversational embodied agent. *In International Journal of Human-computer Studies. Special Issue on Applications of Affective Computing in HCI*, Vol 59, pp 81-118.
- Yuan, X. and Chee, S., 2005. Design and evaluation of Elva: an embodied tour guide in an interactive virtual art gallery. *In Computer Animation and Virtual Worlds*, Vol. 16, pp.109-119.