Abstract

This paper presents a powerful animation engine, called Maxine, to develop applications with embodied animated agents. The engine, based on open source tools, allows the management of scenes and virtual characters, and allows multimodal and emotional interaction with the users. The virtual actors are provided with facial expressions, lip-synch, emotional voice, and can vary their answers depending on their emotional state and on the relationship with the user along the conversation. Maxine virtual agents have been used as virtual presenters in several applications: MaxinePPT is an specific application, developed to allow non-programmers to create 3D presentations easily from classical PowerPoint presentations; a virtual character was used as an interactive interface to communicate and control a domotic environment; and finally, an interactive pedagogical agent was used to simplify and improve teaching and practice of Computer Graphics subjects.

Keywords. virtual worlds, animated characters, natural interaction, multimodal interfaces

1. Introduction

Current research into ambient intelligence deals with multimodality and social interaction. Moreover, most research on social interfaces is related to the design of embodied conversational agents (ECAs) [8]. 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 body movements, etc.

Virtual characters endowed with these new features can be used in a wide range of contexts [21] [7], including education and learning [18] [14] [5] [15] [22], sign language interpretation [24], therapy [20], persuasion [25] [4], entertainment [27] [29], among others.

In order to obtain a more natural and trustworthy interaction, the virtual agents must be capable of responding appropriately to the users with affective feedback [11]. So that, within the ECAs, our research is centered in interactive virtual agents that support multimodal and emotional interaction, for establishing a more effective communication with the user. Special emphasis is done in capturing the user’s emotion through images, and in synthesizing the emotion of the virtual agent through its facial expressions and through the modulation of the voice.

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. The emotional state modulates the presenter’s facial expressions, the answers it gives and the modulation of its voice.
The paper is organized as follows. An overview of the animation engine, Maxine, is described in Section 2. Section 3 presents the system’s input and their management by the sensory/perception modules, while Section 4 presents the possible agents reactions through the deliberative and generative modules. Section 5 exposes the motor module and the system’s outputs. In Section 6, different applications developed with our engine are presented. Finally, in Section 7, some conclusions and future work are commented.

2. Overview of Maxine: the animation engine

Maxine is a script-directed engine for the management and visualization of 3D virtual worlds. In Maxine it is possible to load in real-time models, animations, textures, sounds, etc. as they are needed in the virtual representation. Even though it is a very generic engine, it has been oriented to the work with characters in virtual scenarios. It has the advantage of allowing the specification of contents through an editor but it also makes possible the direct communication with the low-level tools. It has been written in C++ and employs a set of open source libraries. The engine manages scene graphs that can be built in real-time, dynamically creating and manipulating its elements by means of a simple command interface. These commands can be executed via script-files when initiating the application or during execution, or can be introduced through the text console every time. The overall architecture of our system is shown in Figure 1. In the following sections, all the modules that conforms Maxine are explained in detail: 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. The philosophy behind the generation of our system was to use Opensource libraries as far as possible so that we could concentrate our efforts on truly interesting and challenging problems, especially those related to emotional and multimodal user interaction. Table 1 shows the main libraries used, with a specification of their functionality and a brief reference to their purpose within Maxine.

3. System’s Inputs: The 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 knowledge of computers). 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.

![Figure 1. Maxine's Architecture](image-url)
3.1. Interaction via console / mouse

Advanced users can fully control the scene thanks to the scripting language used (LUA, see libraries in Table 1). 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. Due to the power of some of the functions available for the elements and, in particular, of the scripting language used, the options are very varied.

3.2. Audio Speech Recognition (ASR)

The user formulates an order, a question or any sentence in natural language that might be used in a conversation between people. The sound or audio generated by the user is picked up by the microphone and the sound card, as can be seen in Figure 2.

One of the main 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. So, in order to get a text chain from the words said in Spanish by the user, a voice recognition engine has been constructed using the commercial Loquendo ASR software [19]. However, during the development of the recogniser, some problems that are specific to Spanish had to be solved: specifically, Loquendo ASR is not capable of distinguishing between words with or without ‘h’ (this letter is not pronounced in Spanish), with ‘b’ or ‘v’, or with ‘y’ or ‘ll’ (these letter pairs apply to single phonemes). Neither is it useful when it comes to detecting the difference between words that are written in the same way but pronounced differently, where the difference is marked by an accent, such as, for example; the verb form “está” and the pronoun “ésta”. We had to take all these factors into account when writing the words of the JSGF (Java Speech Grammar Format) grammar and when generating the AIML files [2] with the different possible answers.

<table>
<thead>
<tr>
<th>Function</th>
<th>Library</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scripting</td>
<td>Lua and LuaBind</td>
<td>Real-time creation and manipulation of objects</td>
</tr>
<tr>
<td>Animation</td>
<td>Cal3D <a href="http://alpp.sourceforge.net/">http://alpp.sourceforge.net/</a></td>
<td>Local animation of virtual characters (modified to adapt it to Maxine needs)</td>
</tr>
<tr>
<td></td>
<td>STLPort <a href="http://sourceforge.net/projects/stlport">http://sourceforge.net/projects/stlport</a></td>
<td>Multiplatform C++ standard library</td>
</tr>
</tbody>
</table>

Tabla 1. Libraries used in Maxine

XVII Congreso Español de Informática Gráfica
3.3. 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 [12], 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. The face classification implemented is an effective method based on the theory of evidence [10]. 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 recognizing surprise to a rate of 100% in the case of happiness or disgust.

3.4. Other Possible Inputs: Adaptation to the CAVE-like Facility

The system is being used to carry out presentations of precedent projects in different languages in the group’s CAVE-like facility (Figure 4). In this environment, a new kind of input/interaction has been set up:
Positioning System
The positioning system (mixed inertial-ultrasound) gives the position and orientation of the user’s head and hand. This information enables the presenter to “look at” or “speak to” the user, or the user to point at and select objects in the presentation. The “main” user also has a “mouse” with four programmable buttons to reproduce the interaction via keyboard/mouse on console.
Special care has been taken in discarding incorrect information (due to drifts or incorrect measurements) and, on the other hand, detect any appreciable change of position or orientation of the user’s head or hand, and also to identify gestures such as nodding or shaking of the head, raising of the hand to signal, point, etc.

“Background” Microphones
They monitor the sound level (and therefore, the level of attention) in the room.

4. 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 or the estimate of the user’s emotional state in image interaction.
These modules are basically in charge of generating the answer to the user’s questions in text mode (see Figure 2), and they are 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 [1] and CyN [23]. 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 main differences lie in being able to work with accents, dièresis and the “ñ” character, and in enabling the use of opening interrogation and exclamation marks.
The knowledge of the virtual character is specified in AIML (Artificial Intelligence Markup Language) [2]. 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.
The AIML interpreter has 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.
5. **System’s Outputs: The Motor Module**

5.1. **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 and actor loading process via script file. As well as general animations, which may be cyclic or otherwise, and pose animations, which are postures that are maintained for a certain length of time, the system automatically adds background and periodic animations. The former are secondary animations that gradually modify the character’s posture and some parts of its body (see Figure 5), the implementation of which is based on the use of Perlin noise. The latter are animations that are executed automatically every so often, the typical application of which is the generation of blinking or of breathing movement. Regarding facial animation, we work with the six basic expressions defined by Ekman and the nomenclature followed is that of the VHML standard [28].

In all the cases, animation blending was achieved with the help of the Cal3D library (see libraries at Table 1).

5.2. **Speech Synthesis**

In the developed system, the speech synthesis is done with the packages of Spanish voices that offer Loquendo TTS [19], although SAPI5 shareware software (see Table 1) is used for getting information about the visemes (see Section 5.3). In order to avoid that the voice sounds artificial, it has been equipped with an emotional component. The voice’s emotions follow the six universal Ekman’s categories: joy, sadness, anger, surprise, disgust and fear, plus the neutral one. SAPI5 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. These emotional parameters were configured considering several studies [6] [13] [17] and the process carried out to find the values at which these parameters must be fixed was done through voice assessment by users, following three assessment paradigms: Forced Choice, Free Choice and Modified Free Choice [3].

![Figura 5. Automatic Position Changes of the Virtual Character during a Presentation](image-url)
5.3. Expression/Speech Coordination: Lip-sync

A lip-sync module specially developed for Spanish language has been implemented. In order to solve the problem of the labial synchronization, this module uses SAPI5 (see Table 1) to obtain the information about the visemes (visual phoneme) that take place pronouncing the phrase that is wanted to be synthesized. The visemes have been used to model the different movements of the actor’s mouth.

In Figure 6 the visemes used for synchronization in Spanish are shown.

6. Maxine’s Applications

The system described in the previous sections has been used in different applications:

- MaxinePPT: Virtual humans for PowerPoint-like Presentations

This application allows PowerPoint information to be presented through a like-life character on a graphic display. This kind of presenter has demonstrated to be specially 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 [26] 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, user intervention is not necessary. 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 and Figure 8 shows some screenshots of a virtual presentation.

- **Control of a domotic environment**

A virtual avatar, called Max, was created and is used as an interactive interface for the access and remote control of an intelligent room [16]. The user can communicate with Max through natural language in Spanish and can ask to do different tasks within the domotic environment, or, also, can do queries about the different devices of the intelligent room (see Figure 9).

In order to allow all the power of the Maxine’s system, Max incorporates an emotional voice with lip-synchronization. Also, the answer’s engine is asked for getting an answer for the user. And, the answers can dynamically vary, depending on the emotional state of the avatar and on the kind of conversation with the user.

- **Interactive Pedagogical Agent for teaching Computer Graphics**

Maxine was also used for the developing of a learning platform to simplify and improve teaching and practice of Computer Graphics subjects [9]. By carefully orchestrating the different possibilities that offers our animation engine: facial expression, body placement, arm movements, hand gestures and conversational signals, the embodied pedagogical agents may encourage the students to care more about their own progress, may convey the enthusiasm in the learner for the subject matter and may simply make learning more fun.

The interactive pedagogical agent will help students in two ways: for exposing some specific topics, acting as a virtual teacher, and for understanding some difficult topics of CG subjects by allowing the interaction and handle of a 3D environment (when a relevant concept is being explained, the teacher or the student may manipulate the environment modifying different parameters, points of view, lights, etc, and interactively visualizing the results.

The use of Maxine in our courses has revealed to be a very useful instrument for the improvement of the comprehension of the most complex
concepts and the results of the students’ opinion are promising.

7. Conclusions and Future Work

Maxine, a powerful animation engine for managing virtual environments and virtual actors was presented. The system allows the development of new applications where interaction is based on virtual agents supporting multimodal and emotional interaction.

The main features that make our system stand out are that it supports real-time interaction with the user through different channels (text, voice, mouse/keyboard, image); that the voice interface enables communication in natural language and in Spanish; that the discourse is reinforced with referential acts (pointing gestures and/or looking at objects) and interaction (following the user’s movements with eyes and body orientation, allowing the user to interrupt, etc.); that the system is capable of adding character movements that increase its naturalness and credibility (blinking, position and expression changes, etc). The communication between the user and the 3D character is done through the virtual character’s facial expressions and with its “emotional voice” since the answers to questions and the modulation of the voice are adapted to the emotional state.

The potential of the animation engine was shown through different applications that use virtual humans for enhancing interaction: as virtual presenters in PowerPoint-like presentations, as interface for controlling a domotic environment and as virtual teachers for improving theory and practice in a Computer Graphics course.

Several other research lines also remain open, most of which focus on enriching interaction between the presenter and the spectators:
- to consider not only emotion but personality models for the virtual character
- to enrich virtual character behaviour moving from reactive to cognitive schemes

Nowadays we are working in the validation of Maxine system and their characters by the users.

Acknowledgments

This work has been partially financed by the Spanish “Dirección General de Investigación” (General Directorate of Research), contract number Nº TIN2004-07926 and by the Regional Government of Aragon through the WALQA agreement (ref. 2004/04/86).

The authors thank Sergio García Masip and David Anaya for their collaboration in this work.

Bibliography