Developing multimodal interactive systems with EyesWeb XMI

Antonio Camurri
antonio.camurri@unige.it

Paolo Coletta
paolo.coletta@unige.it

Giovanna Varni
giovanna@infomus.dist.unige.it

Simone Ghisio
simone@infomus.dist.unige.it

InfoMus Lab – Casa Paganini
University of Genoa
Piazza Santa Maria in Passione, 34
16123 Genova, Italy
+39 010 2758252

ABSTRACT
EyesWeb XMI (for eXtended Multimodal Interaction) is the new version of the well-known EyesWeb platform. It has a main focus on multimodality and the main design target of this new release has been to improve the ability to process and correlate several streams of data. It has been used extensively to build a set of interactive systems for performing arts applications for Festival della Scienza 2006, Genoa, Italy. The purpose of this paper is to describe the developed installations as well as the new EyesWeb features that helped in their development.

Keywords
EyesWeb, multimodal interactive systems, performing arts.

1.INTRODUCTION
EyesWeb is nowadays a settled platform to perform research and analysis on human expressive gestures features, as well as to design interactive systems based on natural and expressive interfaces [1].

Initially focused on video analysis of human movements, the platform has evolved toward multisensorial analysis [5] and, more recently, in the direction of multimodal and cross-modal processing. Multimodality relates to the ability to integrate the results of analysis of different multimedia streams, whereas cross-modality concerns the ability to apply the same algorithm to different modalities [2].

The result of this evolution has brought to the current version, named EyesWeb XMI (for eXtended Multimodal Interaction), which has been extensively used to setup a number of public demos for the Festival della Scienza event (http://www.festivalscienza.it), which was hold in Genoa, Italy, October 26-November 7, 2006. The installations where designed to be used beyond the festival deadline, thus, they were open to the public up to January 2007. During the festival, Casa Paganini (an international center of excellence for research on sound, music, and new media, where InfoMus Lab has its main site,) was visited by over 2500 persons spreading different ages.

In this paper, we give a description of both the new characteristics of the EyesWeb XMI platform and on the installations designed and developed with it. Thus, in Section 2 we give a brief description of EyesWeb, focusing on the novelties with respect to previous releases; in particular, we focus on the new features that support synchronization of different media sources and parallel signal-processing in each different CPUs. In Section 3 the installations developed with EyesWeb are presented, focusing on the advantages brought by the new characteristics of the software.

NIME07, June 7-9, 2007, New York, NY

2.THE EYESWEB XMI PLATFORM
The EyesWeb evolution from release 3 to release 4 of EyesWeb had already brought a number of new features to improve its usage in the field of multimodal processing. In particular, automatic conversion of datatypes and the availability of blocks able to operate on several different streams of data were the main innovative aspects [2]. Since then, the platform has further evolved, and the new EyesWeb XMI release presents a number of novelties: many of them are devoted to improve support to analysis and processing of several different multimedia streams. The most important characteristics are the presence of sync-(in/out) pins and the support to parallel architectures. The availability of sync-in and sync-out pins allows each block to act as a clock generator, or to get the clock signal from any source. The improved support for parallel architectures enables the exploitation of the characteristics of moder computer systems.

Sync-in pins
Multimodal processings requires the ability to handle several data streams, which are often associated to different clock sources. Several approaches to synchronize different streams involving different clocks have been developed in the literature, including the case where the time sources are different even at the conceptual level (see, e.g., [4]). However, our purpose here is not only to provide specific algorithms to re-synchronize signals (for example, a time-based effect, such as time-stretching), but to provide a general framework where the patch designer can specify how the different clock sources interact, and how they are related to data processing.

Sync-in pins represent the clock sources for EyesWeb blocks. Although graphically hidden by default, not to burden the patch layout, they are available for every block and can be exported on user's choice. When a sync-pin is exported, a clock signal may be connected to it such that the block is activated according to the speed of this clock source. The following is a typical example where the *Video File Reader* block (which extracts a video track from a file) is synchronized with an audio signal which acts as clock generator.

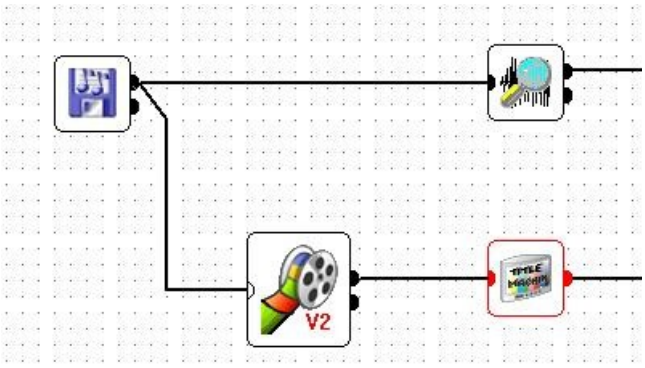

Figure 1: The audio signal is connected to the sync-in pin of the VideoFileReader block.

It is noteworthy that any type of data can act as a clock signal: thus, any stream can be connected to a sync-in pin (e.g., a MIDI stream). Moreover, this type of connection does not cause data transportation, as only information needed for synchronization is propagated. Thus, connecting to a sync-in pin does not imply higher CPU usage.

The sync signal operates by overriding the usual EyesWeb activation rule. Under normal condition the block may be activated according different criteria; the most common one for filter blocks is being activated on when new data is available, whereas source blocks (e.g., the *VideoFileReader*) are activated according to their own clock (the *VideoFileReader* uses the PC clock; the *FrameGrabber* block uses the clock of the video signal, etc.). When a signal is connected to the sync-in pin, the block is activated when the clock signal is raised (if the clock signal is extracted from a data stream, the signal is raised when new data is available). Note that, for the greater flexibility, the sync signal can be mixed with the native clock of the block; thus, the activation rule of the block instead of being overridden is integrated with the clock signal.

Sync-out pins

The sync-out pin let every block to act as a clock source. The sync-out block generates a clock signal, and this clock is raised each time the block is activated. Thus, the clock signal generated by this pin is not directly related to the data stream generated. As a matter of fact, the block might be a sink block, i.e., a block with no output; similarly the block might not generate a datatype for each activation (e.g., the block might subsample input data). Nevertheless, the clock is generated whenever the block is activated.

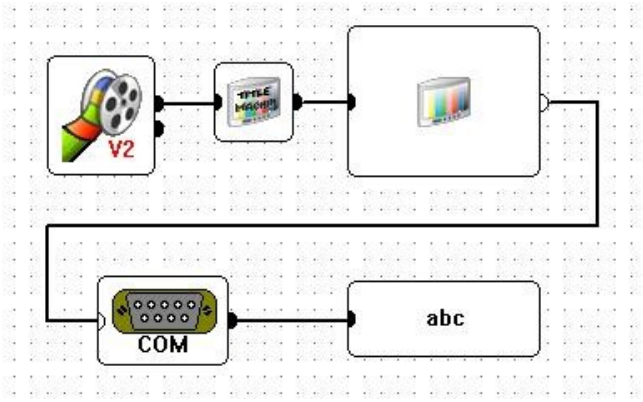

Figure 2: A clock signal is generated by a VideoDisplay, and is used to synchronize a serial input block

Figure 2 shows an example where the *VideoDisplay* block, which has no outputs, exports a sync-out pin used as a clock source for a *SerialInput* block.

Parallel processing

Parallel execution refers to the kernel support for executing the blocks in separate threads, in order to exploit the characteristics of modern computer systems. Multi-processor architectures are now very common. They were usually limited to top-series workstations, and their cost was significantly higher than standard single-processor computers. Nowadays, dual-core architectures are in use even in low-cost PCs, and quad-core systems are already on the market. Further, main CPU producers have indicated in their roadmaps that multicore architectures with higher parallelism degree will be available on the market in a relatively short period. This means that parallel processing is becoming available in most computer systems. Therefore, the exploitation of the characteristics of these architectures can provide benefits in terms of computational power. In particular, since the EyesWeb architecture is focused on multimodal processing, it is straightforward to assume that there are a number of blocks that can operate in parallel, since it commonly operates on several streams of data. However, it must be guaranteed that synchronization is not lost when data needs to be correlated. The XMI platform supports this sort of processing: when a split (or fork) occurs in a patch, or when blocks are unrelated (i.e., no links between them), they are run in separate threads of execution, which practically means that they are executed on different CPUs or on different cores. When different streams of data converge to the same block, then data are resynchronized. In this way, parallelism is implemented in a transparent way: from the user perspective, when synchronization issues are concerned, the system behaves exactly as if processing is serialized.

The degree of parallelism is automatically chosen by the system basing on the number of available processing units, however, users can override this value for special circumstances: for example, when the user knows that there are other processes requiring CPU resources besides EyesWeb on the same computer.

Consider, for example, the following patch, where a video effect is applied to an image. The effect is performed by applying two

different operators to the same image, and then summing the results.

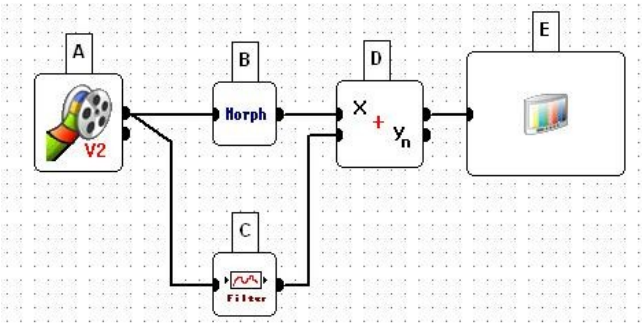

Figure 3: blocks B and C are executed in parallel, on separate CPUs if more than one is available. Data is resynchronized before being sent to block D.

Blocks B and C are executed in parallel, thus, the total time required to process all data is equal to the critical path (i.e., the longest between the upper and the lower path), instead of being equal to the sum of the all execution times.

3.INTERACTIVE MULTIMODAL INSTALLATIONS

The XMI platform has been satisfactorily exploited to design and develop a number of interactive multimodal installations for the science exhibition "Cimenti di Invenzione e Armonia", held at Casa Paganini, Genova, Italy, from October 2006 to January 2007. The exhibition was part of "Festival della Scienza", a huge international science festival held in Genova every year.

The exhibition was organized in a percourse in the auditorium and the museum rooms of Casa Paganini, where the visitor could interact with a number of installations. In the following, we give a brief description of the installations. All of them were designed and developed by means of the new XMI platform.

A whole room of the scientific exhibit "Cimenti di Invenzione e Armonia" was devoted to Tangible Acoustic Interfaces (TAIs). The room consisted of three installations. The first one implements a TAI interface (a table) toward the Google Earth application. An MDF (Medium-density fibreboard) surface is sensorised with Four Knowles BU-1771 accelerometers. Vibrations along the depth of the material, which badly affect performance, were reduced by inserting a panel of phonoabsorbent material below the tangible surface. In this way, vibrations along the surface were emphasized with respect to vibration crossing the surface. Accelerometers are connected through a custom audio front-end to a Firepod PreSonus audio interface, so that data are received as audio signals. Moreover, a B/W videocamera is placed in front of the surface, the major axis of the videocamera normal with respect to the surface. Since we project the Google Earth output onto the tangible interface, which is in the view of the videocamera, we have to filter out the effect of videoprojection. To this aim, the camera is endowed with an infrared filter and the tangible surface is enlightened with infrared light. The movement of the two hands of the user is tracked by means of multimodal integration of visual and acoustic continuous tracking techniques. Both kinematical (e.g., position, velocity, acceleration) and expressive features (e.g., directness, impulsiveness) are extracted from user's gestures. The user can rotate and zoom in and out satellite images projected on the table. Rotation and zooming are controlled by both kinematical and expressive gesture features. For example, zoom percentage is controlled by the distance of the two hands; inertia in the flow of satellite images can be controlled by impulsiveness and hesitation.

In this application, sync-in pins were used to synchronize the results from the analysis performed on the audio and video stream.

The second application is a TAI chair that was used in the music theatre opera "Un avatar del diavolo", composer R.Doati, presented at La Biennale, Venezia, September 2005. Touch gestures of an actor on the chair are localized and processed: the touch position and its qualitative features (e.g., with hard and quick tapping-like movements, or with light and smooth caress-like movements) are used to control sound generation and processing in real-time.

The third application was composed of a painting stand with a white canvas, which was transformed in a TAI. Gesture direction and impulsiveness was detected and used used to control the playback of video sequences projected on the canvas. For example, a fast gesture toward the right triggers fast-forward reproduction of the video. Expressive qualities of the performed gestures were mapped to local expressive variations of the reproduction (e.g., hesitant movement mapped to slower reproduction rate).

These three applications were connected together and their activation was controlled through a small cell-phone like device. Touching the TAI objects (the table, the chair, the painting stand) with the cell-phone provokes the recognition of the object from the noise produced by the touch of the cell-phone on the object and therefore causes the activation of the corresponding application. The cell-phone is also able to recognize other (passive) objects in the room (e.g. a metal handrail) ftom the noise caused by the touch/impact on the object itself. In a separate paper in preparation we describe this promising mobile application.

In the main room of the exhibition (the auditorium of Casa Paganini), another installation is available: the Orchestra Explorer [3]. This interactive system allows users to physically navigate inside a virtual orchestra in order to actively explore the music piece the (virtual) orchestra is playing, and to modify and mould the sound and music content in real-time. The Orchestra Explorer was installed on the stage of the 250-seats auditorium at Casa Paganini. By walking and moving on the stage, the user discovers each single instrument and can operate through her expressive gestures on the music piece the instrument is playing. The installation covered a surface of about 9 m 3.5 m. A single vidocamera observed the whole surface from the top, about 7 m high, and at a distance of about 10 m from the stage. Four loudspeakers were placed at the four corners of the stage for audio output. A white screen covered the back of the stage for the whole 9 m width. A videoprojector projected on such screen the video feedback. Lights were set in order to enhance the feeling of immersion for the users and to have a homogenous lighting of the stage. The music piece "Borderline", by M. Canepa, L. Cresta, and A. Sacco, was selected for the installation. "Borderline" is an original piece of film music, which was never performed in

public. It consists of 26 mono audio tracks and includes the following music instruments: harp, cello, horn, flute, double bass, oboe, bassoon, percussions, piano, violins, and alto pizzicato. The 26 tracks are mixed (including rendering and processing according to gesture qualities and the position on the stage) in real-time, and the mixing strategy determined by the use of the physical space made by the user (see to [3] for more details).

This applications involved playback and real-time processing of a significant number of audio tracks, as well as the processing of expressive full-body movement and gesture which involved the parallel execution of the same algorithm (the same EyesWeb subpath) with different tuning parameters for different physical places on the stage. Clearly, this application benefited of the EyesWeb improvements in parallel processing: the whole application was run on a single workstation (Dell Precision 380, equipped with two CPUs Pentium 4 3.20 GHz, 1 GB RAM).

4.ACKNOWLEDGMENTS

We wish to thank our colleagues at DIST – InfoMus Lab for their support in this work. In particular, we thank M. Peri, A. Ricci, M. Demurtas, and R. Sagoleo for their contributions to the development of the EyesWeb XMI platform, and C. Canepa and G.Volpe for their invaluable contribute to research and to the design and development of the Orchestra Explorer installation.

5.REFERENCES

[1] Camurri A., Mazzarino B., and Volpe G. *Expressive interfaces*. Cognition, Technology & Work, 6,1 (Feb. 2004), 15-22.

[2] Camurri, A., Canepa, C., Drioli, C., Massari, A., Mazzarino, B. and Volpe, G. *A Multimodal and cross-modal processing in interactive systems based on tangible acoustic interfaces.* Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2006), University of Hertfordshire, Hatfield, United Kingdom, September 2006.

[3] Camurri, A., Canepa, C., Volpe, G. *Active listening to a virtual orchestra through an expressive gestural interface: The Orchestra Explorer*. Submitted for presentation at NIME 2007, New York, NY, 2007.

[4] J. Borchers, E. Lee, W. Samminger, and M. M¨uhlh¨auser. *Personal orchestra: A real-time audio/video system for interactive conducting.* ACM Multimedia Systems Journal Special Issue on Multimedia Software Engineering, 9(5):458–465, March 2004.

[5] A. Camurri, M. Ricchetti, R. Trocca, *EyesWeb - toward gesture and affect recognition in dance/music interactive systems.* in Proc. IEEE Multimedia Systems '99, Firenze, Italy, June 1999.