The CINEMA Project: A Video-Based Human-Computer Interaction System for Audio-Visual Immersion

by Renaud Dardenne, Jean-Jacques Embrechts, Marc Van Droogenbroeck, and Nicolas Werner

Numerous studies are currently focusing on the modelling of humans and their behavior in 3D environments. The CINEMA project has similar goals but differentiates itself by aiming at enhanced interactions, the creation of mixed reality, and the creation of interactive and reactive acoustical environments. Part of the project consists in gesture recognition of a user, who is given the real-time control of auralization and audio spatialization processes.

The CINEMA project is a collaborative effort between teams at four Belgian universities. The project foresees the development of a real-time system that includes 3D user and environment modelling, extraction of motion parameters, gesture recognition, computation of depth maps, creation of virtual spaces, and auralization and spatialization. The merging of these techniques is expected to produce a tractable 3D model allowing users to interact with a virtual world.

Although much progress has been made in the areas of video and audio processing, there is still a need to find a way of properly combining these elements. The production of a tool for augmented or virtual reality that mixes video and audio could be used for a number of applications: as an educational tool for example, or at cultural places like museums.

This project started on January 2003 and will end in September 2006. It is funded by the Walloon Region, Belgium.

Following are details regarding the video-handling techniques and the interaction with the sound system:

The Video Analysis Engine

The major goal of the video part of the system is to identify simple human gestures. As real time operation is a compulsory requirement for any interactive system, we committed to the use of simple tools to achieve gesture recognition. Therefore our algorithm detects motion based on a known algorithm for background extraction, combined with a skin detection algorithm to isolate the head and the hands. To increase the robustness of the system, we tested several motion detection algorithms and combined them. For indoor scenes, a background extraction using a static background and a thresholding algorithm in the HSV colour space for skin detection appeared to

suffice, except for the presence of shadows. Fortunately shadows have a specific range of colours and particular shapes that enable us to discard them.

Once regions with skin have been detected, we must distinguish between the head and the hands. It is hard to propose a fast and general method that would consider all possible pathological cases for the relative positions of skinned regions. Therefore we rely on the assumption that the head lies in the middle of the region with motion. The final system is capable of recognizing simple hand positions (both hands raised above the head, both hands on the same side of the body etc), and sending appropriate instructions to the audio subsystem (see Figure 1).

The Audio Subsystem

To build a realistic and immersive world, we have implemented a sound spatialization system. This audio system is able to synthesize localized sound sources, which means that we can decide on the orientation and distance of all the sources. In addition, the system uses real-time software to produce a realistic sound. This last technique, called auralization, is fully configurable and takes into account the room configuration or the virtual audio environment via its impulse response, and the positions of the sound sources and the listener. For the sound rendering, a multichannel setup is used, and the audio signals are distributed to the loudspeakers according to an amplitude panning technique.

The Communication Channel etween the Video and Audio Subsystems

The video and the audio subsystems communicate through a socket interface. This approach offers the choice of running all the algorithms on a single machine (using the loopback) or on separate machines. As all the machines are presumably located on a LAN, a communication channel based on the UDP protocol was chosen. The subsystems interoperate, and in order to allow for further extensions, both understand a language that uses tags and is based on the XML language.

Links:

The institutes involved in this project are:

- the Telecommunications and Imaging Laboratory at the University of Liège, Belgium
  http://www.ulg.ac.be/telecom/
- the Acoustics Laboratory at the University of Liège, Belgium
  http://www.montefiore.ulg.ac.be/services/acous/indexfr.html
- the Electronic Circuits Theory and Signal Processing group at the Polytechnic Faculty of Mons, Belgium
  http://tcts.fpms.ac.be/
- the Communications and Sensing Laboratory at the Catholic University of Louvain-la-Neuve, Belgium
  http://www.tele.ucl.ac.be/

Please contact:

- Nicolas Werner, Université de Liège, Belgium
  Tel: +32 4 366 2652
  E-mail: nwerner@ulg.ac.be
- Renaud Dardenne, Université de Liège, Belgium
  Tel: +32 4 366 26 86
  E-mail: R.Dardenne@ulg.ac.be
- Jean-Jacques Embrechts, Université de Liège, Belgium
  Tel: +32 4 366 26 50
  E-mail: jjembrechts@ulg.ac.be
- Marc Van Droogenbroeck, Université de Liège, Belgium
  Tel: +32 4 366 26 93
  E-mail: M.VanDroogenbroeck@ulg.ac.be

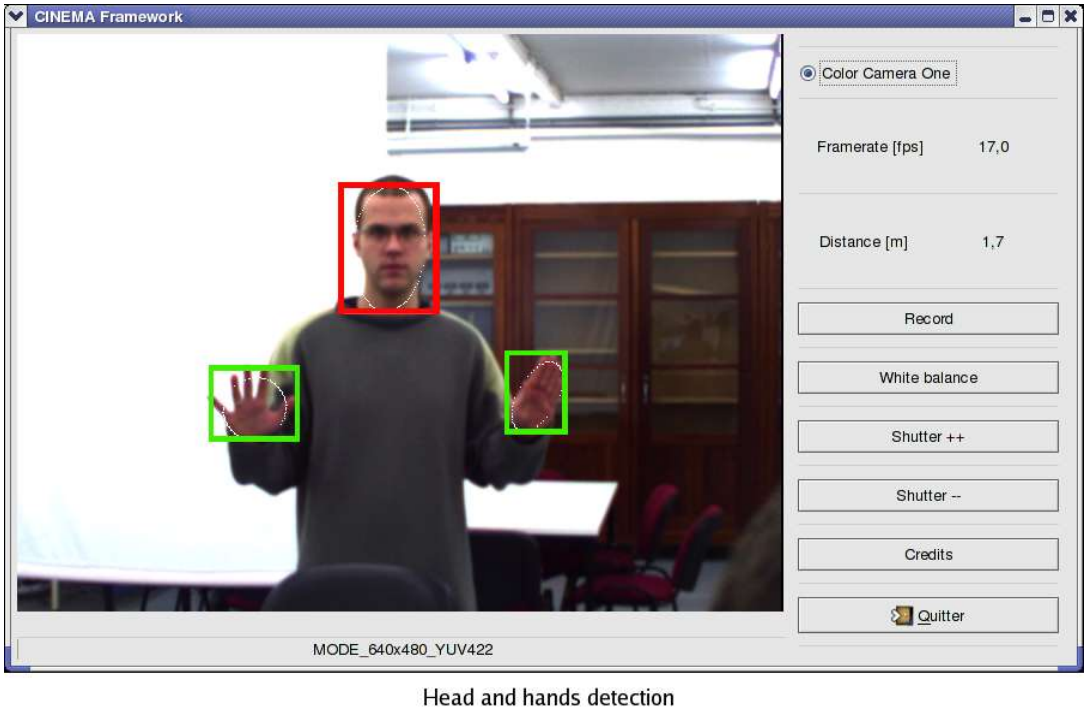

Figure 1: Intermediate result that shows the extraction of skinned regions (hands in green, and head in red) further used to control the audio sub-system.