A VIDEO-BASED HUMAN-COMPUTER INTERACTION SYSTEM FOR AUDIO-VISUAL IMMERSION

R. Dardenne, J.J. Embrechts, M. Van Droogenbroeck and N. Werner

R.Dardenne@ulg.ac.be, jjembrechts@ulg.ac.be,
M.VanDroogenbroeck@ulg.ac.be, nwerner@ulg.ac.be
Univ. Liège, INTELSIG group, Dept. Electrical Engineering and Computer Science,
Sart-Tilman B28, B-4000 Liège 1, Belgium

ABSTRACT

This paper describes the video and audio tools that have been implemented in a real-type system to immerse a user into a virtual scene. The video tools include motion detection, skin detection by thresholding, shadow detection and extraction, and finally the user's head and hands detection. Once this is done, the user (who is surrounded by a matrix of loudspeakers) is able to move a sound source in the horizontal plane around him. Moreover, the sound is *auralized* by convolution with (directional) room impulse responses, which have been pre-computed by a ray tracing method. The different sound contributions are distributed to the individual loudspeakers by applying the VBAP technique.

1. INTRODUCTION

The CINEMA research project has been initiated by the TELE research group of the Catholic University of Louvain-la-Neuve (Prof. B. Macq) and the partners are, besides TELE, two research teams of the University of Liege (the authors of this paper) and the department TCTS of Polytech. Mons (Prof. J. Hancq). The general purpose of the project is to immerse a user into a virtual scene and to create the corresponding audio virtual environment.

This paper describes some methods that have been developed in Liège, in order to mix video and audio tools in a real-time system of augmented or virtual reality. In this system, the user is given the opportunity to interact with some video and/or audio objects of this virtual world. For example, in this paper we describe how he will be able to move a sound source in the horizontal plane around him.

The paper first addresses the video and audio methods individually, and then describes the merging of both techniques.

The system could be used for various applications, as an educational tool or at cultural places like museums.

2. GENERAL PRESENTATION OF THE SYSTEM

Figure 1 represents the visual field of the user, who stands in front of a screen on which he finds himself immersed into a virtual scene.
Two cameras are used. A front camera captures real-time images that are used to delineate the user and to select some parts (like hands and head) used for interaction. After processing, the user's image is projected onto the screen. A second camera, placed above the user, is used to construct a 3D view of some motions. Selected motions are then interpreted to interact with the system. For the audio part, the user is surrounded by six loudspeakers which are equally distant (by 60 degrees) on a five meters large circle. Each loudspeaker is fed by an independent audio signal, leading to efficient sound localization in the horizontal plane.

Figure 1: *Visual field of the user.*

3. DESCRIPTION OF THE VIDEO TOOLS

To move the sound source, we want to capture and interpret the user's arms positions. For this, we will use two color cameras: one camera faces the user, the other captures a top view from above the user's head.
The front camera is used to command the volume of the source and the top camera is intended to move this source in the horizontal plane. Here follows the different steps we perform to get the positions of the user's hands:

Step 1: motion detection.
A motion detection algorithm is a usual preliminary stage to the identification of a head or hands. Many of the numerous techniques for motion detection perform a background extraction, for example by background substraction. After several tests, like a simple difference with the previous images, weighted average of the background, histogram based techniques, etc, it appeared that the simple one consisting in the difference with a static background (equal to the first image) meets our requirement. Keeping

Figure 2: *Skin thresholding in the YUV colorspace.*

Figure 3: *Skin thresholding in the HSV colorspace.*

the computation times as low as possible was an additional reason to be satisfied with that simple motion detection algorithm.

Step 2: skin detection by thresholding.
Human skin has interesting color characteristics in several colorspaces which do not depend on the skin color. Hsu et al. compared face detection algorithms in color images [1]. It seems that the tint-saturation-luma space best fits the purpose of skin detection. Unfortunately cameras do not operate in this colorspace so that a slow conversion process is compulsory to deal with tint and saturation. Our experiments show that the YUV colorspace, although not performing best, is acceptable. Hsu et al. made a similar choice (but for the YC_bC_r colorspace) and justified it by stating that the set of skin colors in such a colorspace form a compact cluster. See figures 2 and 3 for the results of thresholding in two colorspaces.

Step 3: combination of the motion detection and skin extraction.
Even in a laboratory environment, not to speak about less friendly environments, the combination of motion and color informations offers bad performances. There are several reasons for that. Scenes contain objects with colors close to the skin color (for example pieces of furniture like shown on left-hand side of figures

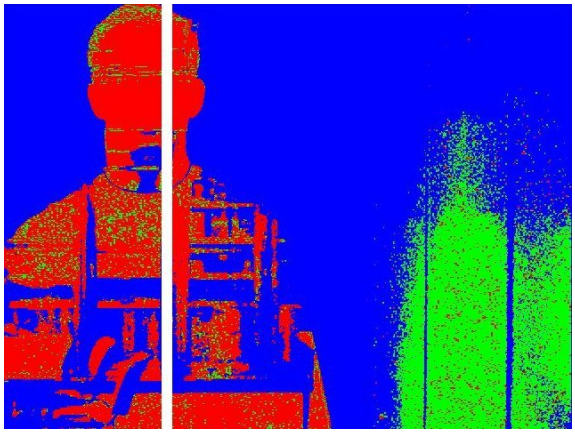

Figure 4: *Shadow Detection.*

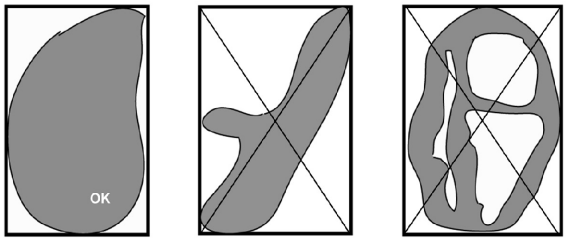

Figure 5: *Geometrical criterions to distinguish between the head and the hands.*

2 and 3), leading to false positives. Moreover the model used for detecting motion is too simple to consider shadows. We solved this problem by implementing a technique of shadow detection and extraction similar to the technique recommended by Prati et al. [2]. Our technique uses the HSV space to separate the moving objects from their shadow: a pixel belongs to a shadow if its hue and saturation components are sufficiently close to the ones of the corresponding background pixel. This method is computationally expensive but it performs well and thanks to the use of a lookup table for the colorspace conversions we manage to keep the CPU load to an acceptable level. An illustration of the extraction of shadows is shown in Figure 4.

Step 4: discrimination between head and hands.
A head is not the only skinned part of a body. We still have to differentiate between a head and hands. First we select areas respecting some geometrical criteria of shape, of width/height ratio, filling and connexity (for example see Figure 5). Under the assumption that there is only one person in the scene, we only keep the 3 regions with the largest areas. Then to distinguish the head from the hands, we state that the head area is the one which is the most in the middle of the moving object. An illustration of this principle is shown in Figure 4, where we have superimposed a white vertical line in the middle of the motion region after removing the shadow.

Once the position of the user's head and hands has been estimated, it is possible for him to start interacting with the system and modify the volume of the source. We will now use the

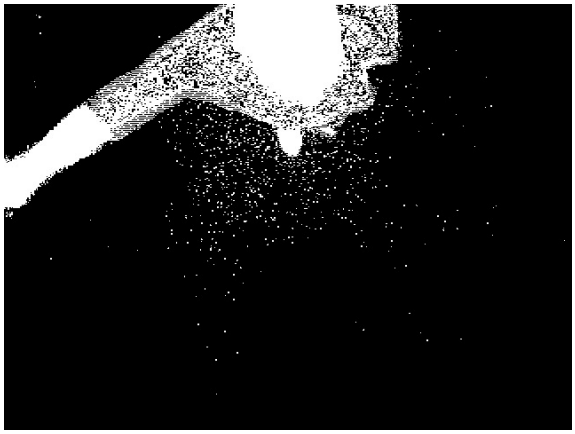

Figure 6: *Foreground detection of the roof camera.*

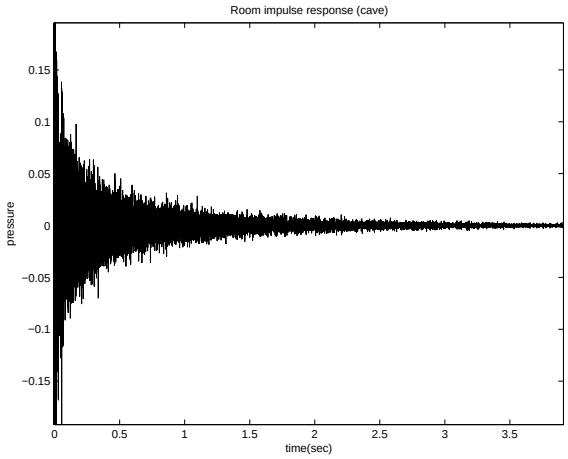

Figure 7: *Room Impulse Response example*

top camera to move this source around a circle in the horizontal plane. As descibed in the step 1, we perform a foreground detection on the image as shown of figure 6.
Finally, we fit an ellipse on the foreground and we consider the major axis represents the user's arm direction.

Once the position of the major axis is known, the system filters all gathered informations over time and it sends instructions to the audio sub-system.

4. DESCRIPTION OF THE AUDIO TOOLS

4.1. Introduction

The final goal to achieve is that all users in the Virtual Reality reproduction room share the same auditory experience as they would have in the virtual environment.

The audio system is composed of one PC, communicating with the video PC via ethernet network, and equipped with an 8 output channel soundcard.

The sound sources signals have been prerecorded in anechoic conditions in mono and are stored on the local hard disk as WAV files. Thus, our purpose is not the modelling/synthesis of virtual sources themselves, it's rather the modelling/synthesis of the room acoustics and the "auralization" process for reproducing the accurate sensations at the user's ears.

4.2. Room acoustics modelling

Given the room geometry and wall acoustical properties descriptions, the source and receiver (= the user)'s positions, the sound pressure signal at the receiver can be seen as a filtered version of the signal emitted by the source.

If we assume both source and receiver to be static, the filter is a LTI system. This is the system we want to model.

The room acoustics modelling method is based on a ray tracing algorithm. Since this algorithm is currently not real time, the modelling will be achieved offline, i.e. impulse responses (like in fig. 7) will be precomputed for each virtual room configuration, with several source and receptor positions.

Note: The impulse responses could also be measured in a real space.

4.2.1. Ray tracing method used in software Salrev [3]

- considers the sound propagation along rays, ignores wave propagation characateristics (thus less valid at low frequencies)
- omnidirectionnal spherical sound sources
- specular and diffuse reflections on walls, with octave band coefficients values.
- frequency dependent air absorption (octave band coefficient values)
- spherical receivers which record time arrival, incoming direction and energy level of the sound rays.
- the results of the ray tracing process are echograms in each octave band.
- since a limited number of rays are traced, statistical errors are inherent on the energy levels. The statistical errors lead also to finite temporal and directional resolution
- Statistical errors can be avoided on the direct component and the first order specular reflections.

From an echogram, an impulse response is created (adding a random phase), and auralization is then possible.

4.2.2. Directional Room Impulse Responses

An idea to enhance sound localization in our virtual system is to model separately the direct contribution of the source and the first order mirror sources.

Another idea that we have introduced is to compute echograms along several directions of incidence at each spherical receiver. This technique of directional Room Impulse Responses has been described in [4]. Twenty-six directional RIR are presently computed at each spherical receiver.

4.3. Auralization system

Assume we have a prerecorded anechoic sound signal and a set of directional impulse responses, with possibly separate direct components and first order reflections from the offline modelling. This prerecorded signal must be processed before sending it to the soundcard outputs driving the loudspeakers (both are supposed to be high quality transducers).

4.3.1. Filtering

The anechoic sound signal must be convolved with the room impulse response. This convolution has to be fast, since the filter has typically several tens of thousands coefficients. The idea is to do it in the frequency domain. However, because the latency must be small, a special version of the fast convolution is being used, similar to [5] or [6].

4.3.2. Reproducing the direction of the incoming sound field at the user's ears

This part is dependent on the loudspeakers configuration, and is realized by implementing a geometrical panning law called VBAP ([7]).

The principle is to distribute the incoming direct sound, mirror contributions and convolved *auralized* signals among the well chosen individual loudspeakers. VBAP suggests the way this distribution should be realized. This panning law has been tested for several loudspeakers configurations, and the present system uses 6 loudspeakers equally spaced on a circle surrounding the user.

5. ACKNOWLEDGEMENT

This part of the CINEMA project is funded by the Walloon Region of Belgium, under contract 021/5392.

6. REFERENCES

[1] R.-L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, May 2002.

[2] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918–923, July 2003.

[3] J.J. Embrechts, "Sound field distribution using randomly traced sound ray techniques," *Acustica*, vol. 51, no. 6, pp. 288–295, 1982.

[4] J.J. Embrechts, N. Werner, and S. Lesoinne, "Computation of directional impulse responses in rooms for better auralization," Audio Eng. Soc. 118th convention, May 2005.

[5] W. G. Gardner, "Efficient convolution without input-output delay," *J. of the Audio Eng. Soc.*, vol. 43, no. 3, pp. 127–136, March 1995.

[6] A. Torger and A. Farina, "Real-time partitioned convolution for ambiphonics surround sound," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2001.

[7] V. Pullki, "Virtual sound source positioning using vector base amplitude panning," *J. of the Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, June 1997.