**Immersive visual and audio world in 3D**

Patrice Bouvier, Pascal Chaudeyrac, Raphaël Loyet, Benoit Piranda, François de Sorbier de Pougnadoresse

SISAR team, Université de Marne la Vallée
{bouvier, pchaudey, piranda, fdesorbi}@univ-mlv.fr

**Abstract**

This article presents the approach we followed for the creation of our virtual reality room, both from the hardware point of view and the software one.

Our main goal to build a virtual reality room as cheap as possible and easily transportable was constrained by our knowledge of the mechanisms of human preception. Like any virtual reality system, our room aims to immerse the user in a place where he will be able to feel the presence of the virtual objects and his self-presence in the virtual environment. To recreate this dual feeling of presence we use the less intrusive as possible sensory interfaces and the most natural and intuitive motor interfaces.

We achieve this goal by diverting the user’s attention to the application itself. We present applications in various fields: art, games and archaeology.

**Keywords:** virtual reality, immersion, presence, spatial auditory display

---

**1. Introduction**

The aim of a virtual reality system is to immerse one or more users in an artificial environment where he will feel and interact in real-time thanks to sensory and motor interfaces. The experience will have to be credible enough to gull user’s senses in order to create, the ultimate goal, a feeling of presence of the virtual objects, but also a feeling of his presence in the virtual environment. This will be the only condition for the user to accept to take part of the game.

Among the classic virtual reality systems, let us quote the CAVE [CNSF92] and the Workbench [KBF95]. The Cave Automatic Virtual Environment, is composed of a small room of roughly five square meters, where each wall and the ground are screens on which are displayed synchronised images. The Workbench is a virtual work environment composed of two perpendicular screens, a liquid crystal display which provide depth perception, a head motion capture device and a six degrees of freedom stylus to manipulate objects. At last another method is to use a head mounted display (HMD) with liquid crystal display.

Those solutions are both very expensive and/or hardly transportable. Targeting different objectives, the Sisar team of Marne la Vallée University has built a virtual reality room with consumer grade components. Our objective is to dispose of an installation at low cost and easily transportable. It allows us to test various user interfaces, as less intrusive as possible, and different visual and sound rendering in order to immerse the user more and more.

After a brief outline on the methods used for visual and sound immersion we'll focus on our interfaces managing the interactions. Finally, before we conclude, we’ll present some of our realizations. It is important to note that proposing applications with a strong focus on the playful aspect, build on a strong scenario or bearer of feeling help to forget the interfaces and develop a feeling of presence.

**2. Immersion**

Within the framework of our virtual reality room, the immersion of the user is based on the visual rendering in 3D based on stereo-vision and a 3D sound environment reproduced thanks to an 8.0 sound system. After a brief theoretical reminder, we’ll describe these devices.

**2.1. 3D visual display**

From a visual point of view, the immersion of a user is based on three fundamental points: the field of view of the installation, the depth perception and the real-time display of the
virtual environment. The first point is easily solved by respecting a correct ratio between the distance separating the user and the screen, and the screen’s dimensions (see figure 1). For the second point, depth perception is created by stereoscopic images [Oko]. They simulate the human vision, that is based on the perception of a scene through two different points of view: each eye having a specific image. This is this rule that the stereoscopic system (see figure 2), the usage of filters and polarizing glasses, targets to recreate. The two images are the result of two OpenGL processing of the same scene; these images are afterward displayed with two video-projectors equipped with polarizing filters. The separation of the two images on the same screen is made with polarizing glasses beared by the users.

![Figure 1: Our 6x3 meters screen.](image1)

![Figure 2: Stereoscopic system.](image2)

The usage of a home-made 3D engine allows us a total control of the parameters. Among others, the management of the distance between the two cameras, or their orientation allows us to adapt the quality of the depth perception according to the distance of the user. Our room, made with large audience transportable material has been employed during the last "Journée du Patrimoine" in a theater (see section 6). The management of these parameters allowed us to switch easily between a public of 200 people sitting on terraces and a unique user at 2 meters from the screen.

Another great point of a home-made system is that it allows us to switch easily from a hardware solution to another so that it is possible to experiment algorithms linked to recent evolutions on graphical cards. The massive use of shaders programs [Ros] in our rendering algorithms allows us to guaranty a satisfying framerate without making concessions on the visual effects that are essential to the immersion. For example, the usage of shader-based blur algorithms [PdSdP04] allows orienting the regard of the user to relevant zones, but also gives a bit of rest to the observer’s eyes and increases his visual comfort (see figure 3).

![Figure 3: Blur in virtual scene.](image3)

2.2. Spatial sound

2.2.1. Theory

First of all let us remind that the audible range of frequency for human’s ear is going from 20Hz to 20kHz. In 1907 Lord Raleigh [Ral07] defined the Duplex theory on the sound perception in three dimensions. This theory approximates human head to a sphere with two holes diametrically opposed representing the ears. Besides the case where the sound source is located on the median plane between the two ears, the distance from the sounds to the ears ipsilateral and contralateral are different. This difference, because the transit time, is the source of the interaural time delay (ITD). However the ITD is only effective for frequencies above 1500Hz†. Under this frequency, the waves aren’t disturbed by obstacles of the size of a human face. It’s interesting to note that the ITD occurs particularly for the localization of sounds with fast attack. For continuous sounds or whose attack is slower, the brain will use the interaural phase

† That corresponds to a wave-length of about 22cm, that is the width of the head
difference (IPD) that match the difference of phase at each ears. Above 1500Hz the interaural level delay (ILD) starts to operate. Due to their wave-length these waves bounce on the head so they are largely declined (through diffraction) or not transmitted to the contralateral ear. According to Carlile [Car96], an attenuation of 40dB if obtained for a frequency of 3kHz.

The theoretical limit of the Duplex theory is that it makes us feel we hear in 3D because we have two ears. To get convinced let us study the case of a person who is deaf of one ear, we will talk about monaural audition, it is possible for this person to localise a sound in three dimensions [WK97]. It rely on the learning of extended spectral signatures, not only determined by the sound itself but also by it’s relative position regarding the audior. Those spectral variations are based on interferences caused by the shoulders, the head and the pinna\(^\ddagger\). This ability also contributes in classical cases of binaural audition to ripen the localization, to simplify the distinction front / rear and to locate sounds on the median plane. These interferences can be simulated thanks to Head Related Transfer Functions (HTRF) which are used with headphone reproduction. In order to be complete let us add that a priori knowledge of the source position increases the precision of detection.

It is possible to affirm that hearing is a multimodal sense, indeed, human unconsciously performs small movements of the head. By tightening the ear, he tries to obtain other sound pictures of the scene and so ripen the precision of detection. Moreover we feel sound waves with all the body, notably low frequencies that are more felt than ear.

Let us note that there are strong interactions between image and sound perception, here are two examples: Ventriloquism and McGurk effect [MD76] where vision of lips articulating a syllable whereas the sound emitted is another syllable makes the person recognize a third syllable different from the 2 others.

### 2.2.2. 3D sound reproduction device

As we have told in the introduction we are looking for the less intrusive interfaces, that is why we decided to make the sound spatialization with speakers rather than headphone which can become telling and a source of discomfort. We found the following advantages: no need to use generic HTRF that are by definition not those of our user. His ability to locate could taper off. As we’ve told in last paragraph, the user performs some unconscious movements of the head to ripen the localization. These movements will be executed, so we should use a very fine tracking to reflect these variations into the HTRF, so it gives the feeling that the sound comes from inside his head. The sound is a wave that is also felt by the foot, the bones and the body. It is impossible to simulate the very low frequency sounds that are more felt than eared.

![Figure 4: Position and direction of the speakers on the sphere.](image)

Hearing is an omnidirectional sense, so it is natural to represent the sound space by a sphere with user’s head as the center. To keep the fact that the system has to be easily transportable, we have decided to use only 8 speakers at the corners of a cube defined inside this sphere (see figure 4). However it’s clear that with more loudspeaker the reproduction would be more accurate. Our system would allow us to locate the origin of the sound on the plane defined by three speakers.

#### 2.2.3. Our system

The spatialisation of the sounds uses a Terratec 7.1 sound card, which was originally designed to make a planar spatialisation (front centre, front right, front left, surround right, surround right, rear left, rear right and sub-woofer). This configuration has been modified to an 8.0 system that creates a 3D space for the sound. The eight speakers have all the same characteristics and are placed at the corners of a cube, the sound recreated is not a hi-fidelity sound, because no sub-woofer is used, but a great part of the audible spectra (15 Hz - 22 kHz) is reproduced.

Three sound libraries have been tried on Linux, the firsts of a higher level than the third, OpenAL [Opea], FMOD [FMO] and Jack-it [Dav]. Both have systems to read and seek inside a sound file and they all provide buffers to manage communication between audio elements. OpenAL and FMOD provide ready-to-use API to play sounds through a classical sound system such as stereo or 5.1 using planar localization because no third dimension can be transmitted to the user. With Jack-it it is possible to share concurrent\(^\ddagger\) sound buffers with low latency between audio applications and devices. The eight speakers have been defined

\(\ddagger\) The shape of he pinna makes affect sounds with frequencies above 2000Hz.

\(\ddagger\) Operating at the same time
The advantage of Jack-it compared to Alsa or OSS is that it manages an optimized mix of the buffers before sending them to the device. The main drawback is that none of the high level elements such as the sample rate conversion, the management of interleaved audio files, the Doppler, the attenuation of some frequency according to the distance or the reverberation linked to the environment of the sound.

The aim of the spatialization is to place a sound source in a 3D environment, so there is no need to read stereo interleaved files, all files used by the system are mono WAV files.

The sample rate conversion have been made with time variant filters structure [PM96] that has the advantage to integrate the interpolation and decimation processes in the filter so that it provides a really optimized filtering for real-time applications. The conversions have been implemented for the classical frequencies: 44.1 kHz (CD quality), 48 kHz (DAT quality) and 96 kHz (studio quality). This module have been implemented in order to import easily sounds to the software, and to be compatible with the majority of the sound cards operating on Linux.

The modules of spatialization have been implemented thanks to LADSPA [LAD]; it is a simple API to develop audio modules that can be seen in figure 5. These modules have N inputs that can be muted and M Outputs. On each output the level can be checked (to display a vu-meter for example). The processing is the part that determines the modifications to sounds and / or the links between inputs and outputs. Many free modules have been implemented with this structure including a great collection from Steve Harris [Har]. Reverberation has been used to create some ambiances depending on the location of the scene; a great reverberation is used to simulate a cathedral or a dream, a small one for a very small room or to give the user a felling of confinement.

A LADSPA low-pass filter has been used to simulate the attenuation of high frequencies according to the distance between the user and the sound.

The module of spatialization is the last link of the chain, as shown in figure 6. It takes one input, the sound coming from the reverberation module, and has eight outputs linked to the devices on the sound card.

The position of the speakers and of the sound is converted from Euclidian space to spherical space with the head of the player as origin in both real and virtual worlds.

The output level for a speaker $i h_{i,vol}$ is calculated this way: The sound position is defined by $(s_\rho, s_\phi, s_\theta)$ with $s_\rho$ the distance, $s_\phi$ the zenith and $s_\theta$ the azimuth. The eight speakers’ positions are defined by $(h_i, \rho, h_i, \phi, h_i, \theta)$. The zenith and azimuth are then compared and if the following condition is true,

$$\left\{ \frac{-\pi}{2} \leq h_i, \theta - s_\theta \leq \frac{\pi}{2} \right\} \left\{ \frac{-\pi}{2} \leq h_i, \phi - s_\phi \leq \frac{\pi}{2} \right\}$$

Then the next relation is used to defined the amplitude panning :

$$h_{i,vol} = s_{vol} \cdot | \cos (h_i, \theta - s_\theta) | \cdot | \cos (h_i, \phi - s_\phi) |$$

$s_{vol}$ is the volume of the source. At last $s_\rho$ allows to modulate the sound volume functions of the distance between the virtual source and the user.

2.3. Synchronization of visual and sound events

It’s central for the success of a virtual reality application, that the different sensory events are synchronous and to preserve
the causality link between user’s actions and system’s reactions. That’s why our system used 2 threads, the main thread manages the inputs (capture of the LEDs) and the graphical part of the application. When the user or an object moves on the scene a message is send via a FIFO to the audio thread. When a message is received, the audio thread checks if the sound is too far from the player, in this case, the sound isn’t played and the corresponding output is muted. If the sound enters in the listening zone, the process (see figure 6) start. If the sound was already in the listening zone, only the calculation of the spatialization will be made.

3. Interaction

The virtual reality room is fit out two cameras located so that they can capture all the movements of the user. Those cameras are calibrated thanks to the library OpenCV [Open].

permit to extract the position in 3D of electroluminescent diodes places on the polarizing glasses of the user or on various interface objects such as a pen. We can observe that the LED on the glasses (see figure 8) permits to deduce position and movements of the head of the user and therefore user’s interest center. An horizontal variance of the head more significative is associate to user’s will to move in the scene. Our acquisition system being based on diode detection, it is easy to transform any common object to a tool for virtual reality. For example (see figure 9) two small welds can transforms an inoffensive child’s gun to a redoubtable weapon of precision in our room or an old paintbrush becomes a tool of the XXIth century.

Coats of paint are melted when the virtual brush come over older layer, if not dry, integrating the time factor. Always in the field of art, the Amorphous Ubiquity [AP04] is a real time interactive installation where the user is immersed in a perpetually moving 3D maze, the motion of the maze being determined by the user’s motion and speech.

A project in collaboration between the city of Serris, an archaeological team and SISAR has led to the creation of a virtual Merovingian environment whose data are based on vestiges discovered on the site of the city. In addition to the ability for the city to communicate on it’s history, this reconstruction is seen by the archaeologist as a completely new way to emit and to evaluate historical hypothesis.

The virtual reality room allows also to play 3D movies filmed with a device consisting of two calibrated cameras next to each others. The restitution implies a precise synchronization of the video-projectors.

The Puppet is an application that allows to manipulate an avatar. The extraction from a video flow of the user’s profile on a white background is done in real time. The skeleton is then analyzed to interpret the user’s will and the avatar is activated accordingly.

The last application making use of the virtual reality room presented here is an interactive game similar to "Mist" [MYS] (see figure 10). It has allowed to experiment the restitution of 3D scenes in a new fashion, the usual way being to project the three dimensional scene onto the screen. Here, the restitution is based on a dense set of photographs, on traditional screens, this process is known as "mosaic" but we have extended it to panoramic stereo devices. In practice, the scene is modeled in the usual way, a script then construct a dome of images from hundreds of renderings. In the game, during the motion of the user, the two most pertinent images are computed to design a stereoscopic image. With this technique, we have obtained real time frame per second rates ( > 25 fps ) with images for which the quality had necessitated extensive computation time.
5. Conclusion

In this article, we have presented our virtual reality room, its main qualities are its low cost, its mobility and its evolutivity. Apart from the screen, it is only constituted of consumer grade elements. From a computer science perspective, we have made the choice of using open source softwares, available under Linux.

The interest of this device is to allow the visual and auditory immersion of the user and, by the mean of sensors, its intuitive interaction with the environment.

For electronic games, this room allows new types of gameplay and amplify the players sensations. Moreover virtual reality does not limit to stimulating in an independent manner, the different senses of the user, but has to offer a multi-sensorial perception. The aim is there to propose new sensations, of intersensoriality, quasi synestesy.

References


Figure 11: A ludic application.