An Experimental Implementation of a Networked Hapto-Acoustic Virtual Reality Environment Applied to Surgical Education using the High Speed CeNTIE Research Network

Tony Adriaansen[1], Chris Gunn[2]

[1] CSIRO, Telecommunications and Industrial Physics,
P O Box 76, Epping, 1710, NSW, Australia.
http://www.tip.csiro.au
[2] CSIRO, Computing Mathematics and Information Services
Building 108 North Road, ANU campus,
Acton, 2601, ACT, Australia

Abstract. This paper addresses issues affecting the implementation of a networked Hapto-Acoustic Virtual Reality Environment. This is a Virtual Reality system which combines 3D graphics visualisation, audio and touch interaction and which aims to produce a perceived real time interaction by the user with a computer generated model. Networking systems together makes it possible to collaborate in the virtual environment and achieve a sense of perceived virtual presence. We use a hand-immersive virtual environment where the user directly interacts with objects in the 3D virtual world. We are also able to avoid the need for positional head tracking by making use of a space-ball mouse. The interactive space contains a PHANToM haptic device used to provide force feedback. Issues affecting successful implementation of the haptic virtual environment are mentioned and networking of two such systems is demonstrated using the high speed CeNTIE research network.

1 Introduction

Virtual environments (VE) mimic a real world experience to some degree and usually include more than one human sense. Typical Virtual Reality systems include a 3D graphical model with user interaction via some form of position tracking and display. The display device may be either a heads-up display worn by the user where the wearer sees the new space based on the current head orientation and position, a displayed view on a monitor or a larger panoramic projected view as used for example in flight simulator and other VR applications.

Virtual Environments are useful in a variety of domains including military, entertainment, education or medicine. Such virtual environments are especially beneficial for tasks where it is more difficult, more hazardous, where cost constraints would limit the availability of real world interaction or ethical issues preclude performing the task in the real world. Typical scenarios include potentially dangerous activities such as pilot training in a flight simulator, nuclear device handling, bomb disposal, hazardous environment training, medical procedure training and animal or human educational training. Networking Virtual Reality Environment systems opens up the opportunity for remote collaboration between users and makes learning at a distance feasible.

As appropriate human perception is paramount for successful VE implementation, some human factors are discussed in terms of visual, auditory and tactile response and mention is made of how the interaction of these can influence the perceived virtual experience.

2 Goals and Prerequisites

A goal of networking virtual environments is to enable a “sense of presence” between the near and remote users, such that interaction with the virtual world and communication between users is seamless and does not detract from the intention of the task. The users should not be concerned about how the system works, only that their interaction is achieving the correct outcome. Issues for system implementation must therefore include user needs and expectations as well as have sufficient informational cues and detail in the virtual world to enable meaningful interaction.

In terms of visual accuracy, the perceived spatial resolution should be high enough for object recognition by the user and contain enough visual information like colour and texture appropriate to the model. The choice of how the 3-dimensional virtual world is generated and how this is displayed can impact on the degree of realism experienced. It is also important to fulfil the goal of minimising some of the negative but unavoidable features of the technology while still allowing good human–machine interaction, for example the method of generating stereo for perceived 3D vision.

Audio signals need to be of sufficient quality, both for model interaction cues and for user communication when systems are networked. Synthesized sounds can provide information about the force and velocity of interaction gestures, surface properties of the object such as stiffness, roughness and texture, and global properties of the object such as shape and material [1].

For haptics, the interface to the Virtual Environment needs to address the scope and range of (hand) movement, available force feedback, speed of response and ergonomics.

Successful networking of multiple virtual environments depends to a large extent on the capabilities of the technology, in particular a high speed communications network, computer network interfaces fast enough to handle data throughput and enough computer processing power to handle data generated by the near, as well as the remote, user’s interaction while retaining “real time” performance. These goals are addressed in sections 3-7 below.

3 Human characteristics.

The hapto-acoustic virtual environment developed at CSIRO is known as a "Haptic Workbench" and makes use of three of the human senses: sight, sound and touch. We have not yet addressed the sense of smell, however other groups are actively looking at adding the olfactory sense to virtual spaces [4] as well as taste [5]. When designing a haptic system, it is necessary to be aware of a human's (fine) motor control for a successful VE touch interaction. Human performance in the three areas of visual, audio and touch are still in excess of the best achievable by a single machine today, although some machine simulations can be very good. Approximations and shortcuts are sometimes necessary in order to simulate a virtual scene to an acceptable degree for the required task as well as achieve an acceptable balance between model complexity and system response time.

Many Virtual Reality systems incorporate the sense of sight and sound, with the main stimulus and focus directed towards sight. The growth of VR systems using just these two human senses is not surprising as computer output has always included some form of visual information, and digital display graphics capabilities have been steadily improving over time. In fact some graphics cards on the market today are more powerful in terms of processing speed and on board memory than a typical complete personal computer system of only five to ten years ago [1]. Some VR system builders still approach audio as being of minor importance but advances in computer audio hardware including the recent "surround sound" concept [2] are making significant improvements in the field of digital sound reproduction. In order for a human operator to experience a sense of presence, the user's focus should be taken up by the virtual interactive task rather than being overly conscious of the need to conform to constraints made by the system.

4 Computer simulation of human performance

4.1 Spatial resolution for graphics and haptics.

There are conflicting requirements for simulated human perception of visual, touch and audio stimulus. The spatial resolution needed for realistic graphics rendering is much higher than that required for haptics. Typical video graphics cards can now display about 2,000 pixels in both horizontal and vertical dimensions which on a 21inch (53cm on diagonal) monitor translates to about 5pixels/mm. At a typical viewing distance, this resolution accuracy is necessary for good representation of graphics images. In comparison, the typical human fingertip tactile spatial position resolution control capability is about 2mm, or

[1] Graphics cards specifications of the type we use have data throughput speeds of 2GB/sec, include 256MB RAM memory, support 2048x1536 pixel resolution, have onboard hardware processing and can display 24 bit colour

[2] Multiple (6.1) channel "surround sound" with advanced AC-3 encoding (Dolby™ digital), has 16 bit encoding at 44.1 KHz and a maximum dynamic range of 96dB

about 20 times more coarse [3], while typical displacement resolution detection is of the order of 1 micron. The PHANToM we use has a position resolution accuracy of 0.03mm.

4.2 Video dynamics.

There are two refresh rates that are important when refreshing the graphical images: the screen refresh rate and the scene refresh rate. The screen refresh rate determines whether any flicker is apparent to the user's eye. This is important for comfortable viewing even if no objects being depicted are moving. Computer video update frame rates greater than only approximately 70Hz (non-interlace) are sufficient for the perception of flicker free images using displays with short persistence.

Scene refresh rate is the rate at which moving objects in the scene are repositioned. The human eye can resolve a series of stationary images into smooth movement if the refresh rate is not too slow. We have found that scene refresh rates of 20-30 Hz are acceptable, with the lower end of the range satisfying more experienced users, while novices seem to need a faster rate to give the illusion of smooth movement. Displays with longer persistence, like those used for Australian television reception, run at a frame rate of 25Hz (equivalent to 50Hz non-interlaced) and result in adequately smooth movement perception.

For 3D perception, stereo vision can be accomplished by using 3D stereo glasses. These are liquid crystal shutter glasses containing a sensor which alternately block light to left or right eyes, synchronised by a signal from the graphics card. Typical shutter speeds for these are about 100Hz (50 Hz to each eye) which results in no perceivable image flicker and no jerkiness of movement. Passive stereo is an alternative method of producing a moving 3D image. In this case the left and right eye view are simultaneously displayed, but with different polarization of the light. The passive stereo glasses have a corresponding polarizing filter on the left and right lens.

4.3 Haptics dynamics.

Human touch rate capabilities are much faster and frequencies up to 1kHz for tactile detection can be sensed. It has been shown that there are 4 groups of touch sensing nerves showing differing frequency and vibration thresholds, using a 1 micron size peak displacement stimulus [9]. Hand and finger movement and touch is referred to as fine motor feedback control response. This is the response time of tactile feedback by the skin coming into contact with an object and responding to the stimulus. The tactile feedback control response time has been found to be a function of the task, and ranges between 1-10Hz [6]. This is made up of the initial tactile detection, movement and control of limbs together with the deliberate conscious action associated with touch control.

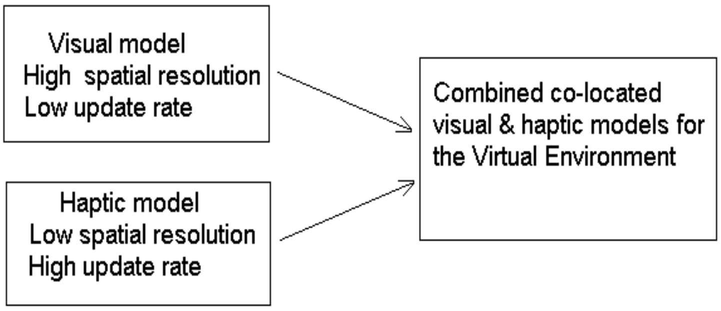
4.4 Sound.

Audio must be sufficiently well synchronised with the visually displayed information and quality of audio should be high enough for the task. It is particularly important to avoid missing data in the audio stream for useful audio in the VE. Signal processing available on many computer sound cards are able to generate 16 bit encoded sound at a digitised sampled rate of 44.1kHz with a maximum dynamic range of 96dB. This combined with good quality speakers results in audio quality equivalent to that found in CD recordings.

4.5 Video, haptic and sound interaction.

Due to the above differing requirements with respect to human perception and limitations in computer power, it is often possible to represent the haptic properties of a virtual object at a lower spatial resolution than its visual representation. Conversely, the update rate for the visual representation can be set much slower than that needed for the haptic representation. In this way, we can free up some valuable computing resources to allow a greater number or more precise objects to be represented while still retaining perceived real time performance. By co-location of the visual & haptic models the user's perception is one of being able to touch the virtual 3D objects seen in the VE. The concepts of co-registration involving scale, orientation and position are described by Bogsanyi & Krumm-Heller [7]. Another important factor in human perception is that the visual cortex is closely involved with certain tactile tasks and that visual processing was shown to be instrumental in ordinary tactile perception [8]. There is also evidence that audio dominates vision in temporal processing involved in sensorimotor coordination [10], so the 3 senses should not be considered in isolation.

Diagram showing visual model / haptic model separation & recombination.



Visual model
High spatial resolution
Low update rate

Haptic model
Low spatial resolution
High update rate

Combined co-located visual & haptic models for the Virtual Environment

5 System Description

5.1 Computer and haptic specifications

The computer system consists of a machine with dual 400MHz R12000 RISC processors with 2MB level2 cache, specialised dual-head graphics card with 128Mb and with stereo capability, 2Gbytes RAM, a Gigabit 1000Base-SX optical fibre network card, 2 high end 21” colour monitors and 2 ultra SCSI 18GB storage disks.

The haptic device is a PHANToM premium 1.5 with an effective workspace size of 19.5cm x 27cm x 37.5cm. The device has 6 degrees of freedom in position sensing (x, y, z, roll, pitch, yaw) and 3 degrees of freedom in force feedback (x, y, z). The PHANToM is capable of supplying a continuous force of 1.4N and a maximum force of 8.5N.

Layout of Haptic Workbench

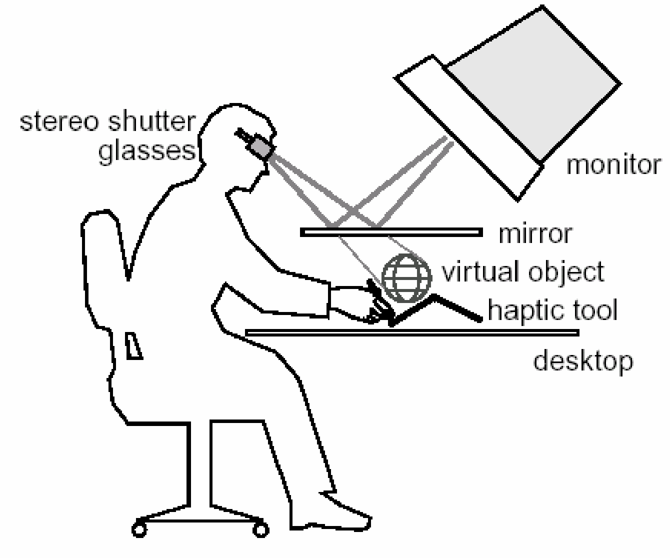


Figure 1: Configuration of the Haptic Workbench.

The haptic workbench combines stereo images, co-located force feedback and 3D audio to produce a small-scale hands-in virtual environment system [2] and fulfils many of the issues relating to successful implementation needed for a VE.

5.2 Network issues

Communication between systems is made possible by the CeNTIE network. This is a high bandwidth fibre optic research network made up of a collection of layered networks currently under construction and nearing completion.

The CeNTIE network uses no routing but is layer 2 switched and capable of network speeds up to 10GBits/sec, although parts of the system are currently configured to enable up to 1GBit/sec bandwidth (Fig 2). The CeNTIE backbone core network links all capital cities in Australia except Darwin and Hobart with network extensions to hospitals, media houses, Universities, research laboratories and commercial partners thus allowing very high speed collaboration between nodes.

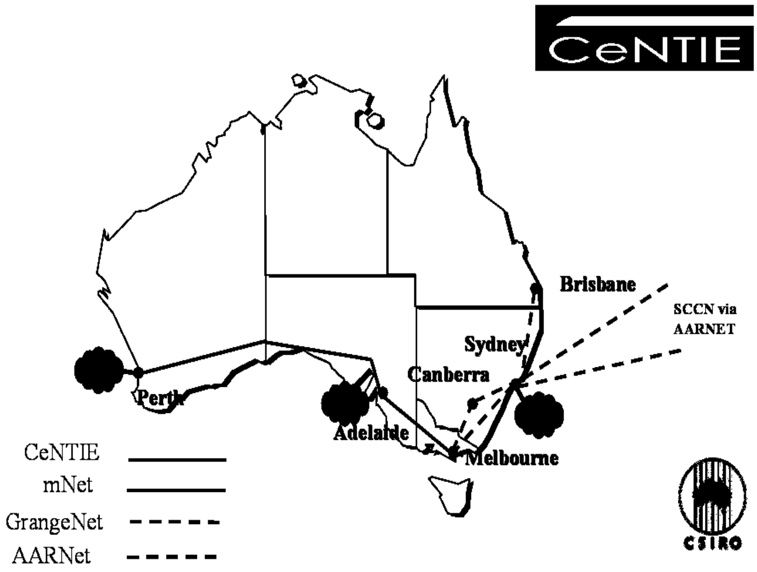


Figure 2. CeNTIE major node locations and extent of network.

6 Implementation

Test models were created using the National Library of Medicine “Visual Human Project” dataset. The Visible Human data consists of about 10,000 images using MRI, CT scans and digital camera colour photos taken from complete cadavers. For the male data set, the body was prepared and then sliced at intervals of 1mm from head to toe by researchers at the University of Colorado Health Sciences Center. A digital colour image was made in the axial plane after each slice was removed, generating about 1,800 cross sectional views of the complete body, (1.8m tall).

Segmentation was carried out on 203 downloaded images using the colour digital image slices, in order to capture a high resolution 3D “point–cloud” data set making up the liver surface. Various methods including Delaunay Triangulation and hybrid techniques were then applied to the surface points in order to generate a high resolution surface mesh of the liver. This resulted in a surface rendered model containing about 30,000 points and about 200,000 polygons making up the liver surface and represents the highest resolution model used in this study. Ten reduced versions of the full resolution (30,000) point model were generated. These have from 5% to 90% of the number of data points in the full model, an example is shown in figure 3 below.

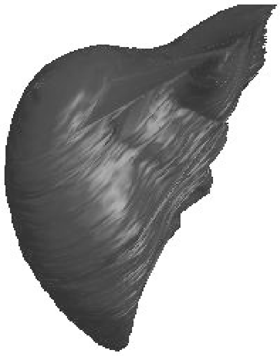


Figure 3. Example rendered male liver surface created using the Visible Human dataset. The model consists of about 30,000 points and 200,000 polygons.

Using the full graphics liver model but without haptics, we found the system response showed a significant rendering time resulting in a noticeably slow graphics refresh rate. Reducing the graphics model to 50% - 60% of the full model or about 15,000 points and about 100,000 polygons enabled a reasonable video response with negligible delay. In terms of the haptics model, noticeable system response delay was evident when the haptic model exceeded about 3,000 to 4,000 points with about 25,000 to 30,000 polygons.

As mentioned in section 4.5 above, the different resolutions are accommodated by the user, it only becomes obvious that perception of a mismatch occurs and the 2 models do not coincide when the difference between visual & haptic models is greater.

7 Network protocols

Networking two Haptic Workbenches can enable two people in different locations to simultaneously see, feel and manipulate objects in the VE. This scenario lends itself to networked surgical training, where an instructor and student may be separated by large distances, and opens up the possibility of distance learning including touch. For successful collaboration, it is necessary to have high enough network features like bandwidth, low latency and quality of service in order to preserve synchronous events in the VE while at the same time maintaining real time communication

We use both TCP (transmission control protocol), a positive acknowledgement protocol where the receiver sends a return signal that it received a data packet and UDP (user datagram protocol) where no acknowledgement is returned. There are two transmitted streams. One carries object movement data updated via TCP at about 20hz. The other carries haptics data and is sent via UDP at 400hz.

Until recently, providing good quality audio communication between users during a networked session was a two-fold problem. Firstly the audio quality was typically poor due to missing data and secondly audio feedback echo from the remote site was evident. Implementing RAT (Robust Audio Tool) for communication between users and using UDP transport has resulted in much better audio. Fitting each user with a headset containing headphones and microphone has practically eliminated feedback echo problems.

8 Results

An example simulated medical procedure was achieved on the CeNTIE network between Sydney and Perth in May 2003. Examples of the collaborative interaction show a 'surgeon' in Sydney guiding a 'trainee' in Perth to perform a simulated cholecystectomy, (gall bladder removal operation) is shown below.

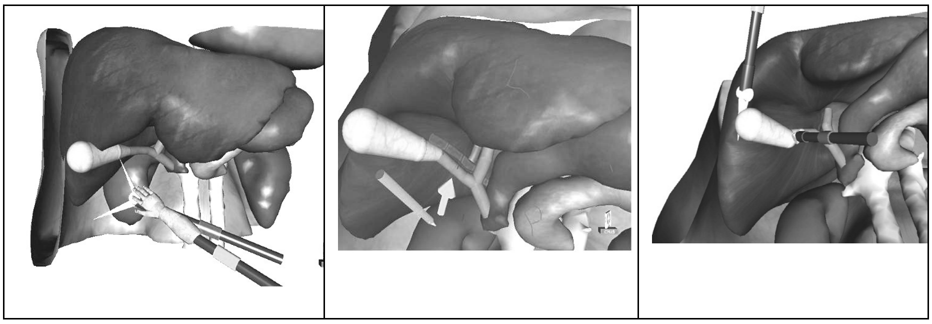


Figure 4. Shared haptic interaction.

(Left) "The guiding hand", here the trainee sees and feels force of movement generated by the instructor. (Center) Interactive drawing in the scene by both users. (Right) Both users applying forces to the virtual gallbladder which deforms in real time.

9 Conclusions

Some issues affecting successful implementation of a networked virtual environment have been discussed, together with some human perception factors. Complex human sensory response, sensory interactions and dynamics have been mentioned. These require varied solutions in terms of adequate simulation with the available computer technology. Some demonstration models have been designed and a successful networked haptic VE has been demonstrated using the high bandwidth CeNTIE network.

References

1. G. Castle, M. Adcock, S. Barrass. Integrated Modal and Granular Synthesis of Haptic Tapping and Scratching Sounds. *Eurohaptics 2002 Conference proceedings Edinburgh*, pages 99-102 Edinburgh 2002.
2. D. R. Stevenson, K. A. Smith, J. P. McLaughlin, C. J. Gunn, J. P. Veldkamp, and M. J. Dixon. Haptic Workbench: A multi-sensory virtual environment. *Stereoscopic Displays and Virtual Reality Systems VI, proceedings of SPIE* Vol. 3639, Pages 356–366, 1999.
3. N. I. Durlach, L. A. Delhorne, A. Wong, W. Y. Ko, W. M. Rabinowitz, and J. Hollerbach, Manual discrimination and identification of length by the finger span method. *Perception and Psychophysics*, 46(1), 29-38, 1989.
4. Y. Yanagida, S. Kawato, H. Noma, N. Tetsutani, A. Tomono. A Nose-tracked, personal Olfactory Display. *SIGGRAPH Sketches and Applications*, 2003.
5. H. Iwata, H. Yano, N. Uemura, T. Moriya. Food Simulator. *SIGGRAPH Emerging Technologies*, 2003.
6. T. L. Brooks, Telerobotic response requirements. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. pages 113-120 Los Angeles, California, 1990.
7. F. J. D. Bogsanyi, A. Krumm-Heller, Scale and Collocation in Hapto-Visual Environments, *Stereoscopic Displays and Virtual Reality Systems VII, Proc SPIE Vol 395*, pages 454-463, 2000.
8. A. Zangaladze, C.M. Epstein, S.T Grafton, K. Sathian. Involvement of Visual Cortex in Tactile Discrimination of Orientation. *Nature*, Volume 401, pages 587-590, 7 October 1999.
9. S. J. Bolanowski, G. A. Gescheider, R. T. Verrillo, C. M. Checkosky, Four channels mediate the mechanical aspects of touch, *J. Acoustic Society of America*, 84 (5), pages 1680-1694, November 1988.
10. B. H. Repp, A. Penel, Auditory Dominance in Temporal Processing: New Evidence From Synchronization With Simultaneous Visual and Auditory Sequences, *Journal of Experimental Psychology-Human Perception and Performance*, Vol. 28, No. 5, 1085–1099, 2002.