The MiniCAVE - A Voice-Controlled IPT Environment

Edward J. Wegman, Jürgen Symanzik, J. Patrick Vandersluis, Qiang Luo, Fernando Camelli, Antoinette Dzubay, Xiaodong Fu, Nkem-Amin Khumbah, Rida E. A. Moustafa, Robert L. Wall, Ying Zhu

George Mason University, Center for Computational Statistics 4A7, Fairfax, VA 22030, USA

Tel.: (703) 993-3786, FAX: (703) 993-1700
e-mail: ewegman@galaxy.gmu.edu, symanzik@galaxy.gmu.edu

Abstract

The computer hardware for the CAVE and similar immersive projection-technology (IPT) virtual reality (VR) environments are normally high-end Silicon Graphics systems with expensive CRT-based projection systems. The idea behind the MiniCAVE concept is to downscale the computing hardware to fast PCs running Windows NT and to downscale the projection systems to relatively inexpensive and relatively robust LCD projection systems.

Much to our surprise, when we started this project in early 1998, the performance of a 200 megahertz Pentium Pro (at $3,000) was competitive with the performance of an SGI Onyx RE2 (at $120,000) when running matrix-oriented mathematics software. This suggested that a much cheaper version of the CAVE might be assembled using Pentium II hardware. Currently, we are working with a PC that has two 466 megahertz processors. Also, both resolution and frame rates of LCD-type projectors have improved substantially during the last few years. Brightness already is substantially better than CRT systems. This makes the development of an IPT environment, the MiniCAVE, based on PC technology and LCD-type projectors feasible. The MiniCAVE should cost less than $100,000 after development.

Another problem that typically occurs with full scale CAVE environments is the large amount of space; often 12x12x12 foot plus extra space for the rear projection. We propose a 6x6x6 foot MiniCAVE. The reduction in dimension allows that this IPT environment can be placed in many regularly sized rooms available at universities and companies. Also, the halving of the linear dimension of the projection walls matches the resolution capabilities of LCD projectors.

As a first software example (and just using one PC), we have successfully ported a stereoscopic fly-through demonstration package from SGI Unix using GL to Windows NT using OpenGL. Windows NT supports the CrystalEyes liquid crystal shutter glasses so that stereoscopic displays are available in a PC environment. While most PC applications use the desktop metaphor for navigation, moving to a virtual reality environment suggests another metaphor that is more appropriate. We have implemented a voice-command interface for the Windows NT environment so that the user can control features of our stereoscopic fly-through demonstration by simply using voice commands.

Keywords

PCs, Windows NT, LCD Projection Systems, Speech Recognition, Virtual Reality.

1. Introduction

Immersive projection-technology (IPT) virtual reality (VR) environments such as CAVE [URL3, 2, 3, 4], Virtual Portal [5], C2 [URL1, 13], CUBE [URL5], CAVEEE [URL4], or CABIN [URL2, 7], are considerable improvements of other VR technologies, e.g., head mounted displays (HMD) and boom technologies, in a number of ways. The HMD technologies usually consist of a piece of headgear with either binocular color ray tube (CRT) or liquid crystal display (LCD) display devices and often binaural audio reproducers. Most users experience discomfort in their weight and often users discontinue the use of these units not only because of their awkwardness, but also because of the inherent latency in position tracking of the head. In addition, there are legitimate concerns about the use of CRT displays so close to the eyes. This implies that additional shielding has to be done – thus adding weight and therefore further inhibiting the use of these devices. Finally, in order to present an immersive view, head tracking is a necessity. The tracking sensors coupled with even modern computing speeds still produce a latency in response often on the order of milliseconds to tenths of seconds depending on the complexity of the image. The implication is that the sense of balance provided by the inner ear and the visual sense are out of sync and the resulting conflict often gives HMD VR viewers cybersickness [12]. The boom is a binocular-like alternative mounted on the end of a boom. This display device allows for a full 360 degree field of view, but restricts the user to motion around the pivot point of the boom. The boom tracks the head position by mechanical means and therefore does not exhibit the same latency defects of the HMDs. Nonetheless, it is an awkward device and never seemed to have caught a wide usage. See [1] and [10] for example for an in-depth discussion of non-IPT VR devices and standard VR input devices.

The CAVE and similar IPT devices are by far the most satisfactory of the VR devices. Consisting of approximately a 12x12x12 cubic foot enclosure of translucent material, the CAVE uses rear-mounted CRT projectors to surround the user with a complete visual environment. Typically, the projectors are high-end systems capable of high frame rates. Used in conjunction with CrystalEyes liquid crystal shutter glasses and a tracking device, the CAVE allows total immersion in a completely three-dimensional environment. The viewpoint is dynamically adjusted so that no matter where the user looks the CAVE provides the view from the user’s perspective. The CAVE is usually driven by high-end multiprocessor computers, usually Silicon Graphics machines equipped with multiple Reality Engine[x] graphics subsystems. The most compelling drawback to the CAVE system is the cost, an estimated $500,000 to $2,000,000, per copy.

Our idea of a MiniCAVE is motivated by several recent developments and our experience [15] with what might be called a PlatoCAVE. The matrix-oriented mathematics software MATLAB was recently released in version 5. We acquired copies for both our SGI Onyx RE^2 computers (at $120,000) and for more pedestrian PCs, e.g. 200 megahertz Pentium Pros (at $3,000). Much to our surprise, when we ran the benchmarks, the Pentium not only held its own, but beat the Onyx in several categories. This suggested that a much cheaper version of the CAVE might be assembled using Pentium II hardware, especially say at the 300 or 466 megahertz speed. Moreover, both resolution and frame rates have improved substantially on LCD-type projectors during the last few years. CRT-based projectors, such as the Stereographics unit we own, must be permanently mounted and are subject to misalignment and damage by shock. The reason is that the three CRTs must be precisely aligned for proper resolution and color. Without the proper alignment, not only resolution is affected, but color fringing can destroy stereoscopic effects.

We are in the process of developing a MiniCAVE based on PC technology and LCD projection systems. Rather than creating a 12x12x12 foot CAVE which, when completely set up, occupies a volumetric space that is beyond the capacity of most normal rooms with 8 to 10 foot high ceilings, we restrict the size of the MiniCAVE to 6x6x6 foot. The halving of the linear dimension quarters the required projector resolution and places the resolution requirement well within the capabilities of LCD projectors. Coupled with added brightness, less sensitivity to misalignment, and relative insensitivity to shock, the LCD projectors would seem to be a great step forward.

Our experience [15] with a one-walled IPT device (a PlatoCAVE - after Plato's allegory of the cave) that does not support head tracking suggests that as long as an individual's viewpoint is not too far from the nominal viewpoint, the stereoscopic effect is not effected too much. This fact is currently exploited in many IPT environments when an expert guide (who is head-tracked) introduces several people at a time (who are not head-tracked) to a particular application – implicitly assuming that the audience stays close to the guide and does not explore the environment by themselves. One constraint of a 6x6x6 foot MiniCAVE is that an individual's movement is more restricted. However, the lack of movement implies that the need for head tracking may be dispensed with.

Less required resolution coupled with no tracking requirement implies much less demands on the computation engine - again suggesting that this is well within the capability of a 300 or 466 megahertz Pentium II. The smaller size also means that the MiniCAVE will fit into normal-sized rooms. The projection systems involved are approximately $7,000 and the computers on the order of about $4,000. This suggests that the of-the-shelf hardware requirements are on the order of $55,000 for a five-walled MiniCAVE. With translucent projection screens, mirrors, and structural hardware, the MiniCAVE should cost less than $100,000 after development.

Compared to this amount, the cost for off-the-shelf speech recognition software, often in the range of $200 to $500, is not a significant cost factor. However, using a speech interface in a VR environment is a far more natural way to operate an IPT device than to use obscure pointers to select options from a standard 2D pop-up menu that has been enhanced to 3D or to locate and operate a complex user interaction menu that is often hidden in a corner of the IPT device and thus distracts the user from the main activity when selecting a new feature. Our experiments suggest that off-the-shelf speech recognition software is ready for use in IPT devices.

In Section 2 of this paper, we will discuss the MiniCAVE hardware. Section 3 deals with its speech recognition component. The current development stage is discussed in Section 4.

2 The MiniCAVE Hardware

2.1 Computer Requirements

Multiple computers with graphics processors are focal components of the MiniCAVE. The combination of multiple computers, each producing stereoscopic images independently, is a key feature of the system. The computers are used to transmit stereoscopic graphical information to a LCD projector. A preferred form of transmission of the graphical information is in the form of a time-sequential (left-eye, right-eye) stereoscopic image signal, but is not limited to this method. Indeed, recently polarized light stereoscopic LCD projection systems have become available. The computers are also used to transmit synchronization signals to a signal emitter that controls the synchronization of stereoscopic images of the viewer.

The transmitted graphical information can be retrieved from a sub-component of the computers, a distant storage medium via computer networking, or it can be generated dynamically by the computer, itself. The graphical information is generated by computer code based on internal components of the software itself, or by software manipulation of data such as numerical or CAD data stored on the computer's storage medium or accessed from remote sites via computer networks, or by interaction of two or more autonomous virtual reality systems via computer networking each system providing the other with stereoscopic images.

In our original implementation, the computer was a single processor 300 megahertz Pentium II machine. Regrettably this machine was stolen while we demonstrated the system at a conference. The replacement machine is a 2 processor 466 megahertz Pentium II system.

2.2 Synchronization Signal Generator

When multiple computers are used to generate and display the three-dimensional environment on a screen, a mechanism is required to synchronize the independent computers. Temporal synchronization is necessary to align images generated by autonomous, multiple, computers so that the viewer is confronted with a continuous display that mimics the real world. Synchronization is needed at two levels. First, synchronization must be achieved so that images displayed by the projector are in sufficiently close temporal alignment so that blending of the images is achieved as perceived by the human visual system. This synchronization requires that the images displayed by each projector be no more than 1/100th of one second delayed from fastest to slowest image. We designate this type of synchronization as "image lock". Second, synchronization between the time-sequential images for left eye and right eye is required so that all projectors display left-eye information simultaneously and similarly display right-eye information simultaneously. The required synchronization is within approximately 1/150th of one second. We designate this type of synchronization as “stereo lock”.

A synchronization signal generator is achieved by subprocesses running under a multiprocessor operating system on the multiple independent computers communicating via ethernet or similar computer networking scheme with speed capabilities of at least two megabits per second. One of the independent computers for the system is designated as the “master” and the others are designated as the “slaves”. The stereo lock is achieved by master computer broadcasting a message via the computer network connection to each of the slave computers indicating which of the left or right eye images are to be displayed. This message needs only to contain a single bit of information plus routing overhead which is limited to a single packet of information. A packet containing 64 bytes or 512 bits would be available in less than 3/10,000th of one second on a two megabit per second computer network easily within the 1/150th of one second requirement for stereo lock. The image lock synchronization works by having each slave computer reporting to the master computer when the slave computer has finished computing its current frame. Until each slave (and master) have completed computing the corresponding current frame, all computers display and re-display the previous frame. When the master computer has received messages from each slave computer that the next frame is computed, and when the master computer itself has completed the next frame computation, the master computer broadcasts a signal to all slave computers to display the next frame. The next frame packet is similar in size to the stereo lock packet so that switching to the next frame can occur within the same 3/10,000th of one second time scale. The computation time of individual frames may vary depending on the complexity of the image from 1/15th of one second to 1/150th of one second.

2.3 Viewer, Signal Emitter, and LCD Projector

In our first implementation, we used a simple CRT screen which was quite successful in demonstrating the Windows NT/StereoGraphics LCD shutter glasses combination. This scaled to our CRT based projection system with no difficulty demonstrating the feasibility of using Windows NT/time-sequential stereo for an IPT environment. However, the first attempts to use LCD (or DLV) projection systems revealed significant differences in operating characteristics between CRT and LCD systems. LCD systems are inherently capable of sufficient switching speed to support stereoscopic displays. The CrystalEyes shutter glasses are themselves LCD based. However, the design of LCD projectors requires a shift in the operating characteristics to make them effective in a stereoscopic application.

In essence, the difference is as follows. We assume first a time-sequential, shutter glasses form of stereoscopic display. As the electron beam begins the scan of the top scan line say for the right-eye image, the bottom line of the left-eye image has faded due to the limited persistence of the phosphor. As the right-eye image develops (at frame rates of 90 to 120 frames per second), the human eye is able to assemble a complete image in spite of the fact that the top of the image has long faded by the time the last line of the right-eye image has been drawn. Thus the fading of the CRT drawn image implies that there is no optical cross-talk. LCD projectors have a different scheme. As a pixel location is turned on by a scan, it remains on until it is rescanned. Thus for an LCD projector, as the top scan line of the right eye is drawn (and presumably the LCD shutter glasses allows transparency for the right eye), the remaining left-eye image is still turned on. This remains so through the scan. Hence it is only when the last line of the right-eye image is drawn that the whole of the right-eye image is available. Thus at any point during the time the right-eye shutter is transparent, the right eye will see a mix of both right eye and left eye images. Because of the persistence associated with human vision at these frame rates (85+), in essence both eyes see both images and sequential stereoscopic devices are useless.

There are two possible alternatives both of which we are pursuing. The first is to modify the BIOS of the LCD projector so as to induce an artificial blanking analogous to that which occurs with a CRT projection system. The alternative is to abandon the time-sequential stereo and use a polarized light stereo. The latter has some appeal because rear projection screens that preserve polarization are available. Polarized light projection systems can be inherently brighter than the artificially blanked LCD projector and also have the advantage of not requiring stereo-lock synchronization. Both of these avenues (LCD blanking and polarized light) are currently under investigation.

3. Speech Recognition

Standard human-computer interaction is typically based on the desktop metaphor, also called WIMP (windows, icons, menus, and pointing devices) user interfaces in Van Dam [14]. In this mode of operation, various “windows” containing graphic, icon, or text information are presented on a two-dimensional screen as if they were sheets of paper sitting on a desktop. Control of the computer is by means of pointing and clicking using a mouse and keyboard. This mode of interaction is suitable for a two-dimensional environment, but inappropriate for a three-dimensional environment. However, as Van Dam [14] points out, a new generation of post-WIMP user interfaces including speech recognition will be available soon for general use. Van Dam predicts that “voice recognition based on (limited) natural language understanding will be a dominant form of user-computer interaction”. Similarly, Patch and Smalley [9] summarize predictions that speech recognition will have a considerable input on everyday computer usage beginning as early as in the year 2000 or 2001.

We include as part of the MiniCAVE a metaphor involving voice interaction. Previous virtual reality systems have included instrumented gloves or wands with triggers that tend to mimic a three-dimensional version of the desktop metaphor. Because these are essentially means for interacting with two-dimensional windows in three-dimensional space, they tend to be awkward to use. In addition, the tracking required for position information on the glove or wand systems tend to be computational intensive and introduce time latency and position inaccuracy into the system making the entire system suboptimal. In our system, described as a voice command metaphor, a limited vocabulary is introduced analogous to the commands found in pull down menus in the desktop metaphor. Spoken commands are used in addition to traditional VR input devices. It is not our intend to completely replace wands, data gloves, or other currently used VR input devices. In reality, there exist scenarios where we cannot purely rely on our voice but where we need our hands as well. Obviously, to create realistic worlds in a VR environment, we have to consider all possible interaction devices that mimic the real world best. Sometimes speech is best, sometimes our hands are best. Also, as Stanney [11] points out, “multi-modal interaction may be a primary factor that leads to enhanced human performance for certain tasks presented in virtual worlds”. In the MiniCAVE, a unique software code layer interfaces standard applications such as the virtual reality and other graphical displays and the speech recognition software.

The use of speech recognition software has the goal of substantiating the use of speech-enabled command-input devices within the MiniCAVE. We hope that the ability to control objects within the environment through voice directives, without having to refocus eyes on a keyboard (real or virtual) and menu, will enhance the human sense of immersive interaction. This hope is borne on some excellent commercial speech recognition technology that is now emerging after 35 years of research.

This section of the paper will briefly review what is going on “under the hood” of speech recognition tools. We do this expressly to gain an appreciation for the difficulties of implementing this technology for a general audience of speech sources. While speech recognition products are fascinating to any end user, the underlying technology will be even more fascinating to scientists and engineers. This is because this technology represents an arguably successful integration of various fields of research performed over the last 35 years: socio-linguistic; articulatory phonetics; acoustical signal processing; sound card and microphone technology; pattern recognition; computer design; software engineering; database design; neural networks; and technology transfer marketing.

3.1 Overview

Many people immediately become fascinated with the idea of controlling a computer application with voice commands from the first moment of seeing an application accurately respond to a stream of voiced directives: opening files; cutting and pasting text; printing documents; closing windows; and copying files. In the same manner, being able to dictate a letter to a machine and watching it being produced without any human intervention other than the speaker’s voice is, perhaps, even more fascinating. For example, speech-activated computers allow a user to dictate between 30 and 70 words per minute [8, p. 2]. This is the average rate of speed for an experienced touch typist!

Slowly, after the novelty of it all fades away, as it does with most new, “cool” computer things, the exclamatory statements turn from “HOW does it do that!?” into “Why did it do THAT?!” We begin to discover that the speech recognition tool has a 5 to 10 percent (or higher) error rate and we need to invest time in training its algorithms to better “understand” what we say. Then we discover that, after extensive training, the application does not work for our co-worker as well as it used to work for her before the training.

3.2 Theoretical Perspective

To begin to understand the speech recognition process we have to start with the mechanics of human sound production and how information is transmitted through a system of utterances. We humans seem to have little trouble deriving a great deal of meaning from our highly evolved speech signaling system. Speech conveys linguistic information (i.e., the meaning of the utterance); socio-linguistic information (e.g., where the speaker comes from and, perhaps, a particular socio-economic class); and personal information (e.g., the identity of the speaker, voice quality and articulatory habits as distinguished from other speakers). We use context to differentiate homonyms and, conversely, words that are spelled the same but have different pronunciations [6, p. 3].

Machine speech-recognition starts with the basic unit that describes how speech conveys linguistic meaning: the phoneme. A phoneme is a group of similar sounds, not a sound per se, that is felt to be the same by the speaker - a set of units required for representing (writing down) utterances in an unambiguous manner (e.g., “thigh” and “thy” are phonemically two different sounds in English). There are, depending on the dialect (e.g., not being able to distinguish between “which” and “witch”), about 19 consonant phonemes in English. Vowels are much more affected by dialect differences and, therefore, are more difficult to list as contrasting items [6, pp. 4-7].

Machine analysis (and synthesis) of speech must utilize what is known about the parts of the human vocal apparatus as it is powered by the respiratory system. For example, there are seven places in the vocal cavity identified with articulation, starting with the two lips (bilabial) and ending with the back of the tongue on the soft palate (velar). Each of these places of articulation has six manners for articulation: stops (e.g., “pie”, “buy”, “guy”); fricatives (e.g., “thigh”, “sigh”, “shy”); approximates (e.g., “you”, “we”); trills (e.g., “rye” in Scottish English); taps (e.g., “letter”, “Betty”); and laterals (e.g., “lull”). There are combinations of these six manners of articulation as well. For example, the end of the word “church” is a stop combined with a fricative. One can chart English consonants into a matrix that maps place of articulation to manner [6, p. 9].

Vowels are better understood by the shape of the vocal tract as a whole: (1) the size of the minimum cross-sectional area Amin; (2) the location of Amin in terms of its distance from the glottis, the space between the vocal cords; and (3) the magnitude of the lip opening. Unlike some consonants (e.g., the bilabial stop “pie” versus “buy”), nearly all vowels are voiced (i.e., they are produced from vibrating vocal cords with a pulse of air from the lungs). A vowel sound, therefore, produces three to five frequency bands simultaneously. These bands represent resonant frequencies whose spectral parameters depend on the particular shape of the vocal tract [6, p. 10].

A very important aspect for deciphering of speech sounds by machine is the determination of the vocal tract resonant bands known as formants. One can come up with mean values for these frequencies in, say, American Midwestern English, but considerable variations from these means will occur for individual speakers even in the same dialect. Individual variations are largely due to differences in head size. For example, female speakers have formant frequencies that are on average 17% higher than males have. These spectral differences between speakers are what makes speaker independent speech recognition systems so challenging [6, pp. 12-16].

In the foregoing discussion, speech could have been depicted as sequences of "spectral photographs" of phonemes, vowels, or words frozen at instances of time. In fact, spectrographic analysis (e.g., Fourier analysis) is used to determine the acoustical parameters of most speech sounds. It is possible to distinguish between sounds such as "heed", "hid", "head", "had", "hod", "hawed", "hood", and "who'd" by examining the frequency components as a function of time. The rub is, formant frequency differences are not even consistent across speaker classes, especially between male and female.

Analyzing utterances discretely (as some speech recognition tools only do) ignores the fact that speakers normally produce quasi-continuous trains of utterances (e.g., phrases or sentences) to transmit meaning. In continuous speech, the vocal tract is no longer at a "steady state" and responds to dynamic forces that distort reference articulation patterns found with discrete speech. Stops may not fully form, fricatives may not produce the same degree of turbulent airflow as under discrete speech, or taps may now operate such that "Betty" goes to "Beddy". Can speech recognition tools compensate for these speaking-rate effects and make running comparisons, over time, of incoming speech with stored reference patterns? The answer is "Yes they can"!

3.3 Commercialization Progress

McPherson [8, p. 82] provides a time-scale for the commercial market to see various aspects of speech recognition challenges to be solved. However, in 1995 McPherson did not see large-vocabulary (10,000 to 30,000 words), speaker-dependent, continuous-speech systems on the market until the year 2004. Then, he also did not see speaker-independent systems for another 15 years. However, in the fall of 1997 both IBM and Dragon Systems, Inc., released new products with claims of being both speaker independent (i.e., no training required) and continuous (i.e., no pauses needed between words). Each had vocabularies exceeding 30,000 words and claims of being 95% accurate "out-of-the-box" (i.e., a proposed definition of "speaker independent").

In a more recent prediction by computer industry experts [9] in 1998, speech recognition technology is seen as captivating information technology managers in a real way within the next few years. In fact soon, Microsoft is expected to start bundling speech recognition with their Office suites or operating system. Also soon, third party speech recognition vendors will be integrating their products with individual applications by way of Microsoft's Speech API (SAPI) or Microsystems' Java Speech API. Some think that this technology has the potential of replacing the Windows desktop metaphor.

McPherson [8, pp. 1-7] recognized several factors acting to create a strong niche for this technology. Among them he described the Pentium factor: the arrival of desktop computers with enough memory and processing speed to perform the necessary operations. He also predicted sound card technology that can offload the CPU from having to perform the A/D or D/A conversions and digital signal processing.

McPherson [8, pp. 80-81] also described the possibility of successfully approaching continuous speech using "phrase technology" which recognizes that users do not actually speak continuously, but in phrases. The length of an utterance is of prime importance to accuracy, as is the ability to sense the spaces between words. The longer the utterance, the more information an algorithm has to resolve ambiguities and word-usage rules and to “sense” small, frequently used, but difficult to recognize, words like “is” and “of”.

3.4 Implementation Implications

In continuous speech, sensing “spaces” is difficult. Discrete speech systems are “word-based” and require pauses between words to assist in pattern matching. A more successful commercial strategy for continuous speech is to sense the phonemes first, then build up the words, and then the sentences (using context and grammar rules). Discrete speech systems require less processing power, work well with small vocabularies, and are very well suited for command-and-control (versus dictation) situations. Continuous speech systems, on the other hand, require more processing power, memory, vocabulary sizes, and rules as they include more extensions: context analysis; sentence analysis; paragraph analysis; and document analysis.

Promising strategies for achieving speaker independence include applying clustering techniques to large collections of isolated word samples used to fabricate speaker-independent templates [6, p. 198]. A large number, say n, of representative users is selected to utter each word of the vocabulary m times and, thereby, assign the (m x n) samples into a single cluster corresponding to a single word. Usually, about 8 to 14 references (users) per word give a fairly good account of the different ways of pronouncing a word [6, p. 199]. This technique does increase the recognition rate, but can be very time-consuming for large vocabulary systems. Commercial approaches to this problem include shipping products with large sets of speaker-independent templates. These are used initially, but the product then allows the user to alter these templates in order to adapt them to speaker-dependent templates. To maintain speaker independence, the user must make sure the original templates are not themselves “adapted”.

4. The Current Development Stage

The current state of the computer and projection system implementation is discussed in Section 2 above. We did our initial experiments with two commercial low-cost speech recognition software packages, i.e., ViaVoice Version 4.1, by IBM, and Dragon Dictate Classic Version 3.01, by Dragon Systems, Inc. The main requirements for the speech recognition software in the MiniCAVE environment are (i) the ability to enter and recognize commands, (ii) the ability to define and limit speech recognition to a small specialized vocabulary, (iii) the ability to associate more than one pronounciation with the same word, and (iv) easy-to-use APIs for incorporating speech recognition into another application. As it turned out, Dragon Dicatate met our expectations while ViaVoice did not (however it could have been augmented with several add-ons that eventually might have provided the missing capabilities).

As a first application, we ported SGI’s SkyWriter fly-through demonstration which has originally been developed for SGI workstations with GL 3D graphics support to the Windows NT platform. OpenGL API calls have been used to replace GL calls and the GLUT OpenGL utility library has been used to handle mouse and keyboard input. The code has been compiled using the Visual C++ 5.0 Compiler on the NT. To produce stereoscopic images on the NT, an above-and-below format has been used for the left-eye and right-eye view. Using CrystalEyes LCD shutter glasses and a screen refresh rate of 120 hertz for the NT monitor allow to produce a flicker-free stereoscopic image on the 300 megahertz Pentium.

In the next step, we linked the fly-through demonstration and the Dragon Dictate speech recognition software through a Parallel Virtual Machine (PVM) interface. The fly-through has been extended to react to a small set of spoken commands such as up, down, left, right, fast, slow, etc. As it turned out, the system worked within the specifications for speakers with regular American accents when it came to the identification of single spoken commands. However, it required a few minutes of training for speakers with different accents.

The main drawback of running speech recognition and an interactive, graphically intense application on a 300 megahertz Pentium was a temporary freezing of the graphical application when a new command has been spoken. All the CPU power has been used to identify the spoken word. However, a much better performance has been achieved on our new Pentium with two 466 megahertz processors. We expect that the next NT hardware generation will be able to handle two complex tasks such as recognizing a spoken command and continuously updating a complex stereoscopic graphic without any visible delay. Based on our initial experiments and developments, we believe that it is technically possible to fully implement the MiniCAVE as a voice-controlled IPT environment that runs on fast Windows NT platforms.

However, there remains one major technical hurdle. In a CRT-based projection system, the CrystalEyes LCD shutter glasses alternate at 120 frames per second. These work effectively because by the time the left-eye view is ready to project, the CRT projectors phosphor trace for the right eye has decayed. However, with digital LCD projectors, there is no phosphor decay. Hence the right-eye view overlaps with the left-eye view destroying the stereoscopic effect. We recently found a vendor that provides LCD technology that seems to solve this problem. We are eagerly awaiting the arrival of this unit to continue our work on the MiniCAVE.

Acknowledgments

Part of this work has been conducted in the CSI/INFT 979 class Virtual Reality/Scientific Visualization at George Mason University during the Spring Semester 1998. The work of Edward J. Wegman and Jürgen Symanzik was supported in part by the NSF with a Group Infrastructure Grant DMS-9631351. In addition, the work of Edward J. Wegman was supported in part by the Army Research Office under grant DAAH04-94-G-0267. CAVE is a trademark of the University of Illinois at Chicago, MiniCAVE is a trademark of George Mason University. A U.S. Patent disclosure has been filed on the MiniCAVE.

References

[1] Cruz-Neira, C.: “Virtual Reality Overview”, SIGGRAPH ’93 Course Notes #23, pp. 1-18, 1993.

[2] Cruz-Neira, C., Leigh, J., Papka, M., Barnes, C., Cohen, S. M., Das, S., Engelmann, R., Hudson, R., Roy, T., Siegel, L., Vasilakis, C., DeFanti, T. A., Sandin, D. J.: “Scientists in Wonderland: A Report on Visualization Applications in the CAVE Virtual Reality Environments”, IEEE 1993 Symposium on Research Frontiers in Virtual Reality, pp. 59-66, 1993.

[3] Cruz-Neira, C., Sandin, D. J., DeFanti, T. A.: “Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE”, ACM SIGGRAPH ‘93 Proceedings, Anaheim, CA, pp. 135-142, August 1993.

[4] Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., Hart, J. C.: “The CAVE: Audio Visual Experience Automatic Virtual Environment”, Communications of the ACM, Vol. 35, No. 6, pp. 64-72, 1992.

[5] Deering, M.: “Making Virtual Reality More Real: Experience with the Virtual Portal”, Proceedings of Graphics Interface ’93, pp. 219-226, May 1993.

[6] Fallside, F., and Woods, W. A.: Computer Speech Processing. Prentice-Hall, Englewood Cliffs, NJ, 1985.

[7] Hirose, M.: “Development of an Immersive Multiscreen Display (CABIN) at the University of Tokyo”, International Immersive Projection Technology Workshop, pp. 67-76, 1997.

[8] McPherson, M.: Executive Guide to Speech-driven Computer Systems, Springer, Berlin, 1995.

[9] Patch, K., and Smalley, E.: “Speech Recognition Makes Some Noise”, Infoworld, pages 69, 74, February 2, 1998.

[10] Pimentel, K., and Teixeira, K.: Virtual Reality Through the New Looking Glass (2nd Edition), Windcrest/McGraw-Hill, Blue Ridge Summit, PA, 1995.

[11] Stanney, K.: “Realizing the Full Potential of Virtual Reality: Human Factors Issues That Could Stand in the Way”, Proceedings: Virtual Reality Annual International Symposium ‘95. IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 28 -34, 1995.

[12] Stanney, K. M., and Kennedy, R. S.: “The Psychometrics of Cybersickness”, Communications of the ACM, Vol. 40, No. 8, pp. 66-68, 1997.

[13] Symanzik, J., Cook, D., Kohlmeyer, B. D., Lechner, U., Cruz-Neira, C.: “Dynamic Statistical Graphics in the C2 Virtual Reality Environment”, Computing Science and Statistics, Vol. 29, No. 2, pp. 41-47, 1997.

[14] Van Dam, A.: “Post-WIMP User Interfaces”, Communications of the ACM, Vol. 40, No. 2, pp. 63-67, 1997.

[15] Wegman, E. J., Luo, Q., Chen, J. X.: “Immersive Methods for Exploratory Analysis”, Computing Science and Statistics, Vol. 29, No. 1, pp. 206-214, 1998.

[URL1] C2: http://www.icemt.iastate.edu/about/selab/index.html
[URL2] CABIN: http://ghidorah.t.u-tokyo.ac.jp/Projects/CABIN/index.html
[URL3] CAVE: http://www.evl.uic.edu/EVL/VR/systems.shtml
[URL4] CAVEEE: http://vr.iao.fhg.de/ccvr/ausstattung/main.htm
[URL5] CUBE: http://www.hlrs.de/structure/organisation/vis/velab/