A Stereo Display Prototype with Multiple Focal Distances

Kurt Akeley* Simon J. Watt† Ahna Reza Girshick† Martin S. Banks†

Stanford University* University of California, Berkeley†

Abstract

Typical stereo displays provide incorrect focus cues because the light comes from a single surface. We describe a prototype stereo display comprising two independent fixed-viewpoint volumetric displays. Like autostereoscopic volumetric displays, fixed-viewpoint volumetric displays generate near-correct focus cues without tracking eye position, because light comes from sources at the correct focal distances. (In our prototype, from three image planes at different physical distances.) Unlike autostereoscopic volumetric displays, however, fixed-viewpoint volumetric displays retain the qualities of modern projective graphics: view-dependent lighting effects such as occlusion, specularity, and reflection are correctly depicted; modern graphics processor and 2-D display technology can be utilized; and realistic fields of view and depths of field can be implemented. While not a practical solution for general-purpose viewing, our prototype display is a proof of concept and a platform for ongoing vision research. The design, implementation, and verification of this stereo display are described, including a novel technique of filtering along visual lines using 1-D texture mapping.

CR Categories: B.4.2 [Input/Output and Data Communications]: Input/Output Devices—Image Display;

Keywords: graphics hardware, hardware systems, optics, user-interface hardware, virtual reality

1 Introduction

Fred Brooks has observed that "VR barely works" [Brooks 2002]. Excessive system latency, narrow field of view, and limited scene complexity are significant problems that limit the utility of virtual-reality (VR) systems, but well-understood approaches to their improvement exist and are yielding steady progress. There are few remaining limitations that are not on such an improvement track. Chief among those is the lack of proper focus cues.

Eye movements and the focusing of the eyes normally work together. Vergence eye movements change the angular difference between the eyes' visual axes. This angular difference determines the distance to the fixation point, which is the point where the visual axes intersect. Accommodation, the focus response of the eye, determines the focal distance of the eye. Fixation distance and accommodative distance are coupled in natural vision [Howard and Rogers 1995]. The coupling is broken by typical stereo graphics displays, which provide correct binocular disparity specifying a range of fixation distances while forcing accommodation to a single image plane.

The consequences of the forced decoupling of viewer vergence and accommodation include discomfort [Wöpking 1995], induced binocular stress [Mon-Williams et al. 1993; Wann et al. 1995], and difficulty in fusing the two images into a stereo pair [Wann et al. 1995]. Another consequence of this decoupling is error in the perception of scene geometry. Estimation of angles in scene geometry is more accurate, for example, when correct focal distance information is provided than when correct binocular projections of an object are viewed at the wrong accommodative distance [Watt et al. 2003]. Accommodation to a single image plane also eliminates a depth cue—variation in blur—and this too causes error in the perception of scene geometry [Mather and Smith 2000].

1.1 Related Work

Other investigators have sought solutions to the issue of incorrect focus cues. From as early as the Mercury and Gemini space programs, out-the-window displays for vehicle simulation systems have fixed the focal distance at infinity by collimating the light from the display [North and Woodling 1970], or by positioning the display surface sufficiently far from the viewer. Because these systems display only the scene beyond the vehicle itself, infinite focal distance is a good approximation, with errors limited to a fraction of a diopter (D).[1] Unfortunately, infinity optics fails for objects that are closer to the viewer, and therefore cannot work for general-purpose VR systems.

Autostereoscopic volumetric displays, which present scene illumination to multiple viewpoints as a volume of light sources (voxels), naturally provide correct geometric and focus cues. However, these displays do not create true light fields. While the inability of voxels to occlude each other is sometimes given as the reason for this limitation [Perlin et al. 2000], the problem is actually more fundamental: voxels emit light of the same color in all directions, so neither view-dependent lighting effects nor occlusions can be represented simultaneously for multiple viewing positions. Autostereoscopic volumetric displays suffer various other practical difficulties, many of which result from the huge increase in voxels that follows the introduction of a generalized third dimension to the display. Because conventional display technology cannot be leveraged to satisfy this resolution requirement, volumetric displays require custom approaches. These include scanned rotating screens [Favalora et al. 2002], laser stimulation of doped media [Downing et al. 1996], and dynamic lens systems [Suyama et al. 2000b]. In addition, volumetric displays are typically small because they are designed to be viewed from various angles and because of the high expense of larger displays.

A commercial product, the DepthCube [Lig 2003], implements an autostereoscopic volumetric display with limited depth and viewing angle using DLP projection [Sampsell 2004] and a stack of electronically controlled screen surfaces. This display is able to

*email: kurt_akeley@acm.org, s.watt@bangor.ac.uk, ahna@berkeley.edu, marty@john.berkeley.edu

[1] Diopters are the reciprocal of meters. In optical systems dioptric measurements are typically relative to the optical center of a lens.

accept image data directly from unmodified graphics applications. While view-dependent lighting effects can be displayed, they are computed for a single viewpoint, and are therefore incorrect when viewed in stereo or by multiple observers.

Luminance-modulation displays present a single viewer with images at two distinct depths, both projected to a cyclopean eye [Suyama et al. 2000a; Suyama et al. 2001]. Thus view-dependent cues, such as occlusion and lighting, are compromised much as they are in autostereoscopic volumetric displays. Binocular disparities are also incorrect.

Non-volumetric approaches to correcting focus cues include displays that adjust the focal distance of the entire image to match the viewer's accommodation, which must be estimated by tracking gaze direction or vergence angle [Omura et al. 1996], and displays that adjust focal distance on a per-pixel basis [Silverman et al. 2003; McQuaide et al. 2002]. A significant limitation of these approaches is that they cannot present multiple focal distances along a visual line (Section 2.4).

Holographic displays with sufficiently high resolution automatically provide correct focus cues. But achieving the required resolution is not currently possible given the computational and optical requirements [Nwodoh and Benton 2000; Lucente 1997; Lucente and Galyean 1995].

Our work follows that of Rolland et al. [1999], who computed the spatial and depth resolution requirements for a multi-plane, head-mounted display. They suggest several possible design approaches, including volumetric display technology, but no implementation or results are presented.

The most similar prior work was performed at Fakespace Labs [McDowall and Bolas 1994]. The Fakespace Boom was augmented with prism assemblies such that image planes at two focal distances were summed. This work was not published, however, and no studies of its effectiveness were ever performed.

Rolland et al. and McDowall and Bolas did not consider depth filtering, which is related to the luminance-modulation function of Suyama et al., and is a key feature for fixed-viewpoint, volumetric displays. We describe the importance of depth filtering in Section 2.3, and our prototype implementation in Section 3.3.

1.2 Goals

Our long-term goal is to enable augmented reality with practical head-mounted display technology, so that viewers experience direct views merged with generated graphics in day-to-day settings. Achieving this goal requires high image quality, image correctness, and viewer comfort. To determine whether fixed-viewpoint volumetric displays offer a path toward this long-term goal, we implemented a prototype display. The prototype was also designed as a test bed for further vision research.

Section 2 describes the principles of fixed-viewpoint volumetric displays, and identifies optimizations that may be employed by practical implementations. Section 3 describes the hardware and software implementation of the prototype display. Section 4 describes our experience with the prototype, including calibration, verification of design principles, and measured improvement in viewer performance. Use of the prototype for further vision research will be presented in other technical papers.

We believe that the fixed-viewpoint volumetric approach can lead to practical head-mounted display technology, but the prototype is not such a device. It is suitable only for laboratory use.

2 Fixed-viewpoint Volumetric Display

Volumetric displays are incapable of presenting a true light field for multiple simultaneous viewpoints, so they cannot correctly represent view-dependent lighting (such as occlusions, specularities, and reflections) when used as autostereoscopic displays. For a single fixed viewing position, however, a volumetric display can provide a completely correct light field, including correct focus cues. Used in this manner, voxels that are occluded from the viewing position are unlighted, and voxel lighting is chosen to correctly represent view-dependent lighting along visual lines.

Because a fixed-viewpoint volumetric display supports only one viewing position, a stereoscopic implementation requires two independent fixed-viewpoint volumetric displays. The added expense and complexity are more than offset, however, by optimizations that are made possible by fixing the viewing position relative to the display. These optimizations are detailed in the following subsections.

2.1 Non-homogeneous Voxel Distribution

The Cartesian and cylindrical voxel distributions that are typical of autostereoscopic volumetric displays [Downing et al. 1996; Favalora et al. 2002; Lig 2003], are not optimal for a display with a single fixed viewing position. An optimal voxel distribution is instead dictated by the spatial and focus resolutions of the human eye.

The maximum resolvable spatial frequency of a human subject is 60 cycles/deg (cpd). This resolution is achieved only along the visual axis, for signals projected on the fovea [Wandell 1995]. In a typical configuration the display will be rigidly connected to the viewer's head, so view direction will be limited by the range of motion of the eye. This range is approximately 100 deg horizontally and 90 deg vertically [Boeder 1961], but viewers typically limit eye rotations to 20 deg from center, rotating their heads to achieve larger changes [Simon et al. 2004]. An ideal display would therefore satisfy the need for 60-cpd visual resolution over the required 40-deg field of view, and provide a coarser resolution over the remaining 160×135-deg field of view of the eye. While the visual cues that are provided by full field of view are known to be important [Tan et al. 2003], for simplicity we will consider as ideal an image plane with 50-deg field of view and a voxel resolution of 120 per degree, to satisfy the Nyquist limit at 60 cpd. Assuming regular voxel spacing, the voxel dimensions of such an image plane are $6,400 \times 6,400$.

Display resolution in the depth dimension affects only focus cues. The degree of blur due to focus error is roughly proportional to the magnitude of the focus error measured in units of diopters [Levick 1972]. So adjacent image planes should be separated by a constant dioptric distance. Rolland et al. [1999] estimate this distance as $1/7$ D, based on the eye's depth of field with a typical pupil aperture.

The range of human accommodation is at most 8 D [Wandell 1995]. With a depth resolution of $1/7$ D, a display with the full 8-D accommodative range requires 56 image planes. Diopters bunch near the eye, so moving the near-field limit of the display slightly greatly reduces the required accommodative range. The 4-D range from $1/4$ m (10 inches) to infinity requires only 28 image planes, for example, and is treated as ideal in this paper because it satisfies most viewing situations (Figure 1).

The voxel depth of such a 4-D-range display is more than two orders of magnitude lower than the $6,400$-voxel spatial dimensions

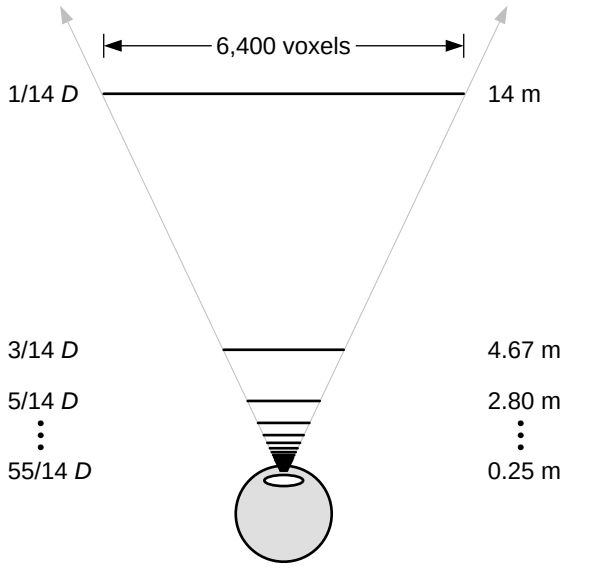

Figure 1: *Twenty-eight image planes at* $1/7$*-D separations form the display that is regarded as ideal in this paper. Each image plane has* $6,400 \times 6,400$ *voxel dimensions, insuring a minimum spatial resolution of* $1/2$ *arcmin.*

of an ideal image plane. Fixing the position of the viewpoint allows a dramatic reduction in voxel count with no reduction in image quality.

2.2 Collapsing the Image Stack

As described thus far the ideal fixed-viewpoint volumetric display has a finite and manageable depth resolution of 28 voxels, but its physical dimensions are not acceptable. Each image plane is centered in its focal range, so the most distant plane is $1/14$ *D* from infinity, or 14 m from the viewpoint. At least two approaches to collapsing the physical depth of the image plane stack, while maintaining the focal distance relationships, have been proposed.

One solution uses dynamic optics to sequentially position a single image plane at multiple focal distances. Wann et al. suggest using an oscillating lens for this purpose [Wann et al. 1995]. Adaptation of the adaptive-optics technology used in large telescopes is also possible. Both approaches have the advantage of requiring only one image plane, and also result naturally in near-optimal spatial distribution of voxels.

The other solution, suggested by Rolland et al. [1999], retains the stack of image planes and requires no moving parts. Instead, the image planes are viewed through a fixed positive lens. When an *n-D* lens is placed close to the eye, the focal distance of an image plane is preserved by moving it *n* diopters closer to the eye. The physical depth of the resulting image plane stack is $1/(n+(1/14))$ m, in the case of the ideal display.

2.3 Depth Filtering

Rendering to a fixed-viewpoint volumetric display requires sampling the scene along visual lines, just as it does for non-volumetric displays. The radiance of an individual sample is then assigned to voxels in one or more image planes, based on the spatial and depth coordinates of the sampled object. Spatial antialiasing within an image plane is possible, and is at least as important as it is for non-volumetric displays. Proper filtering in the depth dimension is critical, however, if visual artifacts are to be avoided.

Not filtering in the depth dimension corresponds to assigning the full sample radiance to voxels in a single image plane—the one that is closest to the sampled object. The left panel of Figure 2 illustrates how such assignment results in a discontinuity within the viewing volume, when a slanted plane is rendered. It is obvious that this discontinuity is visible if there is any misalignment of the voxels in adjacent image planes, relative to the viewpoint. Surprisingly the discontinuity is visible even if the alignment is perfect.

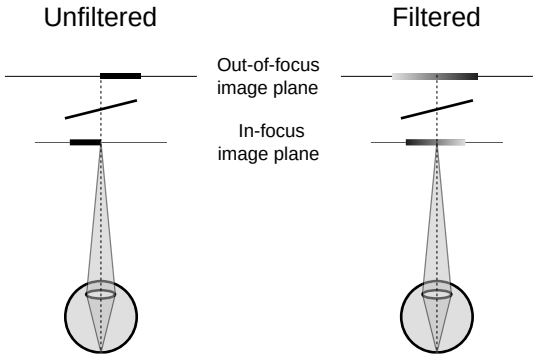

Figure 2: *Voxel lighting with and without depth filtering. The black bars on the image planes represent regions of lighted voxels that result from rendering the object surface that is positioned between the image planes. Not filtering results in a discontinuity of intensity within the voxel volume. Depth filtering replaces this discontinuity with gradual intensity gradients.*

Because the focal distances of adjacent image planes differ, it is not possible for the eye to accommodate to both simultaneously. The retinal images of the voxels on one image plane therefore differ from those of the voxels on the other plane. Figure 3 shows the retinal images that result from each image plane in isolation, and the discontinuity that results when these images are summed to form the actual retinal image. For clarity the figure shows the 32% discontinuity resulting from a $1/2$-*D* difference in focal distances. When it is computed for a $1/7$-*D* difference, however, the magnitude of the discontinuity is still 8% of the magnitude of the signal. This is much larger than the 2% limit of perception [Blackwell 1946], so the discontinuity is visible.

The right panel of Figure 2 illustrates that, when depth filtering is employed, the discontinuity within the volume is replaced with gradual intensity gradients. In this example sample radiance is distributed between the two nearest image planes in linear proportion, with distance measured in diopters. When the depth filtered voxels are convolved and summed, using the method shown in Figure 3, there is no discontinuity, just a gradual change in the magnitude of the voxel images themselves. Diopter-linear depth filtering eliminates the otherwise visible discontinuities due to focus error, as well as discontinuities that would result from incorrect voxel alignment.

Depth filtering adds to the expense of rendering, but it is easily and efficiently implemented on modern rendering systems using 1-D texture mapping (Section 3.3). Accepting this allows significant optimization. Additional MATLAB simulation shows that depth filtering eliminates visible discontinuity regardless of the magnitude of the difference in focus errors. If the separation between image planes could be increased to $1/4$ *D* or even $1/2$ *D*, it becomes possible to construct fixed-viewpoint volumetric displays that span useful depth ranges with fewer than 10 image planes.

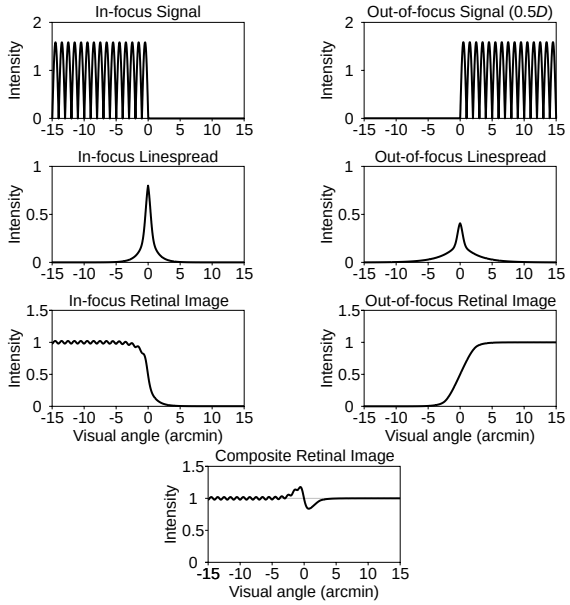

Figure 3: *Modeling the visual discontinuity with no depth filtering. The in-focus and out-of-focus voxels of Figure 2 (Unfiltered) are convolved with the corresponding linespread functions, then summed to give the final retinal image. Voxels subtend* 1 *arcmin with* $1/2$ *sine wave intensity distributions.*

Such displays might make effective use of mainstream planar display technology, allowing them to become commercially viable.

2.4 Multiple Focal Distances

Semi-transparent display is an often-claimed advantage of autostereoscopic volumetric displays. This feature is generally a deficiency, however, because it cannot be disabled while retaining multi-viewpoint viewing. It is an optical rendering technique—important aspects of the image are created by the display itself. But rendering is better implemented and controlled using the software and hardware algorithms that have been evolved over the past few decades. In a fixed-viewpoint volumetric display, the additive nature of light along a visual line allows transparency and reflection to be rendered accurately *and* depicted with near-correct focus cues.

The need for multiple focal distances along a visual line is not as widely appreciated as the need for multiple focal distances in different visual directions. The left panel in Figure 4 illustrates an example of this need. The surface of the cube scatters light (from a matte component of the surface) and reflects light (from a glossy component). Because the cube's surface is flat, the correct focal distance of the reflection of the (diffuse) cylinder is the sum of the distances from the eye to the cube and from the cube to the cylinder.[2] The correct focal distance of the scattered light is the distance from the eye to the cube. The right panel of Figure 4 illustrates how this scene is rendered into a fixed-viewpoint volumetric display. The reflection is drawn deeper into the display, at the focal distance that is the sum of the eye-to-cube and cube-to-cylinder distances.

Subtractive image planes, such as stacked LCD flat panels, cannot directly implement a true volumetric display. Both depth filtering

[2]If the reflecting surface is not planar the reflected light typically spans a range of focal distances.

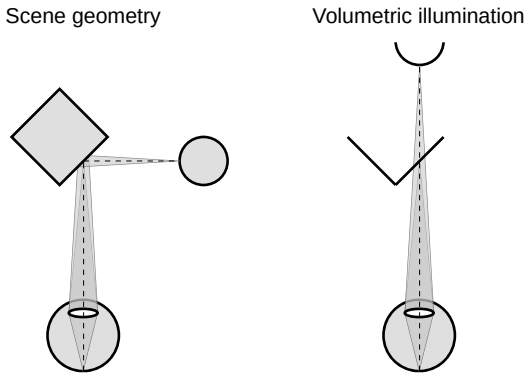

Figure 4: *Multiple focal distances along a visual line. Scene Geometry: the reflection of the cylinder has a longer focal distance than the surface of the cube. Volumetric Illumination: illustrates how the scene is rendered to a volumetric display with high depth resolution.*

and multiple focal distance rendering depend on summing light along visual lines, which is possible only when voxels are additive light sources. Visible discontinuities due to non-depth-filtered rendering are potentially much greater in subtractive displays, because direct illumination from the back-light becomes visible.

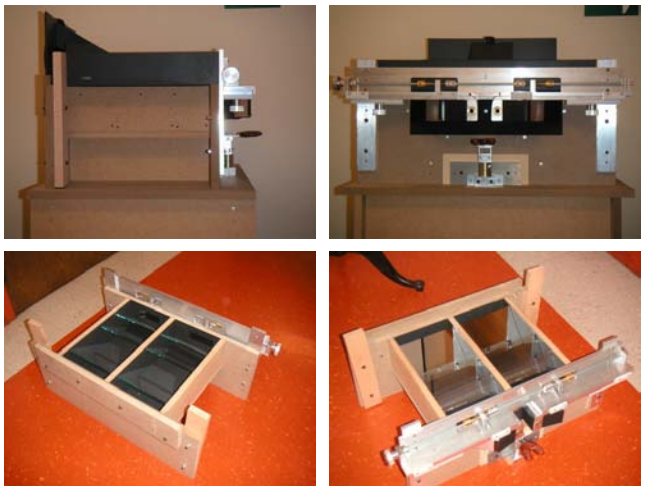

Figure 5: *Four views of the prototype display. The T221 monitor is removed in the bottom two images to expose the beamsplitters and front-surface mirrors.*

3 Prototype

3.1 Design Decisions

The best design is a compromise reached after careful consideration of priorities. Our primary goal was to determine the viability of an optimized fixed-viewpoint volumetric stereo display. We felt that depth resolution was the most important consideration. Our design therefore had to provide several image planes at substantially different focal distances, and these planes had to be additive so that depth filtering could be implemented. Stereo capability with significant overlap of the view frusta was also critical. And the display had to provide good laboratory ergonomics, so that many subjects could be tested with repeatable results.

High spatial resolution was important, but design compromises resulted in lower than ideal resolution (Table 1). Depth of field, field of view, frame rate, and latency were relatively unimportant, as long as they were sufficient to support the required testing. Although motion parallax is an important cue, we felt that the difficulty of implementing a head-mounted display (which would require collapsing the image stack) outweighed the benefit.

Figure 6 schematizes the optics of the prototype display. An LCD flat panel is viewed through plate beamsplitters such that each eye sees three superimposed images. The use of a single flat panel limits the physical dimensions of the prototype display, and thus the available depth range and field of view, but it also has significant advantages.

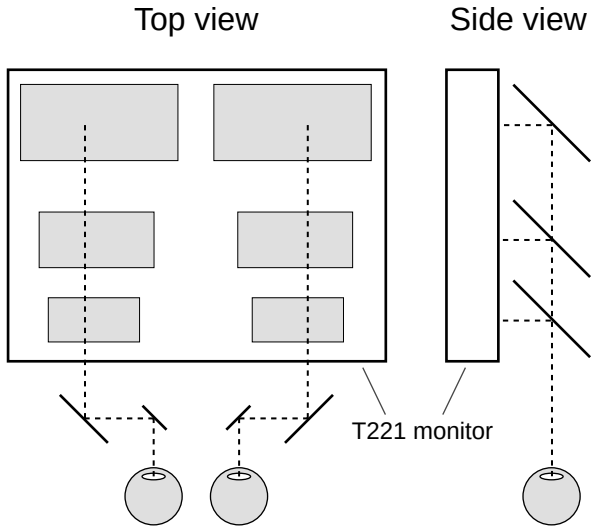

Figure 6: *Each eye views three superimposed viewports. Periscope optics separate the visual axes so that the left and right viewports do not overlap. (The side view is rotated* 90 *deg counterclockwise.)*

The pixel locations are precise because they are fixed by the manufacturer of the LCD flat panel in a regular grid. The digital interface is used to drive the LCD flat panel, so there is no drift in the relationship of the pixels in the frame buffer to those on the panel surface. Thus the locations of the six rendered viewports are known exactly, can be changed precisely and reliably, are coplanar, and cannot be rotated or skewed with respect to each other. The LCD flat panel is driven by a single graphics card, avoiding synchronization and other system complexities associated with multiple graphics cards.

Name	Distance	Diopters	$\triangle D$	Spatial resolution
Near	0.311 m	3.21 D		1.38 arcmin
Mid	0.394 m	2.54 D	0.67 D	1.09 arcmin
Far	0.536 m	1.87 D	0.67 D	0.80 arcmin

Table 1: *Prototype image plane distances.*

No optical elements other than beamsplitters and mirrors are used, so the focal distances of the three image planes are equal to their measured distances (Table 1). The $2/3$-D separations of the image planes are wide compared with the $1/7$-D separations of the ideal display. This spacing tests the effectiveness of depth filtering, and was driven by our concern for adequate vertical field of view, which would have been reduced from the current +/-4.4 deg had the spacing been tightened. The horizontal fields of view are 6.1 deg outside and 12.6 deg inside, resulting in substantial overlap of the view frusta.

The IBM T221 LCD flat panel [Wright 2002] that is used is currently the highest resolution device of its kind on the market. Its horizontal and vertical pixel densities are both 80/cm. The flat panel dimensions are 0.478×0.299 m, with an overall resolution of 3840×2400 pixels. The resulting spatial resolution differs for each image plane (Table 1), but even the lowest resolution is high enough that observers did not comment on the visibility of individual pixels.

The LCD flat panel is driven by a 128MB NVIDIA Quadro 900 XGL graphics card, manufactured by PNY technologies [NVI 2002]. The card has enough on-board memory to support the 9-Mpixel flat panel with 32-bit double-buffered color and a 24-bit Z-buffer, and to store the required texture images. Rendering performance is more than adequate, but because only two DVI display ports are available, the display frame rate is limited to 12 Hz at full 3840×2400 resolution. The low frame rate is acceptable because LCD flat panels do not flicker.

LCD flat panels cannot switch quickly enough to be viewed through shutter glasses. This, along with the low display frame rate, led to the decision to implement the separate stereo views with non-overlapping viewports. The aluminum periscope assembly that separates the left-eye and right-eye visual axes is visible in Figure 5. The arrangement of the viewports and the paths of the visual axes are illustrated in Figure 6, and example images are provided in Figures 16 and 17.

3.2 Ergonomics

At this writing 18 human subjects have run 618 experimental procedures using the prototype display. While procedure lengths vary, a typical run of the user performance experiment that is described in next section requires almost 600 individual trials. With such extensive use anticipated, the prototype display was designed to be robust and to produce repeatable results. Its design features include:

- **Horizontal, eye-level viewing.** A typical experiment lasts $20-60$ minutes. Although short breaks are taken, the subject must be comfortable for consecutive periods of many minutes.
- **Precise, repeatable view position.** The prototype is fitted with a *bite bar* mount. Laboratory subjects are fitted just once with a personal bite bar, which is calibrated such that the subject's nodal points are centered about an origin point on a line whose position is fixed relative to the bite bar mount. Subjects' view positions vary only as a function of the distances between their eye centers (inter-ocular distance, IOD), and are repeatable without mechanical adjustment.
- **Rapid IOD adjustment.** Periscope separation is adjusted to match IOD in the range [50-70] mm using a lead screw with left threads for the left eye and right threads for the right eye. The periscope remains symmetric about the origin, regardless of adjustment. The IOD of the projections used by the software is specified separately, using the software interface that is described in the following subsection.
- **Non-critical periscope adjustment.** Because the two mirrors of each periscope move as a rigid unit, periscope motions affect only the field of view, and have no effect on visual lines, focal distance, or any other view parameter. Thus periscope separation is non-critical. Only subject IOD and software IOD are critical, and both are exactly repeatable.

3.3 Software

The prototype display is driven by a 10,000-line C program, using OpenGL [Segal and Akeley 2002] and its GLUT utility library [Kilgard 1996]. The code is essentially a loop that computes new values for its state vector, renders the six viewports, renders state information as text, and swaps display buffers. Viewports are rendered independently without culling. Near culling would eliminate necessary occlusions, because Z-buffering is used for hidden surface removal. Far culling, while technically correct, causes variations in rendering performance that are undesirable in our research environment.

The default render loop moves the selected object repeatedly between near and far stops. In addition, an experiment mode with a generic up/down psychophysical staircase core [Levitt 1971] is shared by all the experiments that have been implemented. An experiment is configured with 1–20 independent staircases, one of which is randomly selected per trial. Each staircase includes static state, such as the fixation distance and focal distance to the object's origin, as well as a variable that is updated based on the binary responses. The experiment descriptions in Sections 4.2 and 4.3 include specific staircase details.

Depth filtering is implemented using 1-D texture mapping, allowing intensity to be computed separately for each rendered pixel. Three separate 1-D textures, one for each image plane distance, are precomputed and used repeatedly during rendering. The selected 1-D depth texture is indexed using the eye-coordinate Z value, normalized to the [0-1] texture-coordinate range. The coordinate mapping is implemented with the OpenGL TexGen mechanism and the texture matrix, which are both initialized prior to rendering. Because TexGen extracts the vertex coordinates after transformation by the ModelView matrix, modeling transformations do not upset the texture coordinate mapping. Figure 7 illustrates the three diopter-linear depth filter functions that were used for all of the work described in this paper.

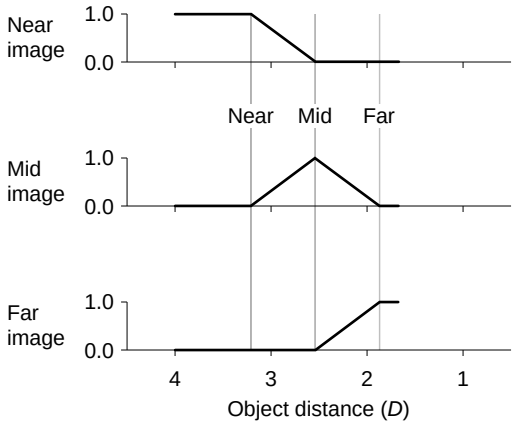

Figure 7: *Depth filter functions for the three image depths. These functions, warped from dioptric to metric distance, define the three 1-D depth texture images.*

4 Experience

Two critical questions needed to be answered to establish the effectiveness of the prototype display:

- **Is device accuracy sufficient?** Can image intensity be held constant as an object is moved nearer or farther? Is geometric alignment accurate enough to prevent visible alignment errors?
- **Is user performance improved?** Do users perform better when fixation distance and focal distance are matched? (Exact matches are possible for fixation distances that match the focal distance of one of the three image planes.) More important, do users continue to perform better when inter-plane fixation distances are used, and image intensity is distributed between the two nearest image planes?

The following subsections detail our approach to answering these questions.

4.1 Intensity Constancy

Two factors affect the constancy of display intensity as an object moves nearer or farther from the viewer:

- **LCD intensity response.** OpenGL color arithmetic assumes a linear relationship with display intensity. Because texture and frame buffer blending were required, gamma correction was done post-frame buffer using the NVIDIA driver and its control of the display hardware. The 2.09 curve gave the most linear results.
- **Beamsplitter light loss.** The reflectance/transmittance ratios of the beamsplitters were chosen to minimize the differences in intensity of the three light paths, but significant differences remained. These were eliminated by attenuating all rendering with one of three viewport-specific factors: 0.365 (near image), 0.615 (mid image), and 1.000 (far image).

To confirm intensity constancy, we measured the intensity of a viewport-filling white rectangle as it traversed the depth range of the prototype display. Luminance measurements were taken with a Minolta CS-100 Chroma-meter, sighted through the left aperture. Multiple measurements were taken at 1.25-cm intervals, and the average values at each distance were found to fall with 2% of the overall average of 2.60 cd/m^2 (Figure 8). The variation is explained in part by loss of resolution due to the viewport-specific attenuation of the 8-bit frame buffer components. Near-image attenuation by 0.365, for example, leaves only 94 distinct frame buffer values.

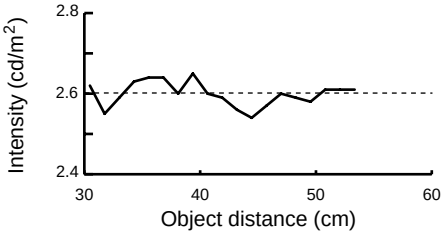

Figure 8: *Intensity constancy. Gamma correction and per-viewport rendering attenuations combine to provide near-constant intensity as an object traverses the depth range of the prototype display.*

An instantaneous 2% change in luminance is the smallest perceptible under optimum conditions [Blackwell 1946]. So continuous variation within 2% of an average value is not a visual distraction. In practice no variation in luminance was visible.

4.2 Geometric Alignment

The software allows subpixel adjustments to viewport position and size, which would otherwise be limited to integer pixel locations by

OpenGL, by making small adjustments to the projection frustum. No provision was made for rotational adjustments. The position and size factors of each viewport were adjusted to meet two goals:

- **Correct binocular disparity.** The far viewports are positioned such that an object rendered directly ahead of an eye is viewed with zero rotation of that eye. A special fixture is used to move the viewing position 1 m back from the usual position, but aligned in all other respects. From this position each eye views only the portion of the far viewport that is directly ahead. A cross-hair target is rendered directly ahead of each eye, and the viewport positions are adjusted until each target is centered in the corresponding 10-mm aperture. This adjustment corrects primarily for slight rotation errors in the periscope mirrors, so it was made only once, for a subject with typical (62 mm) IOD.
- **Exact viewport alignment.** The mid and near viewports are exactly superimposed over the far viewport. This alignment depends critically on the subject's IOD and bite bar calibration, so we developed an automated method to adjust it prior to each experimental session. To align the left-mid viewport to the left-far viewport, for example, the subject views three vernier indicators: two horizontal, one positioned directly ahead and the other 10 deg to the right; and one vertical, positioned directly ahead. Each vernier indicator comprises two collinear line segments, rendered as texture images so that they are subpixel exact. One line is rendered to the mid viewport; the other to the far viewport. The subject adjusts the lines on the mid viewport by rotating them about the view position until they are in exact alignment with the corresponding lines on the far viewport. (The three adjustments are done individually.) After the subject has adjusted them to exact alignment, the software uses the adjustment angles to automatically compute the correct position and size of the left-mid viewport, such that these angles would be zero if the adjustment were done again. If the adjustment angle of the vertical vernier indicator is 3 arcmin (0.05 deg), for example, the viewport is moved 0.344 mm down, the product of 0.394 m (the distance to the mid viewing plane) and $\tan(0.05)$. This process is repeated four times: left-mid to left-far, right-mid to right-far, left-near to left-mid, then right-near to right-mid.

To confirm the geometric accuracy of the prototype display, we implemented an alignment experiment using a two-alternative, forced-choice procedure. The task was to determine the direction of alignment error in a vernier indicator which spanned two image plane depths. (The indicators and adjustment angles are identical to those described above.) Subjects completed trials for 20 separate staircases, each with a different combination of orientation (horizontal or vertical), position, and depth (mid-to-far or near-to-mid) of the vernier indicator, which was initialized with a random alignment error. An incorrect response increased the magnitude of the alignment error, while a correct response adjusted the error toward (and potentially past) zero. Adjustments were made in units of $1/4$ arcmin. The amount of the adjustment began at 16 units, and was halved after each response reversal (that is, a staircase change in one direction followed by a change in the other direction) to a minimum adjustment of one unit. Each of the staircases was run until the eighth reversal occurred. The staircase values at the last four reversals were averaged to give the estimate of the alignment error.

Two subjects, both under 30 years old with normal vision, completed separate left-eye and right-eye versions of the experiment. The results are plotted in Figures 9 and 10. The RMS errors across both subjects are 0.62-arcmin horizontal and 0.93-arcmin vertical, roughly $3/5$ and $4/5$ of the 1.09-arcmin angle subtended by a voxel at the center of the mid image plane. The greater error in the vertical direction may be due to the weight of the subjects' heads resting on the bite bar.

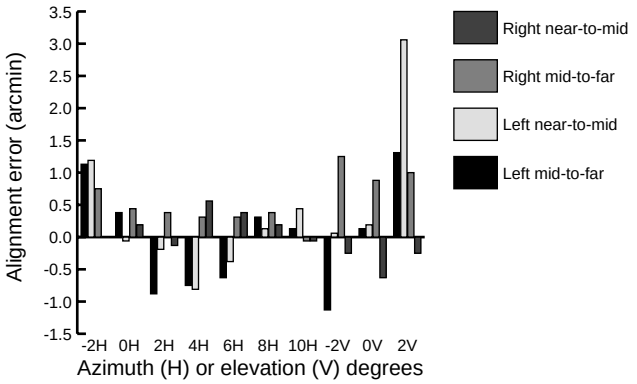

Figure 9: *Alignment errors of subject 1, a co-author. Tests were for horizontal (H) and vertical (V) alignment. Horizontal vernier indicators were positioned on the horizon, at the indicated degrees of azimuth. (Positive angles are inward, negative outward.) Vertical vernier indicators were positioned with zero azimuth, at the indicated degrees of elevation.*

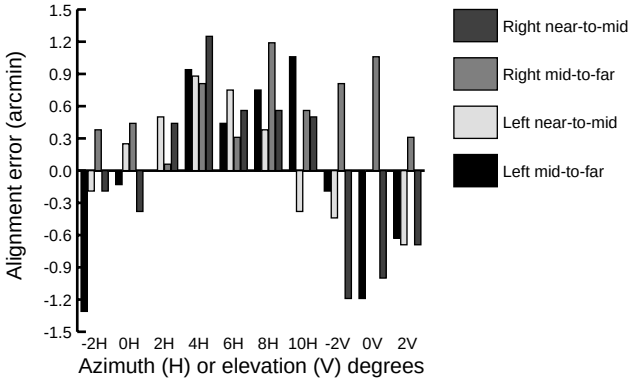

Figure 10: *Alignment errors of subject 2. See Figure 9 for notation details.*

4.3 User Performance

To quantify the effect of the modified focus cues on user performance, we devised an experiment that measured the time required to fuse (perceive the depth of) a stereo scene under various consistent and inconsistent cue conditions. This experiment was designed to be analogous to the typical viewing situation of looking around a scene in which objects are at various distances. We expected that fusion would be faster when fixation and focal distances were nearly matched. Figure 11 provides examples of the scene geometry in this experiment.

Intensive testing of a small number of subjects is standard practice in vision research because it allows reliable measurements under controlled conditions. We tested three subjects, who were all unaware of the purpose of the experiment. The subjects were young (19, 19, and 24 years old) and all had normal vision and stereoacuity at least as good as 40 arcsec as assessed by the TITMUS stereo test. Subjects completed a series of trials, each beginning when the subject pressed a key to respond to the previous trial. After a short

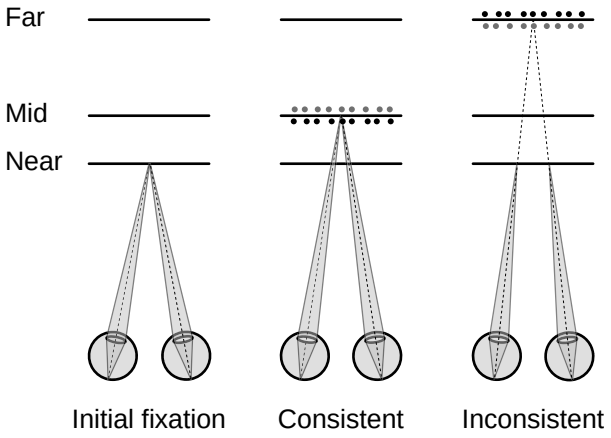

Figure 11: *Each trial begins with the subject viewing a target object on the near image plane (Initial fixation). The subject then views two nearly adjacent frontoparallel planes of randomly positioned dots, one plane red, the other green. The fixation and focal distances of the dot planes are independent, and differ from trial to trial. Consistent: a trial in which these distances are consistent, both at the mid image plane. Inconsistent: a trial with fixation distance at the far image plane, but focal distance remaining at the near image plane. The dots are rendered on the near image plane, with disparities consistent with far fixation distance.*

delay, a target was briefly displayed, bringing subject fixation and accommodation to the center of the near image plane. Then the object to be fused was shown for a specific number of frames. The display was then blanked.

The experiment used a two-alternative, forced-choice procedure. The object to be fused was a pseudo-random pattern of dots rendered on two frontoparallel, closely spaced planes. Dots were red on one plane and green on the other. The subject's task was to indicate whether the plane of red dots was nearer or farther than the plane of green dots. This task is easy once the dots have been fused, and is impossible otherwise.

Care was taken to avoid providing the subjects with any cues other than object disparity and focal distance. Dot positions were randomized for each trial, as were the relative positions of the red and green planes. The planes of dots were clipped to elliptical boundaries, and the sizes of the dot patterns were individually and randomly scaled by up to 5% per trial. The separation between the planes of dots was adjusted as a function of the object's fixation distance from the viewer, such that constant disparity was maintained.

Subjects completed trials for 12 separate staircases, each with a different combination of dot fixation and focal distance. The first trial of each staircase displayed the dots for 40 frames.[3] One incorrect response increased the time allowed to fuse the stimulus, while two consecutive correct responses were required to decrease it. (Without this bias a staircase could become stable at a stimulus duration shorter than that required by the subject.) The amount of the adjustment began at eight frames, and was halved after each response reversal to a minimum adjustment of one frame. Each of the staircases was run until the twelfth reversal occurred. The staircase values at the last four reversals were averaged to give the estimate of the stimulus duration needed to get 71% correct.

[3]This experiment was run at 1920×1200 resolution to increase the frame rate from 12 frames/sec to 41 frames/sec, allowing adequate timing resolution. Because only antialiased dots were rendered, the loss of image resolution was thought to be acceptable.

Each subject completed three repetitions of the entire experiment. The mean of these was taken to give an overall score for each subject in each condition. These values are presented in Figure 12. All three subjects show essentially the same pattern. Given this consistency we chose to collapse across subjects when presenting our results in the following paragraphs, even though the mean subject scores vary across a nearly 4:1 range.

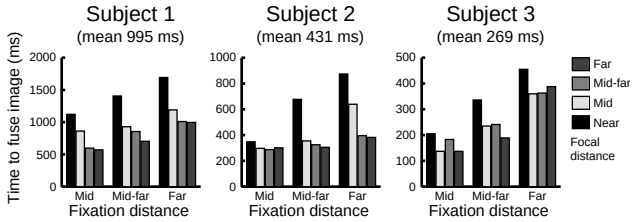

Figure 12: *Experimental results for each of the three subjects. The mean response times vary significantly, but the response patterns are essentially the same.*

A subset of the subject-averaged results is presented in Figure 13. Because subject fixation begins at the near image plane, it is not surprising that more time was required to fuse dots presented at the far fixation distance than at the mid distance. More important for our purposes is the difference in performance between the cues-consistent cases (where focal distance was equal to fixation distance) and the cues-inconsistent cases (where focal distance was at the near image plane). At the mid fixation distance, the cues-inconsistent case required on average 30% more time to fuse than the cues-consistent case. And at the far fixation distance, where the focus inconsistency is doubled to $4/3$ D, the cues-inconsistent case required on average 70% more time to fuse. Cue consistency significantly improves performance, and the penalty of cue inconsistency is related to the magnitude of the inconsistency.

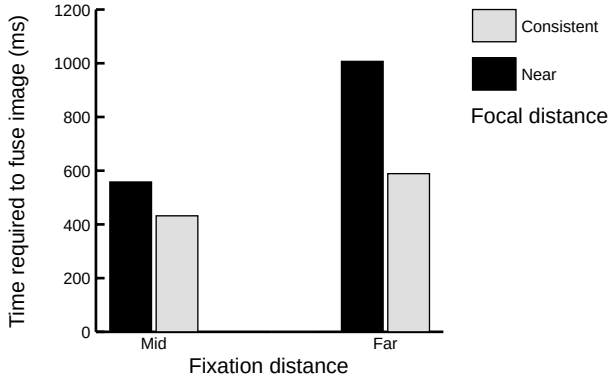

Figure 13: *Average viewing time required for subjects to fuse an object displayed at mid or far fixation distances, with focal distance held to the near distance (black bars) or consistent with the fixation distance (gray bars).*

The experiment also included staircases with fixation distance set to the dioptric midpoint between the mid and far image planes (mid-far). Figure 14 includes subject performance data at this fixation distance for both cues-consistent and cues-inconsistent cases. Performance in the cues-inconsistent case was nearly equal to the average of the performances in the mid and far fixation distance cases, as would be expected. While the cues-consistent case is not actually consistent (because the image energy is divided equally between the mid and far image planes, rather than presented on a mid-far image plane), subject performance in this case also

fell between the performances at the mid and far fixation distances. It is reasonable to expect that the 50/50 intensity distribution is a worst-case focus cue, so this result provides evidence that depth filtering provides a usable, albeit incorrect, focus cue, even with the large image plane separations of the prototype display.

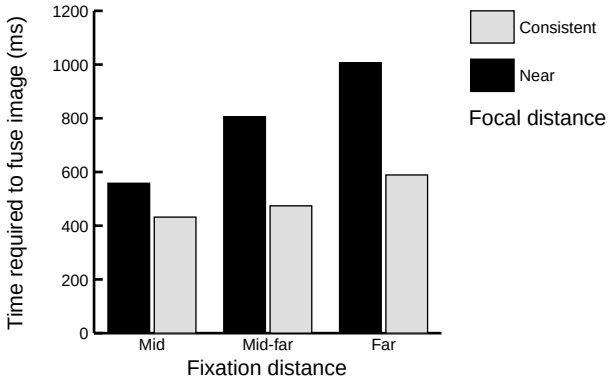

Figure 14: *Average subject performance at the mid-far fixation distance, located at the dioptric midpoint between the mid and far image planes, is added to the data of Figure 13.*

Finally, all of the subject-averaged results are provided in Figure 15. In both cases for which data are available (mid and mid-far fixation distances) subject performance improved slightly over the cues-consistent case when the focal distance was greater than the fixation distance. Because subject fixation and accommodation always began at the near image plane, and moved rapidly toward the far image plane during the fusing period, this result may be related to the subjects' eye dynamics.

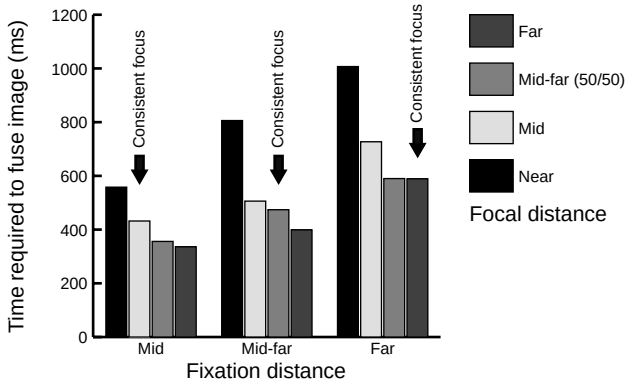

Figure 15: *Average subject performance for all twelve fixation / focal distance combinations.*

5 Discussion and Future Work

We have demonstrated that 3-D geometry can be displayed with high quality using a sparse stack of additive image planes. The fixed-viewpoint volumetric approach leverages current rendering and display technology, allowing focal distance to be closely matched to fixation distance on a pixel-by-pixel basis within a tolerance determined only by the number of image planes. Because there is no need for eye tracking, the approach is inherently robust. Our subjective and experimental results show that the prototype display successfully addresses real deficiencies in current virtual-reality display technology.

Many additional experiments could be performed using the prototype display. To address practical concerns, it would be useful to measure performance of complex tasks, both with and without consistent focus cues. The first author's Ph.D. thesis [Akeley 2004] includes an analysis of multi-image accommodation cues. The authors know of no other research that describes human accommodation in the circumstances of the multi-plane display. Measurement of accommodation while viewing multi-plane images would guide the optimization of image plane separation, and might also lead to improved depth filter functions.

Finally, one interpretation of the data in Figure 15 is that stereo fusion time is minimized when focal distance exceeds fixation distance. To test this possibility, we are developing a fuse performance experiment with initial fixation at the mid image plane, allowing symmetric changes to fixation and focal distance. Preliminary results confirm that matched distances yield better user performance than nearer focal distances, and also outperform farther focal distances.

6 Conclusion

Despite remarkable progress in the past few decades, 3-D graphics is still a specialized activity, rather than a part of our everyday lives. In addition to addressing the poor ergonomics of current stereo graphics displays, we hope our research will lead to developments that allow graphics, in the form of augmented reality, to be integrated into our daily experience. Such integration demands that the ergonomic issues of current display systems, including focus cues, be tailored to human requirements.

7 Acknowledgments

We thank Pat Hanrahan and the anonymous reviewers for their careful reviews of this manuscript, and Pat for his support throughout our work. We also thank Mike Cammarano for creating the high-resolution image pairs used in Figures 16 and 17. This research was supported under NIH grant R01 EY14194-01.

References

AKELEY, K. 2004. *Achieving near-correct focus cues using multiple image planes*. PhD thesis, Stanford University.

BLACKWELL, H. 1946. Contrast thresholds of the human eye. *Journal of the Optical Society of America 36*, 624–643.

BOEDER, P. 1961. Co-operation of the extraocular muscles. *American Journal of Ophthalmology 51*, 397–403.

BROOKS, F., 2002. VR presentation at Hewlett Packard, Palo Alto, Jan.

DOWNING, E., HESSELINK, L., RALSTON, J., AND MACFARLANE, R. 1996. A three-color, solid-state, three-dimensional display. *Science 273*, 1185–1189.

FAVALORA, G. E., NAPOLI, J., HALL, D. M., DORVAL, R. K., GIOVINCO, M. G., RICHMOND, M. J., AND CHUN, W. S. 2002. 100 million-voxel volumetric display. *Proceedings of the SPIE 4712*, 300–312.

HOWARD, I. P., AND ROGERS, B. J. 1995. *Binocular Vision and Stereopsis*. Oxford University Press.

KILGARD, M. J. 1996. *OpenGL Programming for the X Window System*. Addison-Wesley Publishing Company.

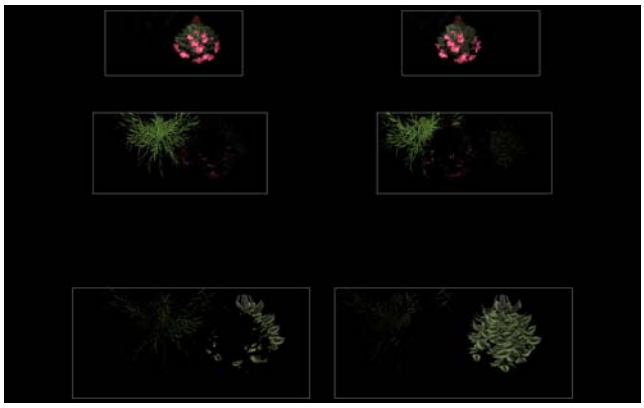

Figure 16: *Full screen image of the T221 monitor. The scene was rendered using RenderMan©, producing separate* $4,500 \times 1,500$ *left-eye and right-eye views with both color and depth information. These image files were read by the prototype software, then remapped and depth filtered to generate the six viewport images. The software is able to read and display short sequences of such ray traced image files, providing a movie loop capability to view highly detailed scenes. The viewports are outlined in white for clarity—these outlines are suppressed in actual use. (Plant models courtesy of Xfrog Public Plants.)*

Figure 17: *The images of Figure 16, superimposed and flipped as they would appear to a subject with infinite depth of focus.*

KOOI, F. L., AND TOET, A. 2003. Additive and subtractive transparent depth displays. *The International Society for Optical Engineering*.

LEVICK, W. R. 1972. Receptive fields of retinal ganglion cells. In *Handbook of Sensory Physiology*, vol. VII/2. Springer Verlag: Berlin, 538–539.

LEVITT, H. 1971. Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America 49*, 467–477.

LIGHTSPACE TECHNOLOGIES. 2003. *DepthCube techology white paper*. Available at www.lightspacetech.com/.

LUCENTE, M., AND GALYEAN, T. A. 1995. Rendering interactive holographic images. In *Proceedings of ACM SIGGRAPH 95*, ACM Press / ACM SIGGRAPH, New York, R. Cook, Ed., Computer Graphics Proceedings, Annual Conference Series, ACM, 387–394.

LUCENTE, M. 1997. Interactive three-dimensional holographic displays: seeing the future in depth. *Computer Graphics* (May).

MATHER, G., AND SMITH, D. R. R. 2000. Depth cue integration: stereopsis and image blur. *Vision Research 40*, 3501–3506.

MCDOWALL, I., AND BOLAS, M. 1994. Fakespace labs accommodation display research. Unpublished report.

MCQUAIDE, S. C., SEIBEL, E. J., B., R., AND III, T. A. F. 2002. Three-dimensional virtual retinal display system using a deformable membrane mirror. *2002 SID International Symposium Digest of Techinical Papers 33*, 1324–1327.

MON-WILLIAMS, M., WANN, J. P., AND RUSHTON, S. 1993. Binocular vision in a virtual world: visual deficits following the wearing of a head-mounted display. *Ophthalmic & Physiological Optics 13* (Oct.), 387–391.

NORTH, W. J., AND WOODLING, C. H. 1970. Apollo crew procedures, simulation, and flight planning. In *Astronautics & Aeronautics*, vol. March. Available at http://history.nasa.gov/SP-287/sp287.htm.

NVIDIA CORPORATION. 2002. *NVIDIA Quadro4 XGL The Standard for Workstation Graphics*. Available at www.nvidia.com/object/LO_20020215_7302.html.

NWODOH, T. A., AND BENTON, S. A. 2000. Chidi holographic video system. In *SPIE Proceedings on Practical Holography*, vol. 3956.

OMURA, K., SHIWA, S., AND KISHINO, F. 1996. 3-D display with accommodative compensation (3DDAC) employing real-time gaze detection. *SID 96 Digest*, 889–892.

PERLIN, K., PAXIA, S., AND KOLLIN, J. S. 2000. An autostereoscopic display. In *Proceedings of ACM SIGGRAPH 2000*, ACM Press / ACM SIGGRAPH, New York, K. Akeley, Ed., Computer Graphics Proceedings, Annual Conference Series, ACM, 319–326.

ROLLAND, J. P., KRUEGER, M. W., AND GOON, A. A. 1999. Dynamic focusing in head-mounted displays. In *SPIE Volume 3639*, 463–470.

SAMPSELL, J. B. 2004. An overview of the performance envelope of digital micromirror device (dmd) based projection display systems. Tech. rep., Texas Instruments. Available at http://www.dlp.com/dlp_technology/.

SEGAL, M., AND AKELEY, K. 2002. *The OpenGL Graphics System: A Specification (Version 1.4)*. OpenGL Architecture Review Board. Editor: Jon Leech.

SILVERMAN, N. L., SCHOWENGERDT, B. T., KELLY, J. P., AND SEIBEL, E. J. 2003. 58.5l: Late-news paper: Engineering a retinal scanning laser display with integrated accommodative depth cues. *SID 03 Digest*, 1538–1541.

SIMON, A., SMITH, R. C., AND PAWLICKI, R. R. 2004. OmniStereo for panoramic virtual environment display systems. In *Proceedings of VR 2004*, IEEE, 67–73.

SUYAMA, S., TAKADA, H., UEHIRA, K., AND SAKAI, S. 2000. A novel direct-vision 3-D display using luminance-modulated two 2-D images displayed at different depths. *SID 00 Digest 54.1*, 1208–1211.

SUYAMA, S., DATE, M., AND TAKADA, H. 2000. Three-dimensional display system with dual-frequency liquid-crystal varifocal lens. *Japanese Journal of Applied Physics 39* (Feb.), 480–484.

SUYAMA, S., TAKADA, H., UEHIRA, K., AND SAKAI, S. 2001. A new method for protruding apparent 3-D images in the DFD (depth-fused 3-D) display. *2001 International Symposium Digest of Technical Papers 32*, 1300–1303.

TAN, D. S., CZERWINSKI, M., AND ROBERTSON, G. 2003. Women go with the (optical) flow. In *CHI 2003*.

WANDELL, B. A. 1995. *Foundations of Vision*. Sinauer Associates, Inc.

WANN, J. P., RUSHTON, S., AND MON-WILLIAMS, M. 1995. Natural problems for stereoscopic depth perception in virtual environments. *Vision Research 35*, 2731–2736.

WATT, S. J., AKELEY, K., AND BANKS, M. S. 2003. Focus cues to display distance affect perceived depth from disparity. *Journal of Vision 3(9)*, 66a.

WÖPKING, M. 1995. Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of the SID 3(3)*, 101–103.

WRIGHT, S. L. 2002. IBM 9.2-megapixel flat-panel display: Technology and infrastructure. *SPIE Proceedings 4712* (Apr.), 24–34.