A TRUE VIRTUAL WINDOW

A Thesis

by

ADRIJAN SILVESTER RADIKOVIC

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2004

Major Subject: Computer Science

A TRUE VIRTUAL WINDOW

A Thesis

by

ADRIJAN SILVESTER RADIKOVIC

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

John Leggett (Chair of Committee)	John Keyser (Member)
Roger Ulrich (Member)	Valerie Taylor (Head of Department)

December 2004

Major Subject: Computer Science

ABSTRACT

A True Virtual Window. (December 2004)

Adrijan Silvester Radikovic, B.S., Texas A&M University

Chair of Advisory Committee: Dr. John Leggett

Previous research from environmental psychology shows that human well-being suffers in windowless environments in many ways and a window view of nature is psychologically and physiologically beneficial to humans. Current window substitutes, still images and video, lack three dimensional properties necessary for a realistic viewing experience – primarily motion parallax. We present a new system using a head-coupled display and image-based rendering to simulate a photorealistic artificial window view of nature with motion parallax. Evaluation data obtained from human subjects suggest that the system prototype is a better window substitute than a static image and has significantly more positive effects on observers' moods. The test subjects judged the system prototype as a good simulation of, and acceptable replacement for, a real window, and accorded it much higher ratings for realism and preference than a static image.

DEDICATION

To my parents who taught me to want more.

To my girlfriend who waited for me.

To my advisor who gave me a vision.

And to the people at the Computer Science Department who gave me a fellowship.

I hope I have fulfilled everyone's expectations.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. John Leggett for this idea that came across in a conversation with him and his help through this research. I would also like to thank my committee members, Dr. Roger Ulrich and Dr. John Keyser.

In addition, I would like to acknowledge some people who helped me during my research in one way or another. I am grateful to:

Vesna and Dragan Vuckovic for treating me like a son.

Dr. Liliana Beltran for the test subjects for the evaluation.

Phillip Mattingly and his staff for technical support.

Dr. Richard Volz for the discussion on signal filtering of tracking data.

Dr. Bart Childs for believing in me even before I knew what I was doing.

Mirko Mandic and Nikola Milosevic for being my pre-test subjects.

Dragana Djordjevic for helping me with this thesis.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

1 INTRODUCTION

"Another criterion for successful window design might be a dynamic one – i.e. the amount of change in the view that takes place for a given change in the viewing position of the observer. As a result of this movement parallax, not only do objects at a different distance within the view change their relative position, but also the window-view relationship changes. This is why two-dimensional artificial windows, even when very carefully contrived, are unrealistic and soon cease to satisfy; they lack 'depth' within the view and the parallax of window aperture-view is also absent." [32]

In his 1967 study on windows, Markus gave a compelling reason for the failure of artificial windows to satisfactorily substitute for real windows. Still images and video lack three dimensional properties necessary for a realistic viewing experience – primarily, and most obviously, motion parallax. If observers move, even an inch, they will see that the scene does not change as it would in a real window view.

Research from environmental psychology shows that human well-being suffers in windowless environments in many ways and a window view of nature is psychologically and physiologically beneficial to humans. The current lack of good window view substitutes is a major problem for people in windowless or strictly urban environments.

Advancements in computing, display technology and computer graphics now allow us to address this decades old problem. The goal of this research was to create an artificial window view with true motion parallax. Requirements, aside from motion

---

This thesis follows the style and format of *CHI Letters*.

parallax, were: photorealistic view, unlimited viewing area, large but thin display, unobtrusive tracking, and low cost standard equipment.

A prototype system was created and evaluated to test the importance of motion parallax to the observer in a simulated window view. The first evaluation was based on psychological ratings of the effects of the artificial window prototype and a static display on subjects' moods. The second evaluation was a subjective comparison of the artificial window and the static display with respect to realism, preference, and enhancement of the indoor environment. The third evaluation was a task-based testing of the artificial window user interface.

2 RELATED WORK

We could find no previous research concerned with the addition of movement parallax to two-dimensional artificial windows. Ongoing research that partially belongs to the artificial window category is on a live video "augmented" window [20]. However, it is 2D video only and does not consider motion parallax. Furthermore, its primary focus does not rest on results from experimental psychology, but on the effects of live video.

This chapter presents previous work in several relevant areas. Presented first is the research on windowless environments from environmental psychology which is the basis of this research. It gives motivation and significance to the research presented in this thesis. This is followed by two areas from computer science and engineering that make this research technically possible. Image-based rendering deals with photorealistic scene reconstruction and head-coupled displays enable simulation of motion parallax. Together, the previous research presented in this chapter gives the motivation and the methods for our research.

2.1 Windowless Environments

2.1.1. Introduction

Many places in our every day lives have no windows. How these windowless spaces affect people has been well-studied. A summary of the most relevant research findings and a quick overview of some of the most relevant studies is given in this section.

2.1.2. Overview

Much research in environmental psychology shows that having windows is judged as very important by persons in indoor spaces ranging, for example, from workplaces, to healthcare buildings, to residences. Windows are strongly desired by those who lack them [37]. Having a window, however, may not be enough. People also desire to sit close to windows [32]. It is not surprising that qualities such as the perceived pleasantness and spaciousness of rooms are strongly and positively affected by windows [37]. Compared to employees with windows, those without windows have lower job satisfaction, job interest, are more negative about working conditions, and report higher levels of work-related stress [19]. These are alarming findings when we consider the great number of offices without windows in workplaces [3,25,64]. The most extreme cases are found in underground buildings. The subterranean work environment can have significant negative effects on human physiology, hormones, sleeping habits, emotional well-being, and health status [29]. This is somewhat surprising since we know that artificial light is close to real light [37,42] and that fresh air and temperature regulation are adequately provided through air conditioning. These ill effects are at least partly linked to the elimination of a view out. Of all window functions, the view out is among the most important and affects the perception of other functions which may be objectively well substituted for, such as lighting quality [37].

The importance of view lies in the connection to the outside world. A large amount of research has shown that a window view of nature is psychologically and physiologically beneficial [27,28,66,67,68]. Hospital patients recovered faster from

surgery and required fewer strong pain medications if they had a bedside window view of trees [66]. Several well-controlled studies have found that recovery from daily stressors such as noise and auto commuting occurs faster and more completely when one is exposed to a view of nature [68]. Looking at nature scenes diminishes negative or stressful emotions such as tension or anger, while levels of positive feelings increase (pleasantness, for example). Nature views elicit beneficial physiological changes, for instance, in blood pressure and heart activity [48]. A window view of only human-built elements has been found to be almost as bad as no view at all [27]. Based on the sizes of current cities and their growth rates, this is a problem which will only grow with time.

There are two explanations for this effect of nature: Ulrich's Psychoevolutionary Theory [68] and Kaplan's Attention Restoration Theory [28]. The main difference between these is that Kaplan relies on a cognitive voluntary appraisal process, while Ulrich relies on involuntary autonomic nervous system response. Whether it is the structure of natural scenes or the residuals of our evolution that make nature so beneficial makes no difference to the fact that it is beneficial.

Windowless employees have been found to use three times more nature-oriented visual materials in their décor than employees with windows [25]. Interestingly, not all nature scenes are equally effective. Park-like or savannah-like scenes with prominent water, foreground spatial openness, and background scattered trees have been found to be consistently pleasing across different cultures [67,65]. Images and video seem to be the best replacements at this time for a window view of nature [67,65]. As a part of a larger study, all possible objects were considered for a window replacement [3] and the

best, from nature-benefit point of view, were pictures, video, and potted plants. However, potted plants are far from the perfect natural scene and, in fact, have been found to have a far weaker effect than a window view of nature [27]. The problem with pictures and video was pointed out decades ago. It is the lack of depth [32] or in other words primarily the lack of motion parallax. If observers move, even an inch, they will see that the scene does not change as it would in a real window view. This is why still images and video, even though they are the best that currently exist, are poor substitutes for a window.

Finally, other reasons exist that motivate the search for a better substitute for a window. One is the fact that windows are not perfect. Disadvantages of windows include glare, undesirable heat gain or loss, and lack of privacy [37]. Artificial windows could mitigate these disadvantages. Also, people have a strong need for individual control [37,71] and artificial windows could allow them to control what they see.

2.1.3. Selected Cases

Finnegan and Solomon (1981) [19] hypothesized that workers in windowless environments would have less positive attitudes toward their jobs than workers with windows. They developed a 33-item questionnaire from previous literature with a five point agree-disagree scale. The questionnaire was divided into six factors: job satisfaction, interest value, time sense, space sense, anxiety and physical working conditions. It was tested on 10 pretest subjects and items were adjusted based on correlation coefficients. The revised 19-item questionnaire was completed by 110 female and 13 male test subjects, 32 of them working in windowless offices. Results showed

that windowless employees were significantly less positive on job satisfaction, interest value of the job, physical working conditions, and total scores ($t = 2.09, 2.21, 4.10, 2.98$; $df = 121$).

Nagy (1998) [37] presented several studies comparing Japanese workers in underground offices with those in above ground offices with windows. In her studies, she used a 7 point Likert scale to test employees working for one company that has both types of offices. In the first study, she examined whether office workers distinguish between different functions of a window and whether underground and above ground employees had different attitudes towards windows. A total of 86 above ground and 22 underground employees were sampled. She found that the only clearly distinguished function was light, while the others were mixed. View and sunshine were grouped with the overall importance of windows. On the importance rating, view was the highest for both underground and above ground employees (6.9, 5.9) and was most closely related to overall importance. However, the means for both groups were significantly lower than overall importance (5.2 vs. 6.4, 6.6 vs. 6.9) which suggests that the overall function of a window was considered to be more than the sum of its physical functions. In the second study, Nagy investigated how the two groups perceived lighting conditions. The survey was completed by 77 above ground and 18 underground employees. Underground employees were less satisfied in all lighting aspects than above ground employees. However, above ground employees had the same mean as their overall satisfaction (4.5), while underground employees had a significantly lower overall satisfaction than the mean of all of the components (2.5 vs. 3.6). Since general level and quality of lighting

was the same for both groups, it can be concluded that the absence of windows affected the underground employees and made them believe that their lighting conditions were unsatisfactory. In her third study, Nagy examined the general perception of the office interior which was objectively similar between underground and above ground offices. In this study, 74 above ground and 17 underground employees completed the survey. The two biggest differences in perception were that underground employees found their offices less pleasant (3.7 vs. 4.9) and more enclosed (5.1 vs. 3.9) than above ground employees. In conclusion, she found that in all three cases windows or their absence caused a strong psychological reaction. She found that underground office environments are "inappropriate work places, since they deprive the occupants of their basic need to have windows and visual access to the outside".

Küller and Wetterberg (1995) [29] hypothesized that due to the chronobiological impact of underground spaces, levels of hormones (melatonin and cortisol) would be affected. They studied personnel in two military subterranean installations and two regiments above ground for one year. The two environments were similar in every detail except for windows. A total of 70 male subjects were included in the final results with half from each condition. Hormones were measured from urine. They found that the level of morning cortisol had a much smaller annual variation, the level of afternoon cortisol was much lower, and diurnal amplitude of melatonin was much larger for the personnel underground. Underground personnel also slept half an hour longer every night and had a distinctly different annual illness incidence pattern. Their conclusion that "although underground environment affects human physiology, it is not worse than the

above ground environment" is a little stretched. It is based on the fact that the subjects seemed satisfied with their work environment, but they were all military personnel who are highly disciplined and might think of underground work differently than would an average person.

Ulrich (1984) [66] hypothesized that recovering patients who had access to a natural view would recover better than those without it. He analyzed records for gall bladder surgery recovering patients in a hospital between 1972 and 1981. The rooms were identical and the same nurses cared for the patients. The only difference was that one set of the rooms had a view of deciduous trees and the other a view of a brown brick wall. Patients were matched between the two conditions, i.e. on sex, age, smoking, weight. Only the time of the year when trees are green was taken into consideration. Data consisted of 46 patients grouped into 23 pairs (15 female and 8 male). Analysis showed that patients with the tree view spent less time in the hospital (7.96 vs. 8.7 days, $t(17)=35$, $p=0.025$), nurses made fewer negative notes for them (1.13 vs. 3.96 notes, $t(21)=15$, $p<0.001$), they took fewer number of doses (Hotteling $t^2=13.52$, $F=4.3$, $p<0.01$), and had slightly fewer post surgical complications.

Ulrich et al. (1991) [68] hypothesized that in accordance with Psychoevolutionary Theory, unthreatening natural environments should foster better stress recovery than urban settings. They expected that following a stressor, unthreatening natural scenes would lead to a more positive emotional state and a decline in physiological arousal, accompanied by high level of attention. Subjects, 120 undergraduate students (60 males and 60 females), were shown a 10 min stressor tape (from previous studies) followed by

a recovery environment tape. Each recovery environment (natural vegetation or water, urban heavy or light traffic, and an urban crowded or non-crowded mall) was shown to 20 random subjects. All groups responded equally negatively to the stressor tape with increase in skin conductance (SCR), muscle tension (EMG) and shorter pulse transit time (PTT) (all $p<0.001$). The same was noted on the self evaluation questionnaires. For the recovery conditions, all three physiological measures (SCR, EMG, and PTT) showed significantly faster and more complete recovery for natural settings over urban within 5 to 7 min ($p<0.05$). Self ratings showed lower fear and anger and higher positive affects for natural environments ($p<0.01$).

Kaplan (1993) [27] reported results of two studies. In the first study, 168 employees of a large corporation and two public agencies were anonymously surveyed. Employees with desk jobs with a window view of nature reported fewer ailments and higher job satisfaction ($p<0.05$). In the second study, 615 employees (92% female) were also anonymously surveyed. In addition to questions on health, life satisfaction, job environment, job satisfaction, and so on, they were asked to rate their view with respect to what elements could be seen. Results showed that built elements did not contribute to the satisfaction or restorative value of the view. Nature was found to strongly affect satisfaction and restorative ratings. The more natural elements that could be seen in the view, the higher the ratings were. Rating of view satisfaction was 2.22 for no nature in the view, 2.91 for one natural element, 3.40 for two natural elements, and 3.58 for three natural elements ($F(3,525)=29.07$, $p<0.001$). Overall, subjects with a view of nature felt less frustrated, more patient and enthusiastic, found their job more challenging, and were

more satisfied with life and overall health. In addition, results showed that satisfaction with indoor plants had a much weaker effect than the view satisfaction.

Heerwagen and Orians (1986) [25] hypothesized that windowless offices should have more visual materials on the walls, which should consist of more surrogate views (landscapes and cityscapes) and be dominated by nature content. A total of 37 windowed and 38 windowless offices were sampled at the University of Washington. The number and content of wall décor was recorded for each office. A median number of items per office was 2.44. The windowless sample had 28 offices with more than the median number and 13 with less, while the windowed sample had 13 with more and 24 with less (x^2=9.86, 1df, p=0.005). To see if the windows reduced the amount of wall space, assuming 3 available walls for windowed offices and 4 for windowless, it was calculated that there were 0.88 items per wall for the whole sample. For the windowless sample, there were 24 offices with more and 14 with less, while the windowed sample had 13 more and 24 less than the median (x^2=4.82, 1df, p=0.025). Windowless offices did use more décor than offices with windows, regardless of the space taken by the window. There were six times more landscapes than cityscapes in windowless offices, as opposed to only two times more in offices with windows. There were also four times as many landscapes in windowless offices, although the number of cityscapes was the same (x^2=5.54, 1df, p=0.01). Finally, windowless offices had three times more nature-oriented visual materials than in windowed offices. There were twice as many nature-oriented than non-nature items in windowless offices, while it was almost equal for windowed offices.

Biner, Butler, Lovegrove, and Burns (1993) [3] tried to extend Heerwagen and Orians [25] by adding other substitutes for a window. In the first part they found and evaluated all possible window substitutes. In the second part they failed to find any significant difference between windowless and windowed offices, opposite of Heerwagen and Orians [25]. A possible reason for failure is pointed out in [18] and it is the fact that Biner et al did not distinguish between natural and built window views. Since they sampled offices in the urban area, it is likely that many windows had urban views, which have been found to be as bad as no view at all [27] (as discussed above). In the first part, which makes this research interesting, they made an extensive list of all possible items that could be a substitute for a window from office surveys and architectural literature. They gave the list of 37 items to 57 undergraduate students to rate and classify. These items were confirmed by 47 office workers who gave the same classification. The final classifications with significant substitute ratings was: other apertures (clear skylight, inside window to window, clerestory, translucent skylight, stained glass window, inside window to no window, and door), paintings and art (of nature, of people, of artifacts, abstract paintings, and sculptures), living things (plants and trees, terrarium, and aquarium), and panels (light panel and video panel).

Previous research presented in this section is by no means a complete list. Only the most relevant and most recent studies found have been presented. For those interested, further reading can be found in the references from the research listed here, however a quick start and an excellent overview of this research area is Farley and Veitch [18].

Ulrich also gives a very compelling overview of benefits of nature in his chapter in *The Biophilia Hypothesis* [67].

2.2 Image-Based Rendering

2.2.1. Introduction

Image-based rendering is a relatively new technique for generating real-time photorealistic images. It requires a constant amount of computation regardless of scene complexity. In image-based rendering different views of an environment are rendered from recorded images of a real world scene. This is the best approach for natural scenes due to the complexity of natural world geometry, which can only be approximated with classical 3D rendering techniques. In addition, image-based rendering does not involve modeling nature, only recording it.

This section starts with an overview of image-based rendering and ends with a more detailed description of the method of choice for our research – the spherical environment map.

2.2.2. Overview

The simplest image-based rendering technique is an environment map. An environment map is all the light that arrives at one point at one time. Everything an observer can see by rotating without translating makes up an environment map for that viewpoint. An environment map can be projected for viewing purposes onto different shapes, most common being a cube, a cylinder and a sphere [7]. The earliest shape and most simple is the cube, but it has appearance problems with edges. A cylindrical map is easy to store, but it only supports limited vertical panning. The best shape for

reprojection is a sphere. It uniformly distributes the map, but it is harder to render (needs to be approximated) and to store (projection of a sphere onto a plane).

Chen extended the idea of a single environment map with QuickTime VR [7] by recording multiple environment maps at different locations and allowing the user to move between them. The problem with using just a discrete set of simple environment maps is that the scene has a fixed number and position of viewpoints. A logical step forward is to interpolate between the captured viewpoints and to reconstruct what the scene would look like from an intermediary viewpoint. To accomplish this, the motion of the pixels between the viewpoints, called the optical flow, must be reconstructed. The first to do that were Chen and Williams [8] with their view interpolation technique. In addition to the light captured in the environment map, the depth of each pixel needs to be captured as well. Chen and Williams used artificial renderings in their examples with known depths, but they suggest using a ranging camera in real world scenes. With depth information, they can resolve what should be seen from intermediate viewpoints. McMillan and Bishop [33] have a similar way of generating intermediate views, but they determine depth from closely spaced images by finding the corresponding points between adjacent viewpoints. However, view interpolation is not perfect because it can easily happen that some elements are not visible from two nearby views. This cannot be avoided without taking more images of the scene.

In addition to their method, McMillan and Bishop related image-based rending to the plenoptic function defined by Adelson and Bergen [1] for the purpose of computer vision. The plenoptic function is a complete systematic description of all of the light

visible in a desired subset of space, time and wavelength. It is parameterized by the viewing position (Vx,Vy,Vz), the viewing direction (θ,Φ), the wavelength (λ), and the time (t). Put together, the complete plenoptic function is denoted by P(θ,Φ,λ,t,Vx,Vy,Vz). A complete sampling of the plenoptic function at a certain viewpoint and a certain time gives a full spherical environment map [33] for the wavelengths that can be digitally recorded and reproduced.

From this framework comes a more complete approach to image-based rendering. At roughly the same time, two systems were presented. The first is the lumigraph by Gortler et al [22] and the second is light field rendering by Levoy and Hanrahan [30]. Both methods start by reducing the 5D plenoptic function at a certain time to a 4D function, where color is the stored value of the function and time is a constant. This is valid if there is no direct occlusion between the viewpoints. Such a space is called free space and exists inside a large environment or around a small object. In the free space, a ray from the outside passing through a viewpoint stays constant as it passes through all of the other viewpoints on that line. In this case, the 5D representation is redundant and can be reduced without loss. Such a sampling allows an accurate reconstruction of all of the possible views in the recorded range. The main problem with this approach is the size of the recorded scene. For example, a scene reconstructed from 8192 images with a resolution of 256 by 256 pixels takes 1.6GB of space [30]. They claim that with different compression methods it could be compressed up to 100 times, but it is still unmanageable for normal viewing resolutions.

Finally, starting from a simple environment map one could go in a different direction. Instead of moving in space, the observer could move in time. Such a technique is called panoramic video or panoramic movie. A different environment map exists for every discrete moment in time, like storing environment maps as movie frames. This is recorded either with multiple cameras or with special lens and mirrors. An example is the immersive panoramic video by Pintaric et al [50]. They used 5 video cameras to record a scene in a cylindrical environment map movie.

2.2.3. Spherical Projection

The environment map projection used in this research is the spherical environment map. As mentioned above, it is a representation of the environment that is projected onto a sphere centered at the viewpoint. A quick summary of the math behind the spherical environment map taken from [53] follows. Schröcker [53] depicts the geometry for a cylindrical projection, but it is given as two separate analyses in two planes. Looking at a cylinder from its center, it consists of a circle in one plane and a line in the other, while a sphere consists of circles in both planes. The goal is to show that the data recorded in the scene, mapped onto a sphere and projected on the screen will look exactly the same as a real photograph taken in that scene with the corresponding orientation and field of view. In the scene recording process (Figure 1 extracted from [53]), the world coordinates (XYZ) are mapped onto the sphere. Looking at a cut of the sphere in the YZ plane, the rotation angle θ on the circle is determined from the location of the point with Equation 1. The sphere radius r is equal to the focal length f.

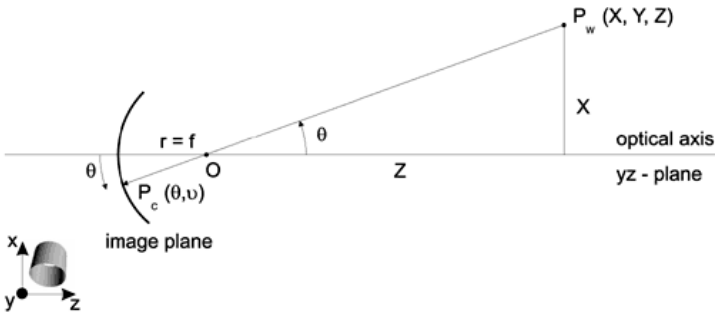

Figure 1: Recording geometry

$$\theta = \arctan\left(\frac{X}{Z}\right) \tag{1}$$

In the reproduction system (Figure 2 extracted from [53]), the sphere needs to be projected onto the image plane which is displayed on the screen. The point P_c on the circle is projected into the point P_p on the projective plane with Equation 2.

$$\begin{aligned} x_p &= f \cdot \tan(\theta) \\ z_p &= f \end{aligned} \tag{2}$$

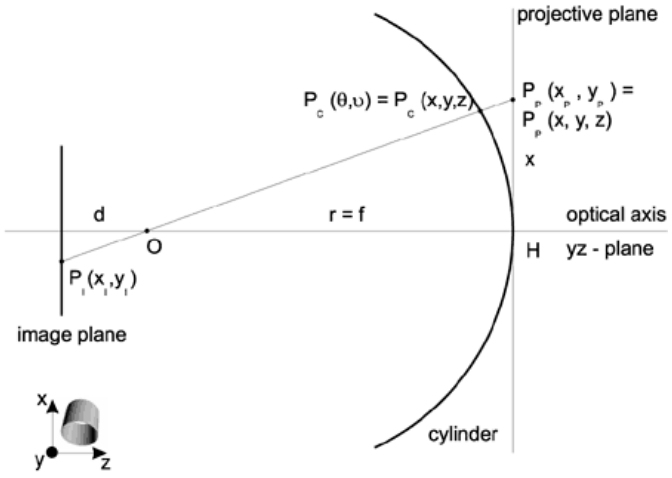

Figure 2: Reproduction geometry

Looking at the recorded point on the circle, its Cartesian coordinates with the origin at the center of the sphere are given by Equation 3. This point is projected onto the image plane with Equation 4. A point on the plane x_p which would be represented with the recorded point is projected with Equation 5. As expected, they are the same, and this proves that the recorded image would be equal to the real scene projected on the screen.

$$\begin{aligned} x' &= f \cdot \sin(\theta) \\ z' &= f \cdot \cos(\theta) \end{aligned} \tag{3}$$

$$x_I' = d \cdot \frac{x'}{z'} = d \cdot \frac{\sin(\theta)}{\cos(\theta)} = d \cdot \tan(\theta) \tag{4}$$

$$x_I'' = d \cdot \frac{x_p}{z_p} = d \cdot \tan(\theta) \tag{5}$$

The scene recording can be done in several different ways. They include fisheye lens, parabolic mirror, panoramic camera, image stitching, or some combination of these techniques [61]. However, the panoramic camera can only be used for cylindrical environment maps as it is a recorded line that rotates. The parabolic mirror requires a significant amount of post-processing as the camera and the operator are visible in the image. This leaves us with the fisheye lens and image stitching. The fisheye lens allows the whole scene to be captured in two pictures, as each captures a little more than 180 degrees in both directions. However, it has some sampling problems because the center of the fisheye image has significantly more data than its edges and the resolution of the scene is limited to the resolution of the camera [49]. Image stitching is the process of combining multiple pictures into a larger one with no visible edges. It consists of three operations: image alignment, image combining, and image blending [49]. Image

alignment aligns the same points in different images with translation and rotation, image combining deals with overlapping regions, and image blending equalizes intensities of the images. Image stitching allows an arbitrarily large resolution (limited only by processing hardware) through selection of a field of view for source images. The smaller the camera field of view, the more images need to be stitched together and the larger the final environment map resolution.

2.3 Head-Coupled Displays

2.3.1. Introduction

Current displays are two-dimensional and are meant to be viewed by an observer seated directly in front of them. When the observer moves in front of the 2D display, the image displayed on the screen remains the same. However, it is not seen as the same, because the surface of the display is no longer seen straight on but in perspective. The closer edge of the screen is perceived taller than the far edge, and the screen width is perceived as narrower. In order to allow the user to move off the central viewing position and still see the correct image, the displayed image needs to be updated to match the user's viewing position. This is exactly what head-coupled displays attempt to provide.

This section starts with an overview of 3D displays, discusses a key component of head-coupled displays – the off-axis projection, and ends with an overview of optical tracking. It describes what we are trying to simulate and provides a background on how it can be simulated.

2.3.2. Overview

An ideal 3D display would emit directional light in such a way that the viewer sees different images from different viewpoints. Referring back to the previous section, it tries to perform a complete reconstruction of the plenoptic function at the surface of the display. Figure 3 (from [39]) illustrates such a 3D display showing how each pixel on the display surface emits a set of directional rays of light called a pencil. Such displays are most often called holographic or autostereoscopic, and there have been many attempted implementations. According to a survey by Halle [24], they include parallax barrier displays, lenticular and integral photography displays, holographic stereograms, and others. However, at this time, none of them have acceptable cost, computational requirements, resolution, or viewing angle. In fact, they will not be available for at least another 20 years due to the computational requirements alone.

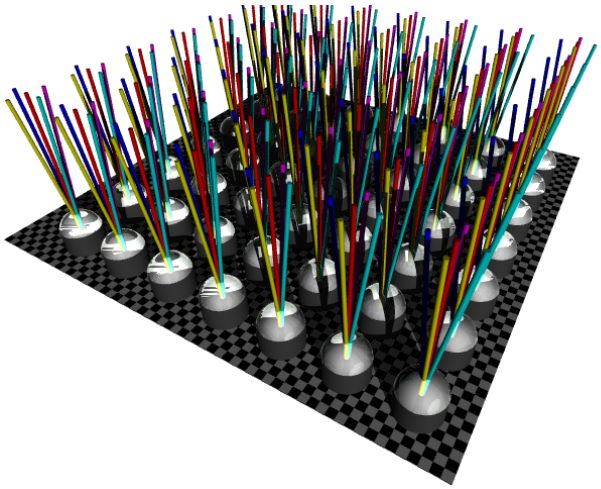

Figure 3: True 3D display

Since such displays are not yet possible, we can only try to simulate them. A simple way to simulate a 3D display is to restrict the number of viewers to a single viewer and render the scene based on that one viewer's position. Such displays are called head-coupled or head-tracked displays [2] and they primarily provide motion parallax. If stereopsis is added (different images for left and right eye), it is called Fish Tank Virtual Reality [70], and the first implementation was described by Deering [12]. He used an ultrasonic head tracking device, a CRT monitor and stereo LCD shutter glasses. Deering described many problems and requirements that still hold. Maximum acceptable lag between the user's position and the display update was found to be 50-100 ms. Viewing parameters had to be adjusted for each viewer and filtering had to be used on the head tracking data. A more thorough examination of Fish Tank VR was done by Ware et al [70]. In their experiments, they found that such displays greatly increase depth determination. This is due to stereopsis and motion parallax, two primary means of obtaining depth in the real world that are not possible with a standard 2D display. Stereopsis is the difference between the pictures that the left and right eye see. Motion parallax is the change in relative position of objects as a result of movement. Ware et al found that the motion parallax was a better depth cue than stereopsis in their system. However, research from psychology shows that in the real world, it is the opposite [17]. In fact, head-coupled systems had problems with correctly displaying stereopsis. This has been mostly resolved with newer implementations, but the viewing angle remains very limited.

Implementations exist which have eliminated the need for glasses and head tracking. They are called autostereoscopic displays. However, they are all limited to a small number of discrete viewing positions and distances. Examples are the Cambridge display [16] with 8 views and the commercially available SynthaGram [59] with 9 views. If stereopsis is not provided in favor of full motion parallax, the user can also be freed of all head-mounted hardware through the use of optical tracking [46]. A camera is mounted on top of the display through which the user's position is determined. This has become possible only recently because of the computational requirements of complex image processing needed to determine the head position. Motion parallax displays are most interesting for this thesis because of their simplicity and the vection effect. Vection is the feeling of self movement induced by movement in our visual field, e.g. looking out a window of an airplane and feeling that we have moved when the nearby plane moves [70]. In vection, images that are perceived the furthest away have the strongest effect.

2.3.3. Off-axis Projection

Off-axis projection is a rendering projection that takes into account the viewer's position and renders the correct view for that position. It is the basis of head-coupled displays. Standard projection that is used with normal 2D displays is called on-axis projection because the center of the image is on the camera viewing axis [15]. In on-axis projection the image projection plane is always considered parallel to the display surface. The perspective of the virtual world is aligned with the image projection plane, but when the observer is off the central viewing position, the virtual perspective image

on the display is already seen in perspective in the real world. The compounding of perspectives results in a distorted image seen by the observer [15].

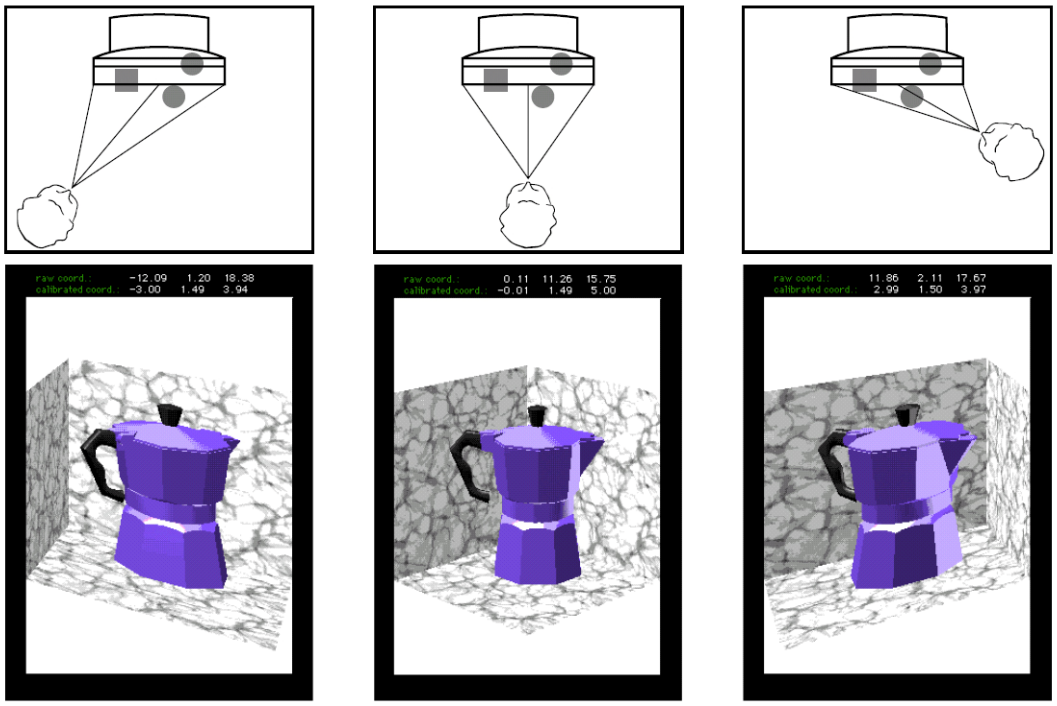

Figure 4: Off-axis projection

In off-axis projection the virtual viewpoint in the virtual world is aligned with the actual physical viewpoint of the user in the real world [12]. The line of sight of the virtual camera is kept perpendicular to the display by translating the camera without rotating it [15]. Now an image meant to be seen from the side looks distorted when seen from the front. This is illustrated in Figure 4 (from [15]). For the left and right pictures, if you move the paper off to the side as illustrated in the diagram above the picture, you will see the correct image of the coffee pot. If you look at it straight on, it will be

distorted. The off-axis projection matrix needs to be adjusted for each viewer and change with the viewpoint in order to obtain accurate images [12]. It has been shown that it is easy to construct such a matrix with basic OpenGL commands [60].

2.3.4. Optical Tracking

In all head-coupled displays, the position of the viewer's eyes is a key parameter to the off-axis projection matrix. Eye position is obtained with a head tracking device, which can be ultrasonic [12], mechanical [70], magnetic [23], or optical [46]. Optical tracking is least intrusive, but also least reliable and slowest (limited by the video frame rate). However, the advantage of having no head-mounted devices makes it the most appealing. With today's computational power, even regular desktop computers can process camera video at a full 30 frames per second.

There are many ways to perform optical head tracking and it is used in many applications such as human-computer interaction, teleconferencing, entertainment, security, etc [21]. The most basic optical tracking is performed by a single static camera and can be roughly divided into global and local methods [21]. Global methods use properties such as skin color, head geometry, and motion. They are more robust, but less precise than the local methods, which use some kind of information on facial features. However, there is no clear cut between actual implementations since most use some combination of global and local properties. The main problem with complex feature trackers is the high error rate (i.e. 3-5% [44]). For head-coupled displays the most important head tracking requirements are robustness, precision, speed, and affordability [21], which eliminate many approaches. Almost all methods developed for head-coupled

displays first find the head position, from which eye position can be estimated [46] or exactly located by finding the actual eye features in the head region [9]. A more recent and accurate stereo tracking approach has been done with two static cameras. Examples include feature based methods using epipolar lines [21], background subtraction followed by neural networks [52], and a method using infrared cameras with feature templates [38].

The problem with most of these approaches is that the user is assumed to be sitting in front of the display [38]. In the best case the user is limited to the field of view of the camera. A newer, less researched approach is the use of an active or pan-tilt-zoom (PTZ) camera [63]. One PTZ camera can be used to cover a large space without loss of detail [62]. A PTZ camera rotates and zooms in to follow the user with a high level of detail. An example method used for surveillance uses motion detection followed by color histogram matching to track the target selected by the operator [63]. Image alignment between two consecutive images is used to find the changes in camera pan, tilt and zoom. In this way, alignment in the real world comes directly from images, requiring no information on camera position and calibration. Simpler methods use a camera interface to obtain the information on camera position [10]. One thing to note is that with PTZ cameras, the background keeps changing and the tracking area can be large which often causes changes in lighting conditions. Methods that cannot tolerate these problems cannot be used. Most obviously background subtraction fails immediately.

Finally, due to the limitations of our research, the method of choice had to be available in some way as downloadable source code or a module. The best match was

found in the Open Computer Vision (OpenCV) library [45], an open-source project initiated by Intel corp. It contains a high-level feature-based tracker and a low-level color-based tracker. The high-level tracker uses a cascade of feature-based classifiers introduced by Viola and Jones [69] and improved by Lienhart et al [31]. The low-level tracker uses an algorithm called CAMSHIFT (Continuously Adaptive Mean Shift) by Bradski [6]. It uses the mean shift algorithm with color histograms to represent color probability distribution and adapts dynamically to changes in the color distributions over time. The X and Y position of the head are estimated from the center of the area matching the facial color and are quite reliable, but the Z position needs to be determined from the color area, which is very noisy and unreliable [6]. This algorithm is extremely robust and handles irregular object motion, image noise, distractors, occlusions and lighting variations.

As pointed out by Deering, the viewer's viewpoint in the real world relative to the head-coupled display has to be exactly matched by the virtual viewpoint in the off-axis projection [12]. This means that the head tracking device needs to be calibrated and aligned with the display. However, when optical tracking is used, the camera image needs to be calibrated with the real world as well – a mapping from the position in the camera image to the real world is needed. Finally, when a PTZ camera is used, the position, orientation, and mapping from zoom settings to focal lengths are also needed [62]. This can be done approximately with manual measurements and calculations, but more interestingly it can also be done with self-calibration. Since the PTZ camera rotates, it can be geometrically described as a rotation around the x-axis for tilt and a

rotation around the y-axis for pan. This is a simple model, which was argued not to be exact [11], but is still good enough. Self calibration involves developing representative equations, which give a cost function that is solved using least squares [62]. The more unknowns involved, the more complicated the formulas. Focal length for a zoom setting can be found by finding the displacement of the principal point between two camera images with different tilt settings. After obtaining several measurements the polynomial mapping between the focal length and zoom settings can be found. Trajkovic found that the polynomial should be of second order [62]. The model he used is given in the Equation 6, where P and P' are two points in a pair of images with the same world coordinates. The focal length can be found using the formula from Equation 7, where d is the distance between P and P'.

$$\begin{aligned} X' &= X \\ Y' &= Y \cdot \cos(\alpha) - Z \cdot \sin(\alpha) \\ Z' &= Y \cdot \sin(\alpha) + Z \cdot \cos(\alpha) \end{aligned} \tag{6}$$

$$f = -\frac{d}{\tan(\alpha)} \tag{7}$$

3 SYSTEM DESIGN

The system design provides details on actual system prototype implementation. The first section provides theoretical analysis of the view rendering and optical tracking. The second section describes the scene capture procedure and hardware/software implementation of the artificial window prototype. The last section provides a description of system operation and instructions on system use for the system operator.

3.1 Theory

3.1.1. Scene Rendering

The goal of our research was to create a view on the display that is as close as possible to the real view that would be seen if there was a window with the exact same dimensions and elevation as the display at the place where the scene was recorded. The scene to be rendered, as discussed in related work, has foreground spatial openness, prominent water, background scattered tress and other savannah-like characteristics. This subsection discusses the theory behind scene rendering.

In order to create the most realistic and efficient window view simulation, we analyzed the properties of a real window view of the target scene. We defined our viewing area to be the entire area from which the artificial window is visible, matching the viewing area of a real window. All of the light from the scene visible inside the viewing area must come through the window surface, so the maximum visible motion parallax is between the viewpoints on the opposite edges of the window. This was also formally discussed in related work under the plenoptic function and its free space properties. Motion parallax is the change in view resulting from change in viewpoint.

The amount of motion parallax depends on the amount of change in the viewpoint, the distance of the objects from the viewer and the distance between the occluding objects. This is illustrated in Figure 5. The first thing to note is that when an object is close to the viewer, its image changes significantly as the viewer moves from one edge of the window to the other. As the distance between the object and the viewer increases, the change in the object image decreases until it remains constant. The second thing to note is that the amount of change in occlusion between two objects decreases as the relative distance of the objects from the viewer increases with respect to the distance between the objects. Same as in the first case, when the ratio between the distance from the viewer and the distance between objects gets high enough, the image of the two objects becomes constant between the window edges.

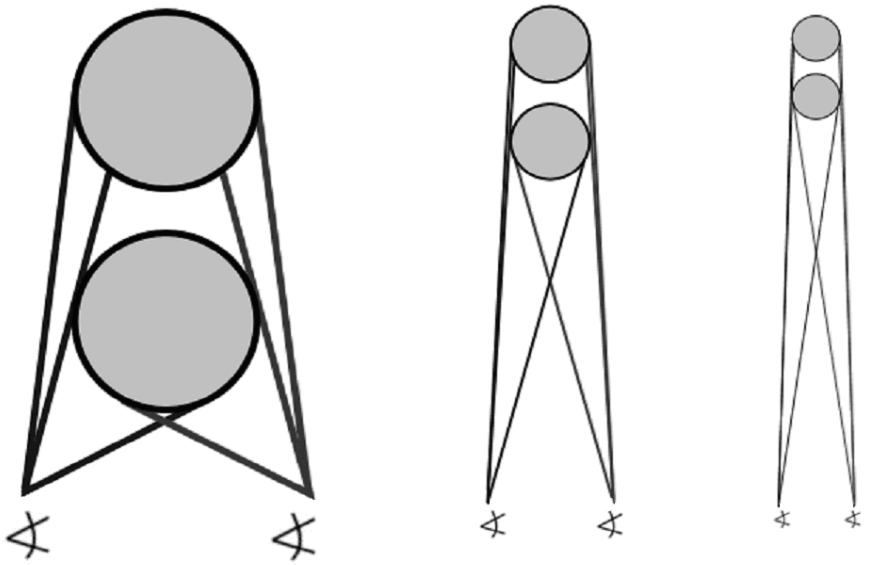

Figure 5: Motion parallax and relative distance

Figure 6: Model scene – Lake Bryan, TX

Using these observations we can analyze the target scene based on scene properties (sample scene is illustrated in Figure 6). Foreground spatial openness implies that there will be no close objects which can exhibit noticeable amounts of motion parallax. Still water and grass which cover most of the foreground area are uniform in appearance. Occlusion between background trees is a potential source of motion parallax, but with their distance from the viewer and relatively small spacing it is not visible. The only noticeable motion parallax is between the trees and the sky, but when there are no clouds to form a distinct pattern, it is not visible either. So, for this particular type of very

beneficial scene, there is no need to model the within-view motion parallax. In addition, the motion parallax between the scene and the wall is amplified by the vection effect. As discussed in related work, vection is the feeling of self movement induced by movement in our visual field. The furthest objects contribute the most effect, so the scene which is perceived as furthest away, magnifies the motion parallax effect with respect to the wall.

Based on the conclusion that within-view motion parallax does not need to be modeled, we can choose the appropriate image-based rendering method. There is no need to use complicated methods, such as light field rendering or even view interpolation. Since the image can be assumed to remain the same over different viewpoints, a simple environment map can be used. The field of view through a window is 180 degrees in both directions and can be best represented with a hemisphere. Geometry of the hemispherical environment map is exactly the same as the spherical map. View projection of the environment map (Figure 7) can be performed in software such as QuickTime VR [7], but it is faster with texture mapping in 3D graphics hardware [53]. In our case it is not only faster, but also better and simpler because we can model the off-axis projection with the same hardware. This allows us to model the scene accurately in the virtual world with 3D transformations that recreate the real world geometry.

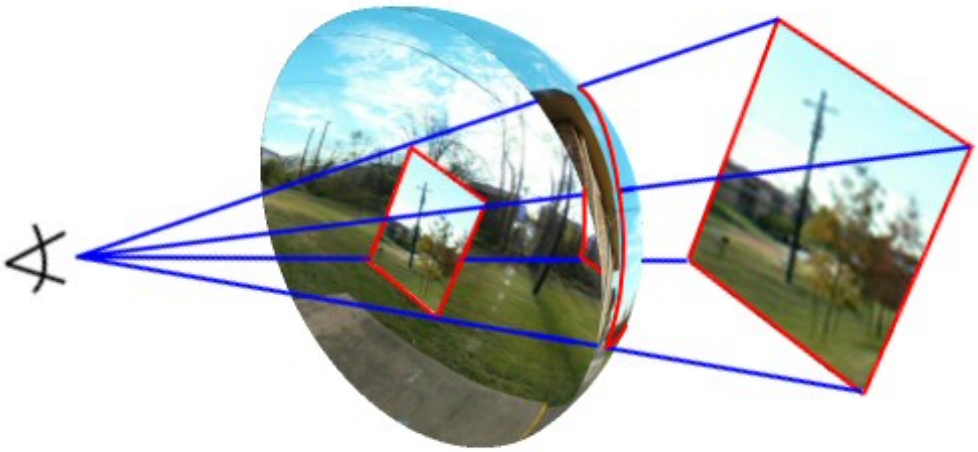

Figure 7: Hemispherical environment map

Next we will look at the theoretical analysis of combining image-based rendering of a spherical environment map with off-axis projection. We want to display the correct section of the environment map with the off-axis camera frustum. The problem is that the captured environment map was taken at the center of the virtual display, while the viewer will be standing somewhere in front of the display and have a different projection of the environment map. This will cause a mismatch in the displayed region between the two environment maps. If we consider the 2D version of the environment map hemisphere and place the display at the center, then all possible views will be coming from one side of the display. Field of view (FOV) is defined by the two end-point rays which are drawn in solid lines for the two viewpoints in Figure 8. The simpler case on the left is a viewpoint on the center-line with only a translation *d*. We can see that the places where the rays hit the captured circle do not correspond to the viewer's circle.

The enclosed circle arc represents the region of the environment map, which is going to be projected onto the display. Arc is defined by the angle α and we can see that on the captured circle the angle (shown in dotted line) is not the same as the angle on the viewer's circle (solid line). We need to eliminate the difference between the arc on the captured circle (*L*) and the arc on the viewer's circle (*l*). We will refer to this difference as the error ε which is derived in Equation 8 (*r* is the radius of the environment map). We can eliminate the error by either making distance *d=0* or the radius *r=infinity*.

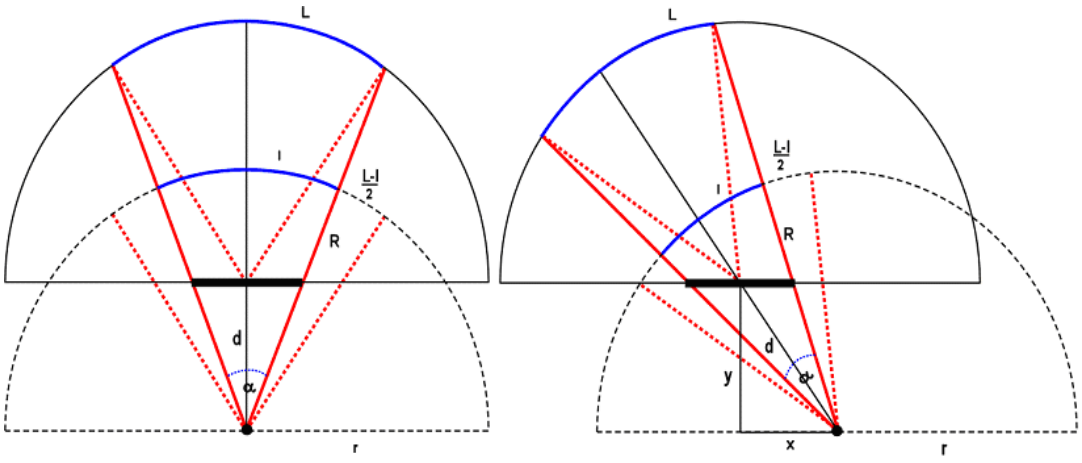

Figure 8: Off-axis environment map

$$
\begin{aligned}
l &= \frac{\alpha \cdot \pi \cdot r}{180} \\
R &= r + d \\
L - l &= \frac{(R - r) \cdot \alpha \cdot \pi}{180} = \frac{d \cdot \alpha \cdot \pi}{180} \\
\varepsilon &= \frac{L - l}{L} = \frac{\frac{d \cdot \alpha \cdot \pi}{180}}{\frac{\alpha \cdot \pi \cdot R}{180}} = \frac{d}{r + d}
\end{aligned}
\tag{8}
$$

In our case of off-axis projection we cannot eliminate *d*, but we can assume that the environment map is infinitely far away. In that case we can set the *r* to be arbitrarily large which will eliminate the translation error, but keep the correct FOV. Given a screen resolution *S* the error can be reduced to sub-pixel size, which will make it unnoticeable to the viewer. For an example, if *d*=60 and *S*=1024 then *r*>30,720.

$$\begin{aligned} \varepsilon &< \frac{l}{S} \\ \frac{L-l}{2} &< \frac{l}{S} \\ \frac{d \cdot \alpha \cdot \pi}{360} &< \frac{\alpha \cdot \pi \cdot r}{180 \cdot S} \\ r &> \frac{d \cdot S}{2} \end{aligned} \tag{9}$$

3.1.2. Optical Tracking

This subsection discusses the theory behind the conversion from camera image coordinates to world coordinates and the calibration of the camera. As discussed in related work, CAMSHIFT is a color-based head tracking algorithm. It gives a quite reliable X and Y location of the head position in the camera image sequence, which can easily be mapped to a line in the 3D world coordinates. The problem is to determine the distance of the head on that line, because it can only be estimated from the size of the head which is very unreliable. This is unacceptable in our system because the image on the display must remain stationary when the viewer is not moving. Otherwise, there will be noticeable shaking in the simulated view which will not only be unrealistic, but also irritating to the user. As we will see in this section, even the X and Y head location will need to be filtered to get stable coordinates. A solution to the distance determination was

found in a restriction of the observer's pose. While using this system, the observer is assumed to be standing. This gives us a constant height of the head from which the actual 3D position can be determined with only X and Y head coordinates from the camera image.

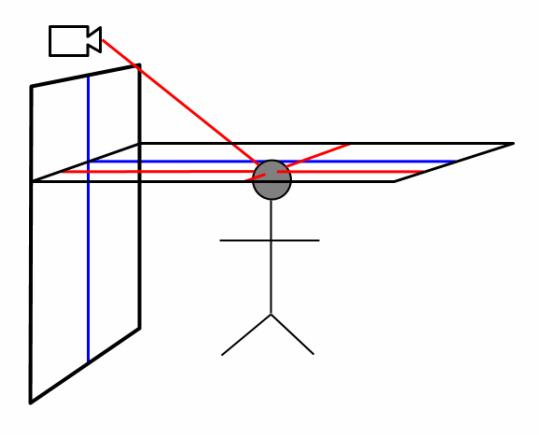

Figure 9: Viewer's head position

The goal of our optical tracking algorithm is to determine the line from the camera lens that passes through the center of the viewer's head and intersect it with the plane of user's eye height (Figure 9). However, since we are using a PTZ camera (rotates horizontally (pan), vertically (tilt), and changes zoom), the camera image has to be related through the camera parameters to the world coordinates. Camera parameters (pan, tilt and zoom) are obtained through the camera interface. Pan and tilt values can linearly be converted to rotation angles in degrees, either from simple measurements or manufacturer specifications. Zoom values however require a much more involved measurement and calibration method. They are related with a polynomial of second

order. In order to find this quadratic function, the relationship between the camera image coordinates and the pan and tilt parameters needs to be measured for several zoom values. The measured data is then used to fit a quadratic curve. The following procedure was used to obtain these values as precisely as possible.

A static object was selected to be tracked at a specific zoom setting. This object was positioned near the edge of the camera image and the camera was panned until the object moved to the other edge, and then back. During this back and forth movement the object's X coordinate and camera's pan value were logged into a file. This file was opened with Microsoft Excel and both movement directions were graphed separately. For each direction a line was fitted through the points (Figure 10 shows one direction). The average of the two slopes was used as the value for that zoom setting. This was repeated for several different zoom settings. The averaged values were graphed and a quadratic trendline was fitted (Figure 11). The final minor adjustments for parameters made sure that the end settings stayed within the measured values. The same procedure was repeated for Y coordinates and tilt values. The choice was made here to use zoom values to map directly X to pan and Y to tilt. An intermediate value, such as focal length, would have been much harder to obtain and use in calculations. This calibration procedure needs to be done only once, as internal camera parameters do not change.

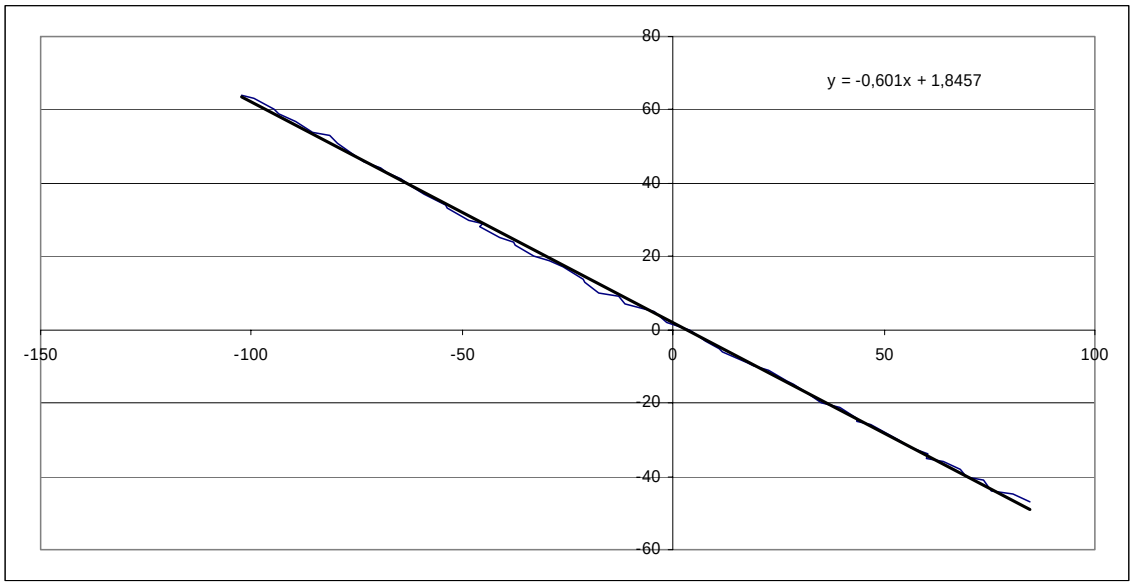

Figure 10: Pan vs. X for Zoom=430

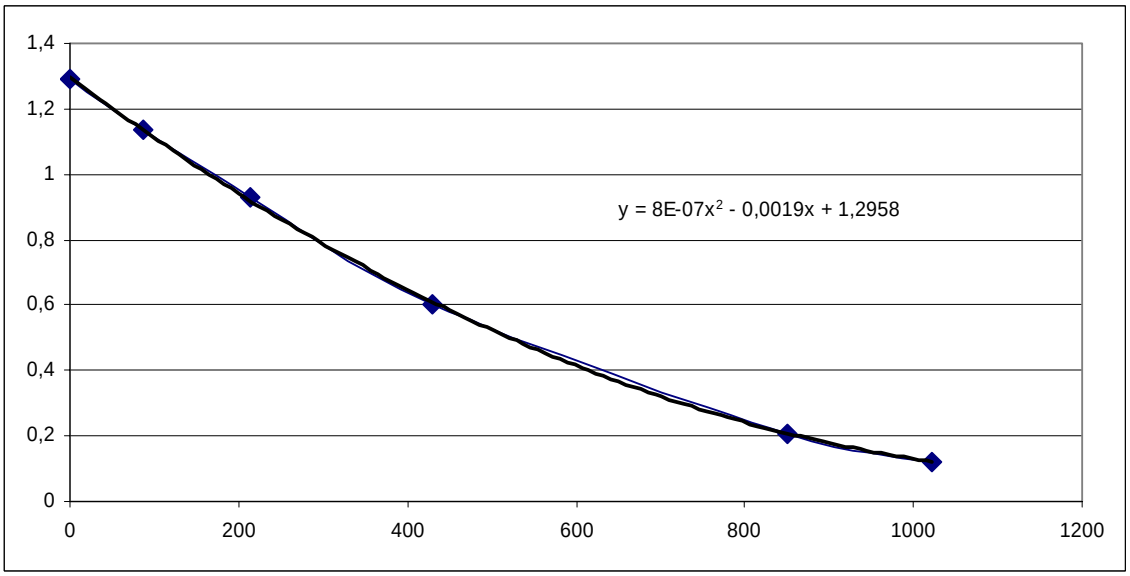

Figure 11: Zoom vs. X-to-Pan conversion factor

Now we have everything we need to find the line from the camera to the viewer's head. From the camera we obtain pan, tilt, and zoom values. From CAMSHIFT we obtain X and Y coordinates of the head in the camera image. Using the zoom value and the mapping function, we convert the coordinates to pan and tilt offsets. Using linear conversion from pan and tilt to degrees, we obtain two angles. These angles can be assumed to be rotations around the axis where the camera is centered at the origin (simple camera model). These two angles can be represented with a unit vector. Since head height is known and constant, this gives us the scalar multiplier for the whole vector and we have the final head coordinates. This is specified in Equation 10.

$$
\begin{aligned}
&\text{headpan} = \text{pan} + \text{mappan}(\text{zoom}) \cdot \text{Y} \\
&\text{headtilt} = \text{tilt} + \text{maptilt}(\text{zoom}) \cdot \text{Y} \\
&\theta = \text{mapdeg}(\text{headpan}) \\
&\varphi = \text{mapdeg}(\text{headtilt}) \\
&\text{x} = \sin(\theta) \cdot \sin(\varphi) \\
&\text{y} = \cos(\theta) \\
&\text{z} = \sin(\theta) \cdot \cos(\varphi) \\
&\text{v} = [\text{x}, \text{y}, \text{z}] \\
&\text{scale} = \frac{\text{headheight}}{\text{y}} \\
&\text{headvector} = \text{scale} \cdot \text{v}
\end{aligned}
\tag{10}
$$

Since the camera used in the actual system has a limited pan range (25 degrees), it was mounted tilted downwards above the display. This introduced an angle offset which was measured during system setup. This angle offset α introduced another rotation resulting in Equation 11. The camera is assumed to be mounted above the display so that the principal point in the central position (0 pan, 0 tilt) is perpendicular to the display.

Any other offsets resulting in translation of the camera with respect to the display center can be dealt with addition or subtraction from the final head coordinates.

$$\begin{aligned} x &= \sin(\theta)\cdot\sin(\varphi) \\ y &= \cos(\alpha)\cdot\cos(\theta) - \sin(\alpha)\cdot\sin(\theta)\cdot\cos(\varphi) \\ z &= \sin(\alpha)\cdot\cos(\theta) + \cos(\alpha)\cdot\sin(\theta)\cdot\cos(\varphi) \end{aligned} \tag{11}$$

The last thing that needs to be dealt with are the fluctuations in CAMSHIFT coordinates. If raw values of head coordinates are used, the image on the screen shakes. This is because, with each frame, the image is just a little different, even if the viewer's head is not moving. These movements, of just a few pixels in the camera image, are visible in the simulated view. Not only is this view shaking not realistic, it is extremely irritating to the viewer and must be eliminated. Obviously, when the viewer is not moving, the image can not move. The simplest way to eliminate this shaking is to reduce the resolution of the head coordinates. This eliminates small movements, but it also eliminates the illusion of continuous change in the viewpoint, due to noticeable jumps. Another problem with resolution reduction is that if the values fluctuate exactly on the border between two different values, even at the lower resolution, the value will keep flipping. A better version of this method is a sliding window. In this case the value changes only if the new value is different from the old one by more than some threshold. A sliding window was used to eliminate sub-pixel movements before further filtering (threshold set to 1.2 pixels).

For more sophisticated filtering, a running average was attempted. The value was averaged over some number of samples. However, it still did not give satisfactory results. It either fluctuated too much or was delayed too much, causing too much of a

lag. A solution to this problem was found in digital filtering and in our case we needed a low-pass filter with a small delay. We chose a Butterworth filter which is a special case of a Chebyshev filter with 0% ripple or, in other words, is maximally flat [56]. The two parameters we needed to determine were the cutoff frequency and the filter order (number of poles). When the user moves his head, it results in slow movements in the detected head location with respect to the oscillations caused by the imperfections in the head tracking algorithm and conditions. When we perform a fast Fourier transform (FFT) and graph it, we can see different frequencies in the signal. The cutoff frequency was determined by recording the X coordinate into a file, importing it into SIGVIEW [54] and looking at the FFT graph. A fast moving head was compared with a static head and the cutoff frequency was set at 0.066 Hz (2 Hz out of 30 Hz sampling frequency). The filter code was adopted and modified from code found in [35]. The order was determined by trial and error. The best tradeoff between roll-off and delay was found to be for a filter order of 2.

All of the parameters discussed so far can be determined for equipment in general, prior to a specific system setup, except for the camera offsets. Once the system is installed at a specific location, the optical tracking head coordinates need to be aligned with the display. This is done by measuring the camera location with respect to the display center. Once these final offset parameters are determined, a guess of the camera mount angle is made. Then the system is initialized and a measuring tape is put on the ground in front of the display. The angle is adjusted until the Z values (display distance) on the display match the actual values read on the measuring tape. This is far from a

perfect method, but for this prototype setup it was more than enough. Head position values from optical tracking were observed to be close enough to the real position.

3.2 Implementation

3.2.1. Scene Capture

In order to determine the required resolution for scene capture we must consider the real window view again. In a real window, field of view (FOV) varies with the position of the viewer. The further away the viewer is, the smaller the FOV, and when recreated on the display, the more the captured scene data needs to be stretched. This magnification is the primary factor in determining the required resolution. Another way of thinking about this effect is to imagine an object which is far away. In a real window, as we step away, the object stays roughly the same perceived size. This is due to the fact that object size is defined by the angle it covers in the image seen by our eyes [26]. When the object is far away, the change in angle is so small that it is virtually unnoticeable. A virtual object is actually displayed close to the viewer and to simulate this size consistency constraint, the object must grow in size. This is illustrated in Figure 12.

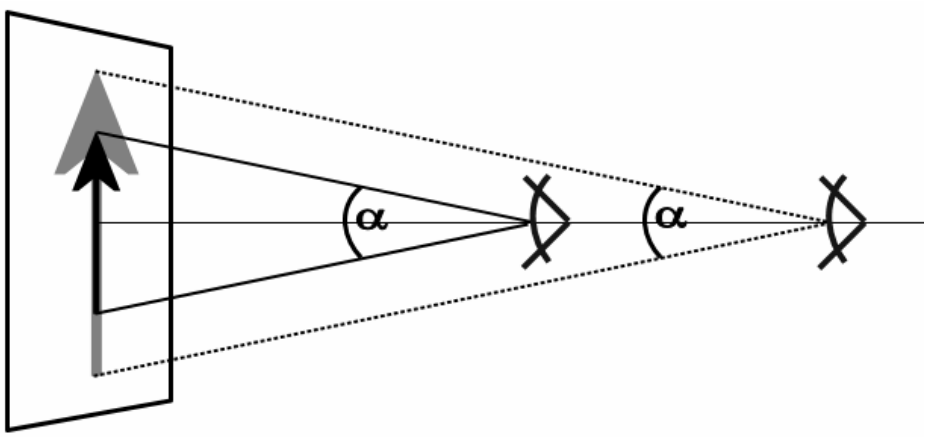

Figure 12: Size consistency of a virtual object

If the scene is not captured in high enough resolution, the image on the screen will start to look blurred as the observer steps away. The desired capture resolution depends on the distance from the display the observer may move before the picture deteriorates. Although human eyes have a limited resolution [26], and at a certain distance picture blurriness becomes unnoticeable, it cannot be reached as we will be limited by the hardware. Using Equation 12, Table 1 was filled out for several distances for the 42" plasma display used in our system. The display's maximum horizontal resolution is 1024 pixels. From the table, we can see that if we want a viewer standing 120 inches away from the display to see an image in full display resolution, the captured environment map of the scene will have to be 10516 pixels wide. Since textures in OpenGL need to be powers of two, rounding it down to 8192, gives us 8192*8192*3 bytes or almost 200 megabytes of image data (in 24-bit color). If we double the resolution, the size grows to 800 megabytes, which is more than what our hardware can handle (P4 2.4GHz with

512MB RAM and 128MB Video RAM). A real-time reproduction of such a large texture would not be possible, even if we could produce it.

$$\tan(\alpha) = \frac{w}{2 \cdot d}$$
$$\alpha = \tan^{-1}\left(\frac{w}{2 \cdot d}\right) \tag{12}$$

d	α	S	S	S	S
0	180	1808	2911	5378	10516
15	102	1024	1648	3045	5955
30	63	636	1024	1892	3699
60	34	344	554	1024	2002
120	18	176	283	524	1024

Table 1: Horizontal resolutions for 42" plasma display

Although there are ways to capture the scene in a single shot, such as fisheye lens, because of the desired resolution, we used image stitching. This is because currently the best digital cameras record at 12 megapixels (MP), which is equivalent to 4048 by 3040 pixels. In order to obtain the scene resolution of 8192 by 8192 pixels, several images at 12 MP needed to be stitched together. Figure 13 summarizes the capture process.

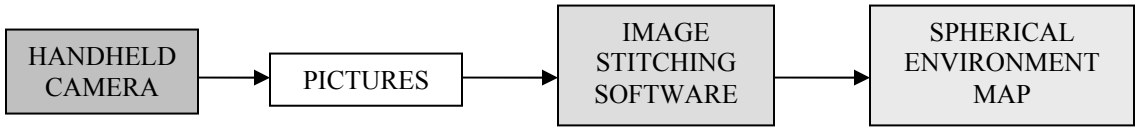

Figure 13: Scene capture process

As discussed in related work, image stitching is the process of combining multiple pictures into an environment map. It involves aligning, combining and blending of input images [49]. In current stitching software, the user roughly aligns the images and specifies a number of matching points between overlapping images. This is tedious, time-consuming work and the number of input images should be minimized as much as possible. Regular cameras have maximum horizontal FOV (HFOV) of approximately 50 degrees. In order to cover the 180 by 180 degrees scene, with at least 30% of overlap between images, it would take many more pictures than necessary for the final scene resolution. This is solved with the use of a wide-angle lens to increase the camera's HFOV. The scene used in our system was recorded with a Fuji FinePix S7000 and Phoenix Super Fisheye 0.25X lens. At minimum zoom this setup produces a circular fisheye image with roughly 140 degrees of HFOV. In the actual recording it was zoomed in to a full-frame fisheye image with maximum HFOV. This was done to reduce the distortion and maximize use of image area. The final stitched environment map was roughly twice the width of the image, which implies that camera HFOV was around 90 degrees.

The image stitching software requirements were: support for spherical projection (equirectangular projection), multi-row stitching, fisheye-lens, lens distortion correction, field of view determination and color correction. A portal for panoramic photography called Panoguide [47] has an excellent comparison of panoramic software across these requirements. The one that satisfies all the requirements is an open source program called Panorama Tools (PanoTools) [13] by Helmut Dersch. A problem encountered

with this software was that the graphical user interface didn't work correctly. The probable reason was a mismatch in Java version, but even after installation of the correct version it still would not work. However, another commercial program called PTGui [51] exists which provides a windows executable graphical user interface to PanoTools.

A tutorial [41] for this software discusses most of the issues encountered in our scene capture process. PanoTools requires a good estimate of the camera HFOV before the user can start arranging images. Using the formula from [41] in Figure 14, a measuring tape was recorded at a known distance and the HFOV was estimated at 90 degrees.

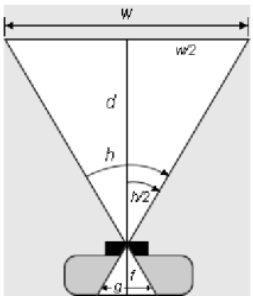

Figure 14: HFOV determination

However, before images could be stitched together they had to be preprocessed. Due to the small incident angle at the edges of fisheye lens, noticeable amounts of chromatic aberration were present in the recorded image. Chromatic aberration is unequal magnification of colors resulting in a shift of a color channel with respect to other colors [41]. In the case of a digital photograph, the three channels are red, green and blue (RGB). The fisheye lens produced radial chromatic aberration (Figure 15 from [41]) or a

shift away from the image center in all directions. This shift is a function of distance, so the further the point is from the center, the more the colors are shifted. This was most noticeable in the corners which are farthest away. The Radial Correct function of PanoTools was used to correct the image. The formula for correction is given in Equation 13.

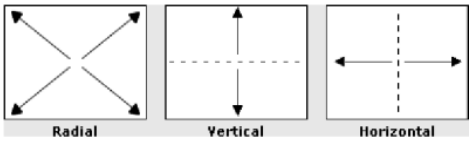

Figure 15: Types of chromatic aberration

$$\mathrm{rsrc} = \mathrm{a} \cdot \mathrm{rdest}^4 + \mathrm{b} \cdot \mathrm{rdest}^3 + \mathrm{c} \cdot \mathrm{rdest}^2 + \mathrm{d} \cdot \mathrm{rdest} \qquad (13)$$

The explanation provided with PanoTools is very vague and summed up as “use b only”. The tutorial explains it in a little more detail. Coefficient *a* affects the outer pixels more than inner, *b* affects pixels closer to the center and *c* affects the center pixels mostly. However, how to determine these coefficients is left unsaid. It took a long time to find the right *a*,*b*,*c*, and *d* parameters for the radial shift function. The first misleading tip was to set *d*=1-*a*-*b*-*c* to keep the size constant. In our case the colors were so misaligned they needed to be moved and resized. After realizing that *d* should stay 1, it took some guessing on where to start the process. Through trial and error the right combination of *a*,*b* and *c* that minimizes the chromatic aberration was found. In the end, in our case, *a* compensates for *b* on the edges, while *c* keeps it all balanced on the inside. The final process involved 10 images with approximately 90° HFOV recorded at 12 MP

and 272 manually selected matching points. From these images and matching points, PanoTools produced a final hemispherical environment map which was touched up in Photoshop. Figure 16 shows the result as an equirectangular projection, which is a way of unwrapping a sphere. The square covers 180 degrees both horizontally and vertically. It has significant distortion at the top and bottom where the image appears stretched. This is due to the fact that the last line actually maps to a single pixel on the pole of a sphere. This is not the most efficient way to represent a spherical map, but it is the easiest to produce and use in OpenGL texture mapping.

3.2.2. System Overview

Our system consists of the hardware listed in Table 2. Once the system is initialized, the camera tracks the observer as long as the observer's face is visible to the camera. The camera turns and zooms to keep the face in view. Below the camera is a large display which shows the window view that should be seen from the observer's position. As the observer moves around, the image on the display changes to match the viewing position

Figure 16: Final environment map

(Figure 17). A light is mounted above the camera to provide constant lighting. The system interacts with the observer and provides the correct image, as long as the observer makes an attempt to face the camera, stand upright, and avoid very sudden movements.

Display:	Plasma Sampo PME-42X6 (42” diagonal, 1024*1024 native resolution)
Computer:	Compaq d325 AMD Athlon XP 2400+ 512MB 80GB
PTZ camera:	Sony EVI-D30 (100° pan, 25° tilt, 12x zoom)
Video capture card:	Prolink PixelView PlayTV PVR
3D graphics card:	ATI Radeon 9200 SE 128MB

Table 2: System hardware

Figure 17: System setup

A functional overview of the system is given in Figure 18. Starting from the observer, the camera provides the image of the observer’s face through the video capture

card. This image is analyzed by the head tracking software and the location of the head in 3D space in front of the display is calculated. This location is used to calculate the correct off-axis viewing matrix. This matrix is combined with the captured scene by rendering a hemisphere textured with the spherical environment map of the scene. OpenGL and 3D graphics hardware are used to render the final image on the display. This process is repeated at the full video rate of 30 FPS, providing a smooth image transition as the viewer moves in front of the display.

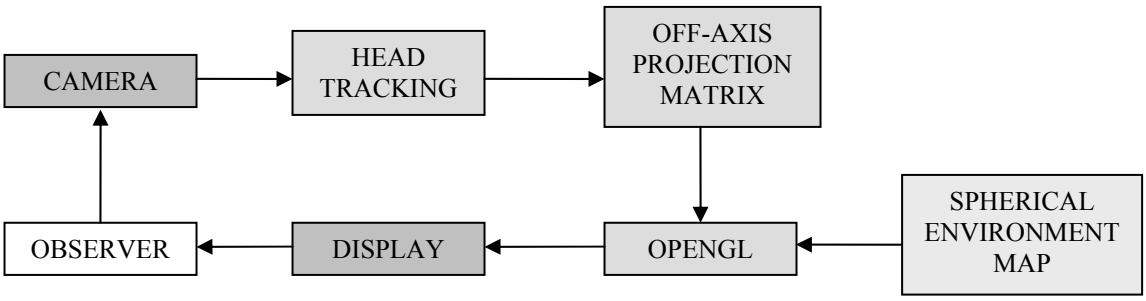

Figure 18: System overview

This process is done with three concurrent threads (Figure 19). The main program thread initializes all program parameters and the other two threads. The camera thread constantly polls the COM port requesting the camera parameters (pan, tilt and zoom) and a DirectShow thread runs the head tracking filter which finds the head location in the camera image. The main thread takes information from the other two threads, calculates the head position, and renders the image. The DirectShow thread is used as the synchronization signal because it updates the head position 30 times per second, while the camera interface thread performs updates only 10 times per second. The main thread

also takes input from the keyboard used by the system operator to calibrate the system and initialize the viewer.

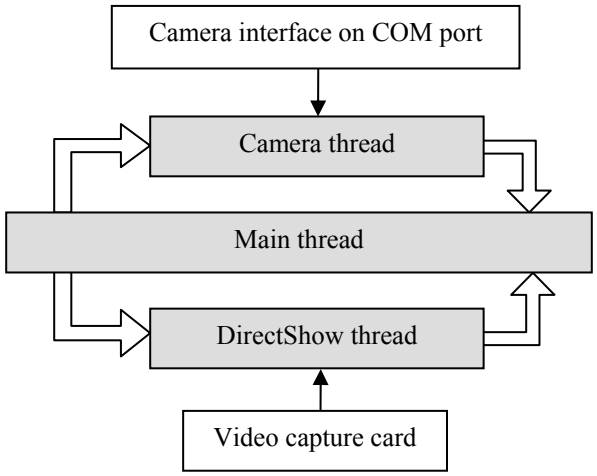

Figure 19: System threads

3.2.3. Camera Interface Thread

The camera interface thread retrieves camera pan, tilt, and zoom. The Sony EVI-D30 camera is connected through an RS-232C (VISCA) cable to the computer serial port. The camera interface thread constantly performs camera queries and updates the program variables. Variable updates are synchronized with the main thread through a mutex. Only inquiries (no commands) are sent to the camera. Tracking of the observer is done with the auto tracking (AT) function of EVI-D30, which allows the camera to track a target autonomously based on color and brightness. In a setting with controlled background and lighting, the camera tracks the observer well enough for our system prototype.

The code used to interface with the camera was extracted from remote camera control software developed at the Berkeley National Laboratory [14]. Their application contains code for giving commands to the EVI-D30, which was modified using the command list manual [57] to send inquiry commands and retrieve data instead. All inquiry commands send a few bytes to the COM port and receive a few bytes back. Maximum speed of inquiry was measured to be 10 times per second for the three parameters. It takes two inquiries to get the parameters, one inquiry for pan and tilt, and one inquiry for zoom.

3.2.4. DirectShow Thread

The camera image processing code comes from the OpenCV library [45] in the form of a DirectShow [36] filter. A DirectShow filter is an excellent choice because it allows direct connection with the video capture card and takes care of all of the synchronization issues. The CAMSHIFT filter is connected with the capture card by forming a filter graph. During the normal operation of the system, video output is not visible on the display. However, during the initialization it is positioned in the top left corner, allowing the system operator to select the target.

In the original filter, once the filter graph is started, there is no way to perform real time frame-by-frame processing and no way to query the filter for the current head position. The CAMSHIFT filter code was modified to take control of this sitatuion. When the filter is being initialized, it pops up its properties window through the DirectShow. In that window, the operator can select the target and start the filter. There is also a structure associated with the filter properties, usually used to set the filter

parameters from the caller thread. The CAMSHIFT filter was modified to update these filter properties with the head coordinates. In addition, at the end of processing of each frame an interrupt (WM_GRAPHNOTIFY) is thrown, which wakes up the main thread and acts as a synchronization signal. In this way, the DirectShow thread functions independently of the main thread, but as soon as a frame is processed, the main thread redraws the view. During the scene drawing in the main thread, the CAMSHIFT filter proceeds concurrently with the next frame, as soon as it is available.

Besides CAMSHIFT, OpenCV contains an algorithm based on a boosted cascade of haar-like feature classifiers [31]. It is one of the fastest and most reliable algorithms currently available. The CAMSHIFT algorithm was chosen due to unacceptably high error rate of the feature-based algorithm. The pre-trained algorithm supplied with OpenCV often didn't recognize the face in the camera image. This is unacceptable for our system, because even a few frames with no data on face position ruin the illusion of a simulated window view. An attempt was made to use the training algorithm from OpenCV to create custom training data, but it failed in the current version (beta 3.1) due to bugs in the code which were beyond our control. Possibly in the next version, if it is fixed, the feature-based algorithm could be used instead of CAMSHIFT.

3.2.5. Main Thread

The main thread starts by loading the system parameters (calibration data, display size, etc). It proceeds with initialization of the Butterworth filter, the DirectShow filter graph, and the camera interface mutex. After creating the OpenGL window and loading the scene texture, the other two threads are started and the main thread waits for

messages. Messages that cause an action are the DirectShow interrupt or one of the operator keys. If a key is pressed, the appropriate parameter is changed. Keys and parameters are described in the next section.

The equirectangular hemispherical environment map is loaded with OpenGL as a texture. However, it is not loaded as one texture, because of size limitations. Instead, it is split into strips, which are matched in the hemisphere rendering process. The problem with splitting the texture is that when mip-mapping is used, edges between the texture strips become noticeable. This is due to the fact that at the edge of a strip there is no data about the neighboring strip to use in anti-aliasing. The best solution is to create an overlap between the strips and adjust the texture coordinates while rendering [53].

The scene is saved in JPEG format and a free library routine is used to read the data from it [55]. In BMP format the scene is too big (190 MB) and, at the maximum quality, the JPEG format has no noticeable degradation while significantly reducing the file size (27 MB). In the final implementation, the chosen number of strips was 34 with an overlap of 8 pixels. This slightly changed the vertical resolution, resulting in the final image of 8192 by 8176 pixels. The original image was shrunk to a height of 8160 pixels and 8 pixels of padding were added on top and bottom. This produces strips with height of 256 pixels and width of 8192 pixels which is an acceptable texture size for OpenGL. The textured sphere rendering algorithm was taken from [5] and modified to render only half of the sphere in texture strips. For each strip the texture is changed and the texture coordinates are modified to exclude the overlap region.

When the DirectShow filter interrupt is received, CAMSHIFT coordinates and camera parameters are retrieved from the other two threads. CAMSHIFT coordinates are filtered with the Butterworth filter to eliminate noise. Viewer's head position is calculated and the off-axis projection is set. The old frame is erased, the hemisphere is rendered, and frame buffers are swapped. Finally, the main thread goes back to sleep.

3.3 System Operation

3.3.1. Operator Interface

The first thing that the system operator needs to do after system installation is system calibration. Then, for each viewer, the system needs to be reinitialized. Finally, some additional functions added specially for testing purposes need to be accessed. A complete list of keys available to the operator and their functions is given in Table 3.

F - show video	Y - increase movement scale	Z - toggle status display
G - hide video	U - decrease movement scale	X - decrease eye height
Q - change filter	P - filter properties	C - increase eye height
S - decrease delay	I - decrease width filter fc	V - start/stop
D - increase delay	O - increase width filter fc	B - flat image only (for testing)
E - change pan freeze	H - decrease move filter fc	T - start timer
W - start/stop write to file	J - increase move filter fc	ESC – quit
M - increase angle offset	K - decreas filter order	
N - decrease angle offset	L - increase filter order	

Table 3: Operator keys

When the system is started, a black screen is displayed with the camera window in the top left corner. Pressing key P causes a filter properties window to be brought up and the operator can select the area belonging to the observer's head and start the system. At

this point, the system will start working. The camera auto tracking must be initialized separately with the camera remote control. Viewer's eye height must be adjusted (keys X and C) to obtain the correct view. During the calibration procedure, the camera tilt angle offset can be changed (keys M and N). Key Q toggles between a simple average filter and the Butterworh filter. It is recommended to stay with the Butterworth filter and find the right settings for it. Movement scale should also be left as is (keys Y and U) to eliminate only sub-pixel movements. Filter cutoff frequency can be modified for the head size (keys I and O) and head position (keys H and J). Filter order is the same for both properties (keys K and L). For the simple average filter, delay can be changed (keys S and D) and higher delay results in less view shaking. For internal calibration data logging can be used, which writes current parameters to a file (key W). Finally, when everything is set, the camera view window can be closed (key G). It can be brought back as needed (key F). During calibration, the parameters are displayed at the bottom of the screen. This status display needs to be removed when done (key Z). For testing purposes, the system can be stopped (black screen) with key V, paused at a random position (still image) with key B, and the timer can be started (5 minutes) with key T. Escape key exits the system.

3.3.2. Results and Limitations

It takes some time to calibrate the system and the calibration is not very exact, but it is good enough for a prototype. Each user needs to be initialized for the camera auto tracking and then again for the CAMSHIFT filter. In addition, each user's eye height needs to be adjusted. However, once the viewer is initialized the system functions

autonomously as designed. For testing purposes, the limitation of a standing pose was not a problem. The generated view looked natural and changed naturally.

However, there were two unresolved problems. The first problem was always present and appeared when the camera moved. Camera position was updated 10 times per second, while head position in the camera video was updated 30 times per second. This resulted in noticeable jumps in the simulated view during camera rotation. This problem was slightly relieved by setting the camera auto tracking mode to move only when necessary.

The second problem occurred for a small number of the participants in the evaluation study of the system. Due to skin color differences, some participants had problems with camera auto tracking. The camera failed to track one participant completely, but most of the participants had few problems. The camera would lose some participants, but pick them up when they returned to the camera viewing area. Participants wearing red clothes had to cover them during system use.

In addition, limitations compared to a real window that were set in the initial design and are not a result of implementation are: one viewer at a time, lack of stereopsis, and a timewise static scene.

4 EVALUATION

A study using a quasi-experimental design was carried out wherein 14 participants/volunteers provided ratings of the effects of the artificial window prototype and a static display on their moods and reported behaviors. Participants also evaluated the artificial window in comparison to the static display with respect to realism, preference, and enhancement of the indoor environment. Finally, participants completed tasks designed to evaluate the user interface of the artificial window.

4.1 Hypotheses

As noted, previous studies have shown that views of nature increase arousal, positive affects or feelings, and interest, and reduce negatively-toned feelings such as tension [68]. We hypothesized that the artificial window view of nature would elicit more positive mood responses than the static nature picture. It was also anticipated that participants would evaluate the artificial window as more realistic and preferred than the static display, and report looking at it and moving in front of it more. Our hypothesis was that the artificial window, in comparison to a static picture, would be more pleasing, improve the environment more, be more realistic, more desired, more involving and a better replacement for a window.

4.1.1. Mood Ratings

The mood measurement questionnaire consisted of established and validated items used in environmental psychology and emotions research [34,58]. Participants rated their moods on five affective dimensions: arousal (stimulated, excited), positive affects (pleasantness, friendliness), interest, tension, and anger. Two questionnaire items were

used to assess each of these five mood aspects. Participants were asked to rate how they felt "right now, at this time" using visual analog scales (VAS) of length 100 mm. The VAS end points were defined as "not at all" and "very much so." Questions are listed in Table 4.

Arousal	Q1	I feel stimulated.
	Q2	I feel excited.
Tension	Q3	I feel on edge.
	Q4	I feel tense.
Anger	Q5	I feel angry.
	Q6	I feel irritated.
Positive Affects	Q7	I feel pleasant.
	Q8	I feel friendly.
Interest	Q9	I feel attentive.
	Q10	I feel interested.

Table 4: Mood questionnaire

4.1.2. Other Evaluation Ratings

In addition to the mood ratings, the participants were asked to evaluate each of the different picture/window conditions with respect to preference, realism, involvement and suitability as a replacement for a real window (Table 5). The same VAS response format was used as for the mood items.

Q11	I looked at the display on the wall (not the camera).
Q12	I moved in front of the display.
Q13	I liked the scene on the display.
Q14	I think the scene made the hallway better.
Q15	I found the scene realistic.
Q16	I would like to have this scene on my wall (disregard equipment).
Q17	The scene made me feel like I was somewhere else.
Q18	I find this scene an acceptable replacement for a window.

Table 5: Evaluation questionnaire

4.1.3. Interface Evaluation Tasks

The artificial window can also be seen as an artificial reality interface that is simulating a real window. Participants were given two tasks to see how well this was accomplished (Table 6). The first task was to find an object in the scene (water tower) by moving in front of the artificial window as one would move in front of a real window. The second task required the participants to count the number of objects (floating markers in the lake) in the scene, thus requiring them to search the entire scene.

T1	Find the water tower in the display and briefly describe its location.
T2	Count the floating markers in the lake and record the number.

Table 6: Interface tasks

4.2 Experimental Setup

Participants were run individually. Each participant provided all data during one experimental session that followed a within-subjects (repeated measures) protocol. The artificial window was hung in a part of a windowless hallway where everything else was removed from the walls. All other conditions or features in the hallway were kept the same except for the window condition (static picture, virtual window), which was varied during each experimental session. The camera was tracking the participants across all conditions, including when its use was unnecessary during the static picture condition.

Participants were met by an experimenter prior to walking into the hallway study area and asked to fill out the mood questionnaire. The experimenter left while the participant provided the mood ratings. (The experimenter was also away when

participants filled out the mood and evaluation questionnaires in later stages of the experimental session.) The first mood ratings were considered the initial state (IS). After the initial mood ratings, each participant walked into the test hallway in front of the system prototype where the camera was calibrated to track their face. Participants spent five minutes in front of the system prototype with a blank display (BD) or control condition. It was expected that the auto-tracking camera would be novel to the participants and accordingly influence their mood. The purpose of the BD condition was to help participants become accustomed or habituated to the moving camera, in order that its effects would not obscure or swamp those of the two main conditions (static picture, artificial window). The camera was not recording and each participant was informed of this several times during the experimental session.

After the five minutes of BD exposure (camera habituation), participants walked back to the desk where they filled out both the mood and evaluation questionnaires. Five minutes was chosen as the minimum amount of time required for the environment to affect the participants, based on previous research [68]. The same procedure was repeated – five minutes followed by the mood ratings and evaluation questionnaire – for the static picture (SP) and the artificial window (AW) conditions. The sequence of measured conditions is summarized in Table 7. Finally, after providing mood and evaluation responses for the AW, participants were given the interface evaluation tasks. The total time for each experimental session was 30-35 minutes per subject.

As earlier noted, a total of fourteen (14) volunteers participated, of which eleven were undergraduate and three were graduate students. Undergraduate students were given extra course credit for their participation.

IS	Initial state (at walk in)
BD	Blank Display (camera effect)
SP	Static Picture
AW	Artificial Window

Table 7: Measured conditions

4.3 Evaluation Results

All VAS responses were measured with a ruler and entered into a spreadsheet. Mood data were summed into two-item affect categories (arousal, positive affects, interest, anger, tension) and statistical analysis was performed. Table 8 summarizes the statistical analysis of both the mood data and the evaluation questionnaire. Given the *a priori* hypotheses, the environmental comparison of primary interest is between the static picture (SP) and the artificial window (AW). The paired samples t-test results for this comparison are given in the last column in bold (SP-AW). For the evaluation questions (Q11-Q18), the last three columns list differences among the BD, SP, and AW conditions. The rest of the table is provided to enhance the discussion. Within-subjects ANOVA for repeated measures shows the probability that all of the measured conditions do not have the same mean.

	ANOVA	Paired samples t-tests					
		IS-BD	IS-SP	IS-AW	BD-SP	BD-AW	**SP-AW**
Arousal	0.003	0.120	0.933	0.008	0.063	0.032	**0.009**
Tension	0.541	0.287	0.480	0.330	0.585	0.985	**0.588**
Anger	0.235	0.044	0.166	0.089	0.532	0.865	**0.665**
Positive Affects	0.011	0.012	0.008	0.298	0.062	0.620	**0.007**
Interest	0.089	0.357	0.211	0.425	0.373	0.026	**0.032**
Q11	0.000	*N/A*	*N/A*	*N/A*	0.005	0.000	**0.000**
Q12	0.013	*N/A*	*N/A*	*N/A*	0.433	0.064	**0.000**
Q13	0.000	*N/A*	*N/A*	*N/A*	0.000	0.000	**0.000**
Q14	0.000	*N/A*	*N/A*	*N/A*	0.000	0.000	**0.052**
Q15	0.000	*N/A*	*N/A*	*N/A*	0.009	0.000	**0.000**
Q16	0.000	*N/A*	*N/A*	*N/A*	0.000	0.000	**0.001**
Q17	0.000	*N/A*	*N/A*	*N/A*	0.016	0.000	**0.000**
Q18	0.000	*N/A*	*N/A*	*N/A*	0.105	0.001	**0.000**

Table 8: Statistical analysis results

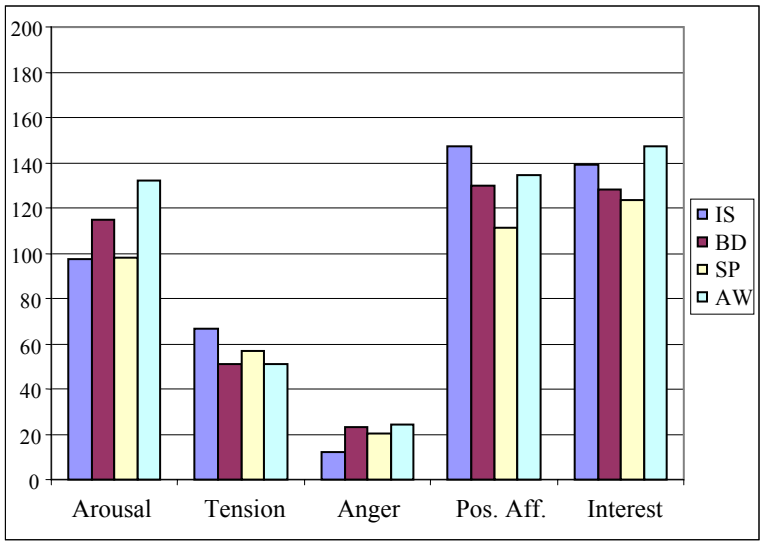

Figure 20: Mood means

Question	0	1	2	3
Arousal	97	115	98	132
Tension	66	51	57	51
Negative affects	12	23	21	24
Positive affects	147	130	111	135
Interest	139	128	123	147

Table 9: Mood mean values

4.3.1. Mood Results

Mood ratings means are graphed in Figure 20 and listed in Table 9. Confidence intervals are graphed in Figure 21. The following moods were all higher for the AW compared to the SP condition: arousal (mean 98 vs. 132, p=0.009); positive affects (mean 111 vs. 135, p=0.007); and interest (mean 123 vs. 147, p=0.032 which is not significant overall given that the ANOVA failed). These findings clearly support the hypothesis that the AW would elicit more positive mood responses than the SP. Tension and anger, however, did not vary significantly as a function of AW versus SP exposure, probably because participants reported low levels of these affects when the session began. Future research on tension or anger mitigating effects of artificial windows should use participants who have been exposed to a condition or manipulation that elicits these feelings [68].

The predicted effect of initial exposure to a novel tracking camera was noticeable in positive affects, where the influence of the BD control condition did not vary from that of the AW. The BD condition with tracking camera, however, was associated with significantly higher positive affects than the SP (0.007). As expected, the participants appeared to habituate quickly to the camera, as evidenced by the drop in the positive affect scores for the SP condition – even though the only change between conditions was a picture on the display. The camera effect was not statistically significant for arousal and interest.

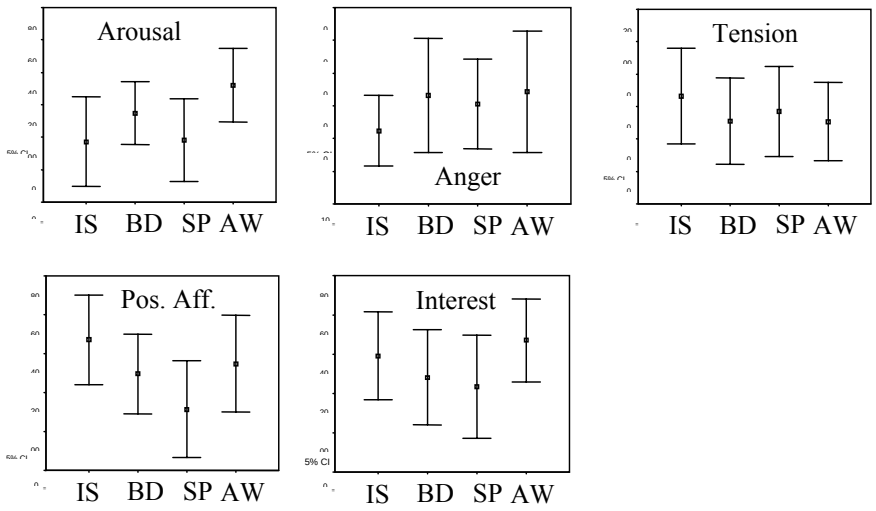

Figure 21: Mood confidence intervals (95%)

In addition to supporting most of the mood hypotheses, other aspects of the mood findings are worth noting. The arousal scores for the SP were at the same level as the IS, whereas they were higher for the AW. Positive affects scores were highest at the IS, perhaps because participants came from outdoors and the surroundings campus has prominent trees and other nature. No comparison is possible for interest scores since ANOVA was not significant.

4.3.2. Evaluation results

Evaluation item means are graphed in Figure 22. For reference, the evaluation questions are listed in Table 5. For nearly all questions, the SP scored between the BD and the AW, just as one would expect from earlier research on static nature pictures. The only items for which this is not true are question 12 – "I moved in front of the display" – and question 14 – "I think the scene made the hallway better".

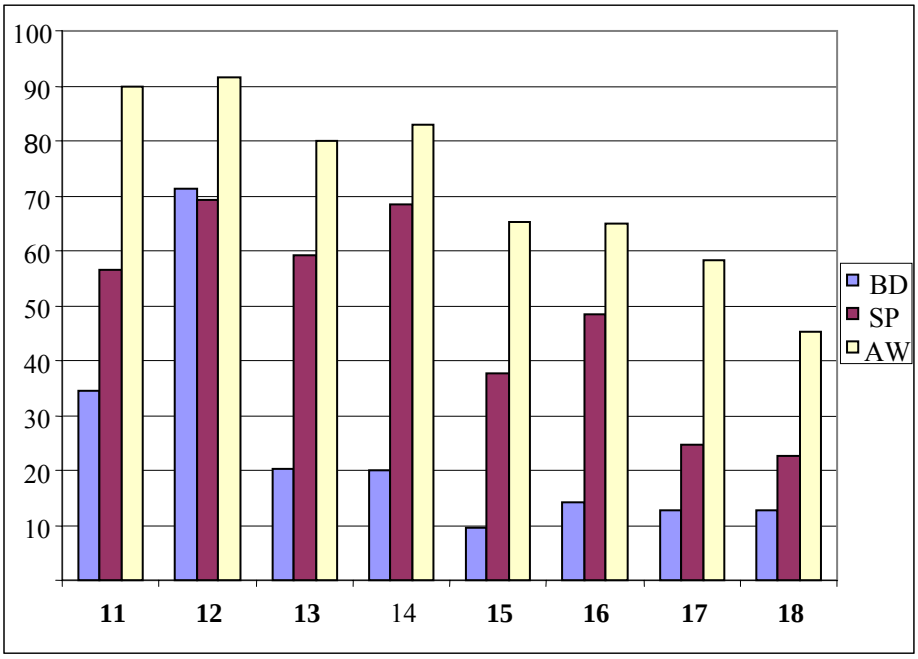

Figure 22: Evaluation means

The BD and the SP received similar scores for movement, reflecting the fact that there was no change in the image resulting from movement. For the same question, there was a nonsignificant tendency for the AW to receive higher scores than the BD with the novelty of the tracking camera (0.064). Question 14 also narrowly missed significance (p=0.052) between the SP and the AW. There are two possible explanations. The first one is that it was too general and might have confused people, thus giving a large variance, although the means were different (68 vs 83). The second explanation is that some people marked the maximum for the SP condition and as a result they could not give a higher rating for the AW.

We can see from the statistical analysis in Table 8, the graph in Figure 22, and the values in Table 10, that the overall evaluation score was far higher for the AW than for the SP ($p \approx 0.000$ and means almost doubled). The positive change or difference in the overall evaluation from the SP to AW is roughly similar to the overall change from the BD to SP.

	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
BD	34	71	20	20	10	14	13	13
SP	56	69	59	68	38	48	25	23
AW	90	91	80	83	65	65	58	45

Table 10: Evaluation mean values

To summarize the evaluation findings, the AW compared to the SP was more preferred (48 vs. 65), judged much more realistic (38 vs. 65), far more involving (25 vs. 58), and a much better replacement for a window (23 vs. 45).

4.3.3. Interface analysis: task results

All participants located the target object successfully in the artificial window, except for one person who misread the question (indicated by their answer). The artificial window interface was clearly recognized as the participants immediately moved the same way a person would in front of a real window. They moved up closer to get a larger field of view and then moved sideways in search of their target.

Twelve out of fourteen participants counted the markers correctly. There were four markers, of which two were very close. The two wrong answers were off by one, raising the possibility that they counted the two markers as one. The participants spent more

time on this task looking in all possible directions to make sure all markers were counted. Success of this task showed that they had a good perception of the virtual space because all markers could not be seen at once. None of the participants needed more than five minutes to complete the tasks.

5 CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

We have presented a new use of display technology – not to display information, but to simulate an environmental stimulation that is psychologically and physiologically beneficial to humans. We have shown a need for a better substitute for a window that was documented decades ago. We have also shown how to construct an artificial window with motion parallax. Finally, we have performed three different evaluations to show that it is a better window substitute than a static image, with stronger effect on human well-being, and a good simulation of a window interface. Our evaluation was a small scale preliminary study with encouraging results. It clearly warrants further research, most notably larger sample size and varying order of conditions, which was not done due to the small number of participants and experiment setup. In our test subjects' opinion, it is clearly a step towards a real window of the same magnitude as the step from nothing to a picture on the wall.

Our prototype did surprisingly well in the evaluation, given that it is far from a perfect window simulation. In the current state, it has some minor problems which could easily be resolved. The current design with a better implementation could be used in some real world windowless settings. Due to the limitation of one viewer at a time, possible applications would be to single person windowless spaces, such as an office or a medical treatment space. In the future, artificial window video with stereopsis would be an excellent replacement for a window in all single-person spaces with a limited view of nature, such as underground, underwater, outer space, or just strictly urban areas. One

could also control window views: perhaps a mountain meadow one day, park with flowers another or a tropical island yet another. Even for people with a view of nature, on a rainy day, an artificial window could show a sunny view, or for people living close to the poles it could simulate normal day and night. The possibilities are endless.

5.2 Future Work

For future work, the simplest step would be minor changes in the implementation. These include elimination of user initialization through a good feature-based tracking algorithm and software camera control, elimination of pose limitation through use of two cameras, and faster camera movement update with a better camera. Changes in the system design are also possible. Currently the scene is designed as static. It is possible to use a panoramic movie approach which is similar to taking different environment maps over time from frames in a movie. Of course, a single movie in a high enough resolution would not be playable in real time, but instead the environment map can be divided into smaller pieces which are encoded as separate movie streams. Depending on the visible region of the scene only some of the streams are actually played [50]. In the artificial window case, the combined scene would have to be split into streams in a mip-mapping fashion, because when the viewer is close to the display, almost the entire scene is visible but only in the display resolution. To record the scene, a hemispherical polydioptric camera [40] would have to be used. With new blu-ray DVD discs [4] 24 hours of artificial window video should fit comfortably on a single disc.

Finally, there are things that are beyond our reach at this time, because they depend on new display technology. Adding stereopsis would be possible, but currently these

displays have a very limited viewing angle. If this improves, the primary assumption of one viewer at a time would be left as the only limitation of the simulated window. In the future, true 3D displays could remove this last limitation, and at this time lenticular displays seem like the most promising technology [43].

REFERENCES

1. Adelson E. H. & Bergen J. R. (1991). The plenoptic function and the elements of early vision. In Michael Landy, J. Anthony Movshon. (Eds.) *Computational Models of Visual Processing*. MIT Press: Cambridge, MA.
2. Arthur, K.W., Booth, K.S. & Ware, C. (1993). Evaluating 3D task performance for fish tank virtual worlds. *ACM Transactions on Information Systems (TOIS)*, 11 (3), 239-265.
3. Biner, P. M., Butler, D. L., Lovegrove, T. E., & Burns, R. L. (1993). Windowless in the workplace: A reexamination of the compensation hypothesis. *Environment and Behavior*, 25 (2), 205-227.
4. Blu-ray Disc Founders, http://www.blu-raydisc-official.org/ Accessed: Oct 2004.
5. Bourke, P. OpenGL sphere with texture coordinates, http://astronomy.swin.edu.au/~pbourke/opengl/sphere/ Accessed: Oct 2004.
6. Bradski, G. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (2), http://developer.intel.com/technology/itj/q21998/articles/art_2.htm Accessed: Oct 2004.
7. Chen, S.E. (1995). QuickTime VR: an image-based approach to virtual environment navigation. *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 29-38.

8. Chen, S.E., Williams, L. (1993). View interpolation for image synthesis. *Computer Graphics Proceedings, Annual Conference Series*, 279-288.
9. Chen, Y.S., Su, C.H., Chen, J.H., Chen, C.S., Hung, Y.P. & Fuh, C.S. (2001). Video-based eye tracking for autostereoscopic displays, *Optical Engineering*, 40 (12), 2726--2734.
10. Comaniciu, D. & Ramesh, V. (2000). Robust detection and tracking of human faces with an active camera. *IEEE Int. Workshop on Visual Surveillance*, 11-18.
11. Davis, J. & Chen, X. (2003). Calibrating pan-tilt cameras in wide-area surveillance networks. *IEEE International Conference on Computer Vision*.
12. Deering, M. (1992). High resolution virtual reality. *ACM SIGGRAPH Computer Graphics*, 26 (2), 195-202.
13. Dersch, H. Panorama Tools, http://www.path.unimelb.edu.au/~dersch/ Accessed: Oct 2004.
14. Distributed Systems Department at Lawrence Berkeley National Laboratory, Remote Camera and Videoswitcher Control Software, http://www-itg.lbl.gov/mbone/devserv/ Accessed: Oct 2004.
15. Djajadiningrat, J.P. & Gribnau, M.W. (1998). Desktop VR using QuickDraw 3D, Part I. *MacTech*, 14 (7), 32-43.
16. Dodgson, N.A., Moore, J.R., Lang, S.R., Martin, G.J. & Canepa, P.M. (2000). 50-in. time-multiplexed autostereoscopic display. *Proceedings of SPIE*, 3957A, 177-183.

17. Durgin, F.H., Proffitt, D.R., Olson, T.J. & Reinke, K.S. (1995). Comparing depth from motion with depth from binocular disparity. *Journal of Experimental Psychology: Human Perception and Performance*, 21 (3), 679-699.

18. Farley, K. M. J. & Veitch, J. A. (2001). A room with a view: a review of the effects of windows on work and well-being, *Research Report, Institute for Research in Construction, National Research Council Canada*, IRC-RR-136.

19. Finnegan, M. C. & Solomon, L. Z. (1981). Work attitudes in windowed vs. windowless environments. *The Journal of Social Psychology*, 115, 291-292.

20. Friedman, B., Freier, N. G., & Kahn, P. H. Jr. (2004). Office window of the future? Two case studies of an augmented window. *CHI Extended Abstracts 2004*, 1559.

21. Gorodnichy, D.O., Malik S. & Roth, G. (2002). Affordable 3D face tracking using projective vision. *Proceedings of International Conference on Vision Interface (VI'2002)*, 383-390.

22. Gortler, S.J., Grzeszczuk, R., Szeliski, R. & Cohen, M.F. (1996). The lumigraph. *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 43-54.

23. Gribnau, M.W. & Djajadiningrat, J.P. (1998). Desktop VR using QuickDraw 3D, Part II. *MacTech*, 14 (8), 26-34.

24. Halle, M. (1997). Autostereoscopic displays and computer graphics. ACM SIGGRAPH, 31 (2), 58-62.

25. Heerwagen, J. H. & Orians, G. H. (1986). Adaptations to windowlessness: A study of the use of visual décor in windowed and windowless offices. *Environment and Behavior*, 18 (5), 623-639.

26. Hubel, D. (1988). *Eye, Brain, and Vision*. Scientific American Library: New York.

27. Kaplan, R. (1993). The role of nature in the context of the workplace. *Landscape and Urban Planning*, 26, 193-201.

28. Kaplan, S. (1995). The restorative benefits of nature: Toward an integrative framework. *The Journal of Environmental Psychology*, 15, 169-182.

29. Küller, R. & Wetterberg, L. (1996). The subterranean work environment: Impact on well-being and health. *Environment International*, 22 (1), 33-52.

30. Levoy, M. & Andhanrahan, P. (1996). Light-field rendering. *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 31-42.

31. Lienhart, R., Kuranov, A. & Pisarevsky, V. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. *Proc 25th Pattern Recognition Symposium*, 297-304.

32. Markus, T. A. (1967). The function of windows: A reappraisal. *Building Science*, 2, 97-121.

33. McMillan, L., Bishop, G. (1995). Plenoptic modeling: an image-based rendering system. *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 39-46.

34. Mehrabian, A. & Russell, J. A. (1974). *An approach to environmental psychology.* MIT Press: Cambridge, MA.

35. Moreau, J.P. SIGNAL PROCESSING IN C/C++, http://perso.wanadoo.fr/jean-pierre.moreau/c_signal.html Accessed: Oct 2004.

36. MSDN Library, DirectShow, http://msdn.microsoft.com/library/en-us/directshow/htm/directshow.asp Accessed: Oct 2004.

37. Nagy, Edit (1998). *Working in Underground Offices*, Almqvist & Wiksell International: Stockholm, Sweden.

38. Nakanishi, Y., Fujii, T., Kitajima, K., Sato, Y. & Koike, H. (2002). Vision-based face tracking system for large displays. *Proc. Ubicomp 2002*, 152-159.

39. Neumann, J. & Fermüller, C. (2003). Plenoptic video geometry. *Visual Computer*, 19 (6), 395-404.

40. Neumann, J., Fermüller, C. & Aloi-monos, Y. (2002). A hierarchy of cameras for 3d photography. *In 1st Symposium on 3D Processing, Visualization, and Processing (3DPVT)*, 2-11.

41. Niemann, T. Panorama Tools Tutorial for Printable Panoramas, http://epaperpress.com/pano/download/PrintablePanoramas.pdf Accessed: Oct 2004.

42. Norris, D. & Tillett, L. (1997). Daylight and productivity: Is there a causal link? *Glass Processing Days Proceedings 1997*, 213-218.

43. Okano, F., Arai, J., Hoshino, H. & Yuyama, I. (1999). Three-dimensional video system based on integral photography, *Opt. Eng.*, 38, 1072-1077.

44. Oliver, N., Pentland, A. & Bérard, F. (1997). LAFTER: Lips and Face Real-Time Tracker. *Proc. CVPR '97*, 123-129.

45. Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary/ Accessed: Oct 2004.

46. Otsuka, T. & Ohya, J. (1998). Scene rendering method to affect motion parallax due to head movements. *Proc. of 8th International Conference on Artificial Reality and Tele-existence*, 117-121.

47. Panoguide, http://www.panoguide.com Accessed: Oct 2004.

48. Parsons, R. & Hartig, T. (2000). Environmental psychophysiology. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Handbook of Psychophysiology (2nd ed.)* Cambridge University Press: New York, 815-846.

49. Peleg, S. & Herman, J. (1997). Panoramic mosaics by manifold projection. *Proceedings of Computer Vision and Pattern Recognition*, 338-343.

50. Pintaric, T., Neumann, U. & Rizzo, A. (2000). Immersive panoramic video. *Proceedings of the 8th ACM International Conference on Multimedia*, 493-494.

51. PTGui, http://www.ptgui.com Accessed: Oct 2004.

52. Redert, A., van Klaveren, J.J. & Hendriks, E.A. (1997). Accurate 3D eye tracking for multi viewpoint systems. *Proceedings IWSNHC3DI*, 224-228.

53. Schröcker, G. (2000). Seamless tiling in OpenGL based Quicktime VR applications. *Web Proceedings of the 4th Central European Seminar on Computer Graphics*.

54. SIGVIEW, http://www.sigview.com/ Accessed: Oct 2004.

55. Smaller Animals Software JpegFile package, http://www.smalleranimals.com/jpegfile.htm Accessed: Oct 2004.

56. Smith, S. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing: San Diego, CA.

57. Sony Corporation, EVI-D30/D31 Command List, http://www.sony.net/Products/ISP/pdf/commandlist/CLEVID30E.pdf Accessed: Oct 2004.

58. Spielberger, C. D., Gorsuch, R. L. & R. E. Lushene (1970). *Manual for the STAI (State-Trait Anxiety Inventory).* Consulting Psychologists Press: Palo Alto, CA.

59. StereoGraphics Corporation, The SynthaGram Handbook, http://www.stereographics.com/products/synthagram/The_SynthaGram_Handbook_v8.pdf Accessed: Oct 2004.

60. Stevenson, A. (2002). Calibrating Head-Coupled Virtual Reality Systems. Masters Thesis, University of British Columbia, Vancouver.

61. Szeliski, R. & Shum, H.Y. (1997). Creating full view panoramic image mosaics and environment maps. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 251-258.

62. Trajkovic, M. (2002). Interactive Calibration of a Pan-Tilt-Zoom (PTZ) Camera for Surveillance Applications, *Proc. Asian Conference Computer Vision*, http://www.ee.usyd.edu.au/~miroslav/recentp/AutoCalibration.pdf Accessed: Oct 2004.

63. Trajkovic, M. (2002). Intruder Tracking with an Active Camera, *Proc. Asian Conference Computer Vision*, http://www.ee.usyd.edu.au/~miroslav/recentp/PTZTracking.pdf Accessed: Oct 2004.

64. Turner, Missy (2001). 'Virtual windows' brighten rooms without a view, *Houston Business Journal*, September 24.

65. Ulrich, R. S. & Gilpin, L. (2003). Healing arts: nutrition for the soul. In Frampton, S., Gilpin, L. and Charmel, P. (Eds.) *Putting Patients First: Designing and Practicing Patient-Centered Care*. Jossey-Bass: San Francisco, 117-146.

66. Ulrich, R. S. (1984). View through a window may influence recovery from surgery. *Science*, 224, 420-421.

67. Ulrich, R. S. (1993). Biophilia, biophobia, and natural landscapes. In Stephen R. Kellert and Edward O. Wilson (Eds.) *The Biophilia Hypothesis*. Island Press: Washington, DC, 73-137.

68. Ulrich, R. S., Simons, R. F., Losito, B. D., Fiorito, E., Miles, M. A., & Zelson, M. (1991). Stress recovery during exposure to natural and urban environments. *Journal of Environmental Psychology*, 11, 201-230.

69. Viola, P.A., & Jones, M.J. (2001). Rapid object detection using a boosted cascade of simple features. *CVPR*, (1), 511-518.

70. Ware, C., Arthur, K. & Booth, K.S. (1993). Fish tank virtual reality. Proc SIGCHI conference on Human Factors in Computing Systems, 37-42.

71. Wyon, D. P. (2000). Individual control at each workplace: the means and the potential benefits. In D. Clements-Croome (Ed.) *Creating the Productive Workplace*. E & FN Spon: New York, 192-206.

VITA

NAME: Adrijan Silvester Radikovic

PERMANENT ADDRESS:

KRALJEVEC 48A

10000 ZAGREB

CROATIA

EDUCATIONAL BACKGROUND:

M.S. Computer Science

Texas A&M University

College Station, TX

Dec 2004

B.S. Computer Science

Texas A&M University

College Station, TX

Dec 2002

AWARDS:

Graduate Merit Fellowship