Scalable Multi-view Stereo Camera Array for Real World Real-Time Image Capture and Three-Dimensional Displays

Samuel L. Hill

B.S. Imaging and Photographic Technology
Rochester Institute of Technology, 2000

M.S. Optical Sciences
University of Arizona, 2002

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Media Arts and Sciences

at the

Massachusetts Institute of Technology

June 2004

Signature of Author: _____
Samuel L. Hill
Program in Media Arts and Sciences
May 2004

Certified by: _____
Dr. V. Michael Bove Jr.
Principal Research Scientist
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by: _____
Andrew Lippman
Chairperson
Department Committee on Graduate Students
Program in Media Arts and Sciences

Scalable Multi-view Stereo Camera Array for Real World Real-Time Image Capture and Three-Dimensional Displays

Samuel L. Hill

Submitted to the Program in Media Arts and Sciences School of Architecture and Planning on May 7, 2004
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Media Arts and Sciences

Abstract

The number of three-dimensional displays available is escalating and yet the capturing devices for multiple view content are focused on either single camera precision rigs that are limited to stationary objects or the use of synthetically created animations. In this work we will use the existence of inexpensive digital CMOS cameras to explore a multi-image capture paradigm and the gathering of real world real-time data of active and static scenes. The capturing system can be developed and employed for a wide range of applications such as portrait-based images for multi-view facial recognition systems, hypostereo surgical training systems, and stereo surveillance by unmanned aerial vehicles. The system will be adaptable to capturing the correct stereo views based on the environmental scene and the desired three-dimensional display. Several issues explored by the system will include image calibration, geometric correction, the possibility of object tracking, and transfer of the array technology into other image capturing systems. These features provide the user more freedom to interact with their specific 3-D content while allowing the computer to take on the difficult role of stereoscopic cinematographer.

Thesis Supervisor: V. Michael Bove Jr.
Title: Principal Research Scientist of Media Laboratory

Scalable Multi-view Stereo Camera Array for Real World Real-Time Image Capture and Three-Dimensional Displays

Samuel L. Hill

Accepted by: _____

Dr. Michael Halle
MIT Visiting Scholar
Brigham and Women's Hospital – Surgical Planning Laboratory

Accepted by: _____

Dr. Wendy Plesniak
MIT Research Affiliate
Research Fellow – Harvard Center of Neurodegeneration and Repair

Acknowledgments

I dedicate this work to Professors Dr. Stephen Benton. I would also like to extend my gratefulness to my advisor Dr. Bove and my other two thesis readers Dr. Michael Halle and Dr. Wendy Plesniak, all of whom I have relied heavily on in the process of completing my degree. Thank you all. I would also like to acknowledge the support of my sister Jennifer Hill, my brother-in-law John Desisto, along with my parents. This work would not be possible without the tremendous support of Steven Smith, Jacky Mallet, and Tyeler Quentmeyer. Lastly I would like to thank the Spatial Imaging Group community for their warm support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 Introduction

The human visual system is quite adept at perceiving depth and making spatial connections between objects in an environment. Many visual cues help define our sense of depth, where the differences and similarities between the two images captured by each eye's retina are compared and fused in the brain.

A variety of three-dimensional displays exist to take advantage of this underlying property of images captured from disparate positions. In these systems, disparity is used to create the perception of depth from two-dimensional media. Providing three-dimensional visualization over traditional two-dimensional representation offers great promise for fields such as entertainment, medicine, surveillance, teaching, design, robotics; just about anywhere spatial awareness is needed

Acquiring image data for these applications and for stereo or multi-view stereo displays requires sophisticated capturing devices. Several stereo camera mechanisms exist that can capture active or non-stationary object scenes. Yet, multi-view stereo devices tend to be regulated to capture static image scenes because such systems require precision camera movements and high capture bandwidth to record all the images at once. Dynamic multi-view capture is thus typically limited to computer-generated imagery, where these physical limitations are non-existent.

This thesis describes the Spatial Imaging Scalable Camera Array (SISCA), a prototype system for active image capture content. SISCA was built using off the shelf components including six USB cameras, vertical and horizontal servos, and a custom software GUI. The system is scalable and easily adjustable to the particular display, image environment, and the overall application. It can capture multiple stereo views for single viewpoint stereo displays, giving the viewer a limited ability to look around a scene from different viewpoints, allowing flexibility to the user. What distinguishes this system from a single precision camera is the ability to capture one moment in time from many spatial locations; permitting moving objects in natural scenes to be recorded. SISCA is based on a highly configurable and scalable camera geometry and software system to encourage experimentation in different applications.

1.2 Motivation

1.2.1 Previous Work

Over the past twenty years several graduate students within the MIT Media Laboratory Spatial Imaging Group have made major advances in three-dimensional display technology such as holographic video (St.-Hiliare, 1994) and most recently a sixteen view-sequential display (Cossairt, 2003) utilizing advanced stereogram rendering techniques (Halle, 1994). With real world real-time image capture the benefit of these displays can be utilized amongst a wider field of applications beyond those that already exist.

1.2.2 Multiple-camera Paradigm

Another motivation of a real-time image capture system for these 3D displays is the ability to use multiple digital cameras rather than a precision moving camera rig because of the rapidly declining costs of CMOS and CCD image sensors. At the same time the performance and accuracy of a multiple camera array was shown in recent experiments using one computer for connection and processing in real time. These advances make it beneficial to acquire active scene acquisition and to look at a much larger image capture paradigm using the inexpensive high quality digital imaging cameras on the market.

Figure 1.1 shows an example of a typical multi-perspective image capture and display system. A collection of small video cameras is arranged in a horizontal array, with each camera positioned immediately next to its neighbors. In front of the camera array is the scene the viewer would like to capture, with a central object a few meters away and a background object several meters further back. The output of the camera array is presented to the viewer using a head mounted stereo goggle system. The system should present to the viewer accurate spatial and depth relationships about the scene, just as if the viewer were looking at the scene directly.

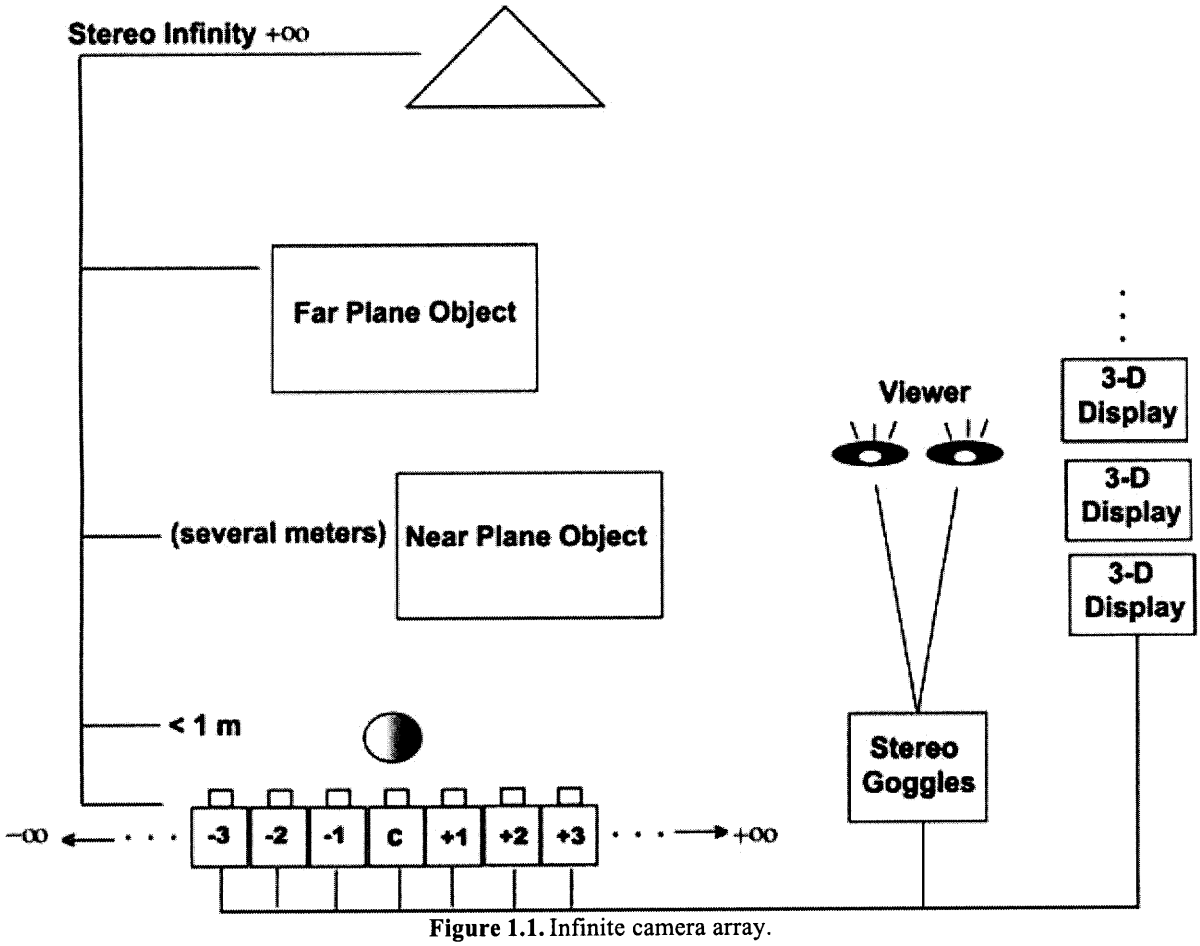


Figure 1.1. Infinite camera array.

With Figure 1.1 in consideration, problems that would be ideally solved by such a system are:

- The system, not the viewer, goes about making the correct representation from world space of the central object and background object to the image plane in two separate cameras to the plane of the stereo goggles in a manner that is pleasing and causes very little strain to the viewer.
- The system allows the user to accentuate features and still provide acceptable images to the eye.
- The user has the ability to transform the system if the 3-D display changes (plug and play).
- The user has the ability to focus and acquire scene depth from objects that appear extremely far away (the triangle represented by stereoscopic infinity).
- The user has the ability to focus and create macro depth from multiple perspectives of an object that appears extremely close to the camera array (the shaded circle).
- The system has the ability to track objects of interest as they move in the scene.

A variety of parameters and sometimes conflicting goals must be balanced when creating three-dimensional imagery of moving objects in real world environments. The goal of the SISCA is to provide a testbed to help understand these parameters and find reasonable solutions for these goals. The thesis will present the paradigm of an infinite camera array that can autonomously choose the best cameras to make up the scene based on the viewer's needs, the viewer's display, and the content that the viewer is looking for. From these parameters the system will choose which views to take from the infinite array. Essentially, the system becomes scalable when the infinite camera array is reduced to a finite number of cameras that can rapidly position themselves based on the information they receive from the program architecture of the entire array.

The SISCA is a prototype implementation of this ideal system based on six USB cameras running in parallel with a restricted user interface GUI. The system has the ability to capture images in a pseudo real-time manner and can correct image distortions for proper output on a display. For now, the user only has the ability to select and record images with a view-sequential or stereo hound mounted 3D display. Also, the user has little artistic control over the system.

Although the SISCA is still in early development, the system reveals promise for answering the higher-level concepts of the multi-camera paradigm while new and improved features are continually added to the camera array and user interface. In the long term the multi capture array can work with a semi-autonomous software platform with higher precision, faster capture/display rates, and can be applied to a variety of different imaging systems. The remaining chapters will delve further into the current and future capabilities of the SISCA.

1.3 Target Applications

An important feature of the camera array is the ability to scale its functionality across extremely varying environments where potential visual attributes of the scene, including physical distances of objects from the camera, differ between viewers. I specifically chose three important applications that entail a variety of parameters in world coordinates, camera positions, the 3D display device and the specific use/content needed by the viewer. The three applications are:

- Aerial Vehicles
- Facial Recognition
- Surgical Applications

1.3.1 Aerial Vehicles

The first field of study is acquiring depth views for hyperstereo navigation and surveillance for unmanned aerial vehicles (UAV), which require controllable distant eyes in dangerous environments. UAVs are remotely piloted or self-piloted aircraft that can carry cameras, sensors, communications equipment or other payloads. They have been used in a reconnaissance and intelligence-gathering role since the 1950s and are now being included in combat missions (Bone, 2003).

Several aerospace companies build UAVs for military purposes, including Lockheed Martin, BHTI, and Teledyne Ryan Aeronautical. Each unique UAV design is closely tailored towards mission requirements and overall flying range. The lightest UAVs can handle close range (within 50 km) missions and short range (within 200 km) missions while larger models can handle endurance missions much further in distance (NASA, 2004).

UAVs could use a drastic overhaul in imaging capabilities especially since they present serious hazards while flying in airspace with other aircraft. First, most UAVs have imaging systems that have a very narrow field of view, which only affords a "soda straw" perspective to the UAV operator. This poor visual perspective diminishes most chances of seeing and avoiding other airborne platforms (Glausier, 1992). Providing a stereo view to the operator is advantageous for a much larger degree of visual and spatial awareness. This is important for avoiding collisions while also potentially allowing for a much higher mission success rate because of greater precision image capture, target location and analysis, and flight maneuvering.

Providing stereo capabilities on a UAV requires that the size and weight of the cameras are minimized to fit within strict weight restrictions. The other crucial factor is the position of the cameras. Most likely, due to high altitudes, the camera array has to be dispersed across the wingspan of the vehicle. An operator depending on the circumstance would use one of the views for navigation or use a head mounted stereo display for navigation. Additional personnel looking at scene content may use some other three-dimensional display like a multi-view sequential display.

1.3.2 Facial Recognition

The second major example was for a mid-level scene, which means that objects of interest are only a few meters from the camera array and the separation between two cameras is less than one

meter. A typical application for a mid-level scene would be precision facial recognition for security and possibly entertainment purposes.

Face recognition systems are highly sensitive to the environmental circumstances under which the images being compared are captured. In particular, changes in lighting conditions can increase both false rejection rates (FRR) and false acceptance rates (FAR) (Heselstine, 2003). Another problem that arises is facial orientation and angle of image capture. When standard 2D intensity images are used, in order to achieve low error rates, it is necessary to maintain a consistent facial orientation, preferably a frontal parallel perspective. Even small changes in facial orientation can reduce the effectiveness of the system (Kroeker, 2002).

This situation is aggravated by the fact that people's facial expressions can change from one image to another, increasing the chance of a false rejection. In order to produce secure site access systems makes it necessary to specify a required facial expression, usually neutral. However, this approach then removes one of the key advantages of face recognition, which is not having subject cooperation, thus making it unsuitable for surveillance applications (Ezzat, 1996).

Stereo image capture provides a larger number of views taken at a multitude of angles making it easier to retrieve a larger array of marker points for facial pattern searching. Stereo images can also be used to create three-dimensional facial surface models to predict facial changes, facial orientation, and possibly add aging effects. Using these techniques also eliminates lighting effects commonly associated with high FFR along with 3-D facial capture systems that use laser scanning equipment. Essentially the system can capture multiple snapshots quickly and undetectable.

1.3.3 Surgical Applications

In the third case I'm interested in hypostereo (for macro close-ups) imaging for recording surgical procedures. This area is important for both teaching procedures and for telemedicine applications. Currently the main forms of three-dimensional visualization in medicine are 3D imaging devices in radiology that ranges from ultrasound, X-ray, nuclear imaging, computed tomography (CT) scanning, and magnetic resonance imaging (MRI). The other form of 3D visualization is taking either 3D or 2D data and creating virtual 3D models. All of these systems are hard to implement in real-time surgical applications along with high volumes of data that would need to be stored for recording purposes.

An area in medicine that could use small real-time three-dimensional imaging is telemedicine programs that use telerobotics to facilitate surgeries between a physician and a patient who may be remotely far away, have limited access to proper facilities or in a dangerous environment. Many physicians rely on the imaging capabilities and the use of peripheral devices such as electronic stethoscopes, otoscopes, and opthalmoscopes to make key decisions for off-site physical examinations (*Telemedicine*, 2001).

However, studies have shown that physicians feel handicapped in operating current medical robots and their surgical tools because of flat 2D images coming from the cameras. The doctor must truly have some way to gauge the robot's physical presence in relation to the patient in three-dimensions. Providing spatial information along with a wide degree of viewing is a must

have for proper diagnosis and surgical procedures (Thompson, 2001). This is true also for teaching medical students proper techniques.

Multiple stereo-views allow for this greater degree of spatial connectedness, awareness, control, and learning to students, physicians, and telerobotic operators. Additionally, a system such as the SISCA would provide students/observers to manipulate the capture streams from multiple viewpoints. Lastly, a stereo camera array takes up considerably less bandwidth and storage memory to display and record surgical operations compared to all the other sophisticated 3D medical imaging devices.

1.4 Common Issues

All three applications bring rise to issues of geometric distortion and camera calibration. A reliable way to reduce these geometric distortions involves detaching the lenses from the image sensor. The result is that all the cameras maintain a parallel projection to the scene and can maintain the object of interest centered on each image sensor. This type of arrangement does introduce a shearing effect to cameras off center with the object but is easily corrected for with image processing.

With real-time capture systems most scenes will probably have moving objects. This entails that most of the cameras will not be parallel to the scene and requires the lenses to have a wide field of view for tracking objects. Tracking involves placing the cameras on horizontal and vertical servos with the ability to pivot towards the object for re-centering and even look around (pitch/yaw). This also mean the lens are attached to the image sensors and that cameras further away from the center camera are toed-in and converging on the object of interest.

Camera convergence foreshortens the perspective projection of the object in image space and creates vertical disparity, also known as keystoning. The disparity changes much like a zoom lens changes magnification, as an individual camera is further away from the central camera. It's important to observe a suitable disparity budget for each of the applications depending on the scene being viewed and the three-dimensional display being used. The final stage of the system is the output of the recorded scenes from the cameras and requires suitable keystone correction depending on which camera view it came from. At the same time, other geometric distortions from radial lens distortion can be corrected before final display.

Another issue of importance is the calibration of the cameras and initial alignment with each other. To synchronize locations between all the cameras for accuracy and reducing distortions involves an initial calibration that provides the camera array system with the intrinsic and extrinsic parameters of all the cameras. Camera calibration is performed in many different ways across a multitude of disciplines but for our case we'll use a standardized method introduced by Zhang (2002) that uses a checkerboard target and non-linear optimization.

The image processing schemes implemented with the camera array are vital in providing corrected image scene output by adjusting the disparity between cameras, aligning cameras, and reversing geometric distortions. It is essential to quantify these issues to optimize all the image processing and for further comparison of the toed-in camera array versus a shearing camera array.

1.5 Goals and Contributions

The main objective of this thesis is to design and test a multi-view stereo capture system for acquiring real-time real-world scenes. The contributions of this thesis include:

- A summary of relevant 3-D capturing systems and 3-D displays.
- Design and implementation of a multi-view camera array. This includes a GUI for viewing the camera output and a limited user interface for setting parameters, calibrating the cameras, and recording input for 3D display output.
- Analysis of multi-view camera capture geometry.
- Implementation of image-processing schemes to adjust image scene output for disparity, geometric correction, and calibration issues.
- Results from a set of experiments indicating the best performance of the multi-view camera array in conjunction with the image processing under several scene environments using either a two-view stereoscopic display or the view sequential display.
- Recommendations on future designs which include tracking user/computer specified objects in scenes, increasing the performance of recording and editing scenes for 3D display output, extending the camera array technology to other image capturing systems, and a more robust user interface for both camera control and camera calibration.

The remaining chapters will provide a breakdown of three-dimensional display and capture. Chapter 2 will discuss the underlying properties and problems in proper image processing through the stereoscopic value chain in both the human visual system and with imaging devices. Chapter 3 will provide a summary of various three-dimensional capture and display systems and how they work. Then in Chapter 4 the functionality, building, and performance of the SISCA is presented. Lastly, Chapter 5 will explore the future of the SISCA and the multi-image capture paradigm.

CHAPTER 2. THE STEREOSCOPIC VALUE CHAIN

This chapter will explore the way a viewer perceives depth and how this relates to the inherent subtleties of making a 3D capture-display system. A brief explanation of visual cues will be presented along with an overview of the physiology of the eye leading to the main causes of strain while viewing three-dimensional displays. Next, there will be an exploration of projective transformation across the various coordinate spaces, which will lead to inherent geometrical constraints and factors involved with image capture.

2.1 Physiology of the Eye

There are a number of depth cues that help people associate spatial relationships between various objects, both image based and real world based. Table 2.1 gives a list of these cues.

Table 2.1. Depth Cues

A. Perspective	Objects get smaller the further away they are and parallel lines converge in distance.
B. Size	We expect a known object to have a certain size and relationship to other objects. If we know box A is bigger than box B but they appear the same size then we would expect box A to be further away.
C. Detail	Close objects appear in more detail, distance objects less.
D. Occlusion	An object that blocks another is assumed to be in the foreground.
E. Lighting/Shadows	Closer objects are brighter, distant objects are dimmer. There a number of other more subtle cues implied by lighting, the way a curved surface reflects light suggests the rate of curvature, shadows are a form of occlusion.
F. Relative Motion	Objects further away seem to move more slowly than objects in the foreground.
Cues not Present in 2D images	
A. Binocular Disparity	This is the difference in the images projected onto the retina and then onto the visual cortex because the eyes are separated horizontally by the interocular distance.
B. Accommodation	This is the muscle tension needed to change the focal length of the eye lens in order to focus at a particular depth.
C. Convergence	This is the muscle tension required to rotate each eye so that it is facing the focal point.

Studies have shown that the dominant depth cue is occlusion. Binocular disparity is considered the second most dominant depth cue however this can be distorted by competing cues such as accommodation and convergence (Bulthoff, 1998).

2.1.1 Binocular Disparity

Binocular disparity is the positional difference between the two retinal projections of a given point in space. This positional difference results from the fact that the two eyes are laterally separated and therefore see the world from the two slightly different vantage points. For the average person the mean lateral separation also known as the interocular is 65mm. Most of the population has an eye separation within ±10mm of the average interocular.

When the two eyes converge at one spot and fixate, the light from this spot will stimulate corresponding points on both retinas. An imaginary ellipse called the horopter can be traced through the fixation point and the nodal points of both eyes. Any object that falls on the horopter

will also have corresponding points on both retinas. All other points inside or outside the horopter will have some difference from each other on the two retinas (Howard, 1995).

Further down the image processing stage of the human visual system lays the ability to correlate the points of both retinas with each other. It is this neurological process that allows a viewer to fuse the retinal images and perceive depth. Unfortunately, there is roughly 8% of the population that is stereo blind and relies on other depth cues (Julesz, 1977).

Our ability to see in stereo is also limited on a limiting distance for stereoscopic acuity. This is based on the stereo base and the resolving power of the system. For the eye, the stereo base is 65mm and the resolving power of the eye, which is approximately one minute of arc (Lipton, 1982).

$$D_x = \frac{t_e}{\tan \Delta\alpha} \tag{2.1}$$

2.1.2. Accommodation and Convergence

The eyes have both accommodation and convergence features. The muscles that focus or accommodate the eyes are controlled by neurological systems separate from those that converge the eyes. When the eyes converge for a given distance, there is a zone of single, clear binocular vision. Once we move out of the zone, or exceed the limits of the accommodation/convergence relationship, one of two things will happen. Either fusion breaks down in which case there is double vision and accommodation remains. Otherwise, fusion is maintained and accommodation is lost because the image is out of focus. The result to the viewer is that the depth perception will be exaggerated or reduced, the image will be uncomfortable to watch, the stereo pairs may not fuse at all, and the viewer will see two separate images (Howard, 1995).

Several factors of three-dimensional image formation by the human visual system have been described. The visual system is extremely advanced and good at simultaneously capturing and displaying a real time event. Imaging systems on the other hand have their own issues as the system tries to accurately relay capture data to display format. The following sections explore these issues.

2.2 Projective Transformation

To capture a real-scene which isn't computer generated involves several coordinate transformations. Firstly from X, Y, Z coordinates in object/camera space to X and Y positions on the two camera imaging sensors (CMOS), secondly from the two sets of CMOS coordinates to X and Y positions of the left and right images on the stereoscopic display, and thirdly to a set of X, Y, Z coordinates in image/viewer space. A breakdown of the coordinate transformation is shown below (Woods et al., 1993).

Object Space → Camera Coordinates→ Screen Coordinates → Image Space

The projective transformation from real world coordinates in 3D geometry to 2D camera coordinates can be accomplished by one of two methods. A parallel projection involves parallel projection lines while perspective projection incases non-parallel projection lines. Table 2.2 gives a brief listing of projections used to transform from 3D to 2D coordinates.

The type of projective transformation to use will be based on the scene capture arrangement of the cameras. Many authors such as Rule (1938) who believe there shouldn't be any divergence involved with stereo viewing will arrange all image sensors in a parallel fashion to each other and will thus use a parallel projection. In such a situation, to keep objects in the center of projection, Halle (1996) recommends a shearing method by displacing the imaging lens from the sensor as seen in Figure 2.1. The two images from both cameras would overlap but their general volume would be distorted as seen from an above position.

Table 2.2. Projective Transformations

CLASS	SUBCLASS
A. Perspective	i) 1,2,3 point (Number of axes intersecting the projection plane)
B. Parallel	i) Oblique (Projection lines are not perpendicular to the projection plane)
	1. Cavalier
	2. Cabinet
	3. Shear
	ii) Orthographic (Projection lines are perpendicular to the projection plane)
	1. Elevations
	2. Axonometric
	2a. Isometric

Another common arrangement is to organize the image sensors with a toe-in effect or crossed-lens effect, which can be represented by a perspective transformation. The problem with the toe-in can be seen in Figure 2.2, which can cause vertical disparity. The perspective projection and its inverse can be derived as functions of the viewing and translation parameters. Further tests outlined will show the degree of vertical disparity based on the degree of toe-in and translations. The inverse transformation applied to the original image can correct for the keystoning.

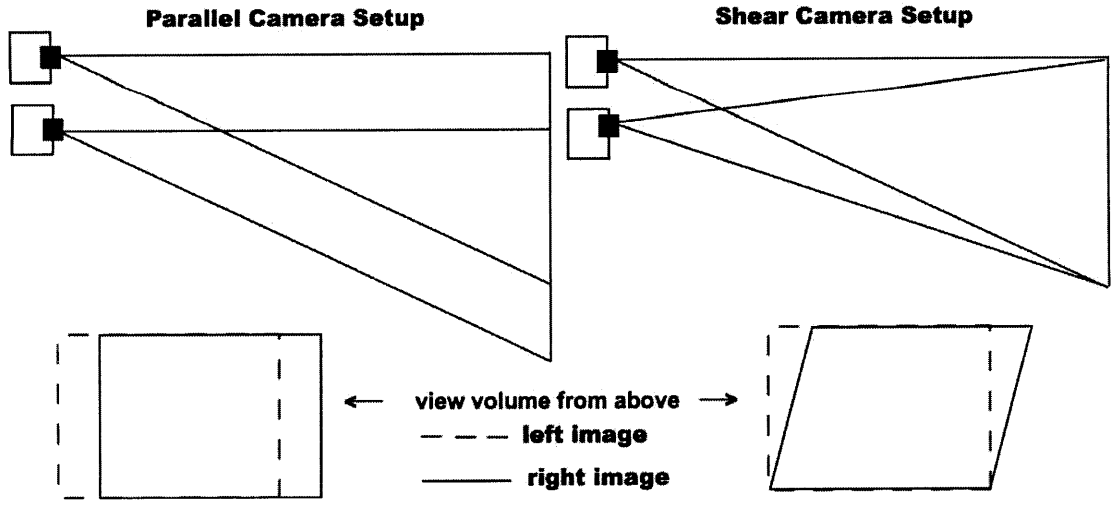


Figure 2.1. Parallel projection and camera sheer.

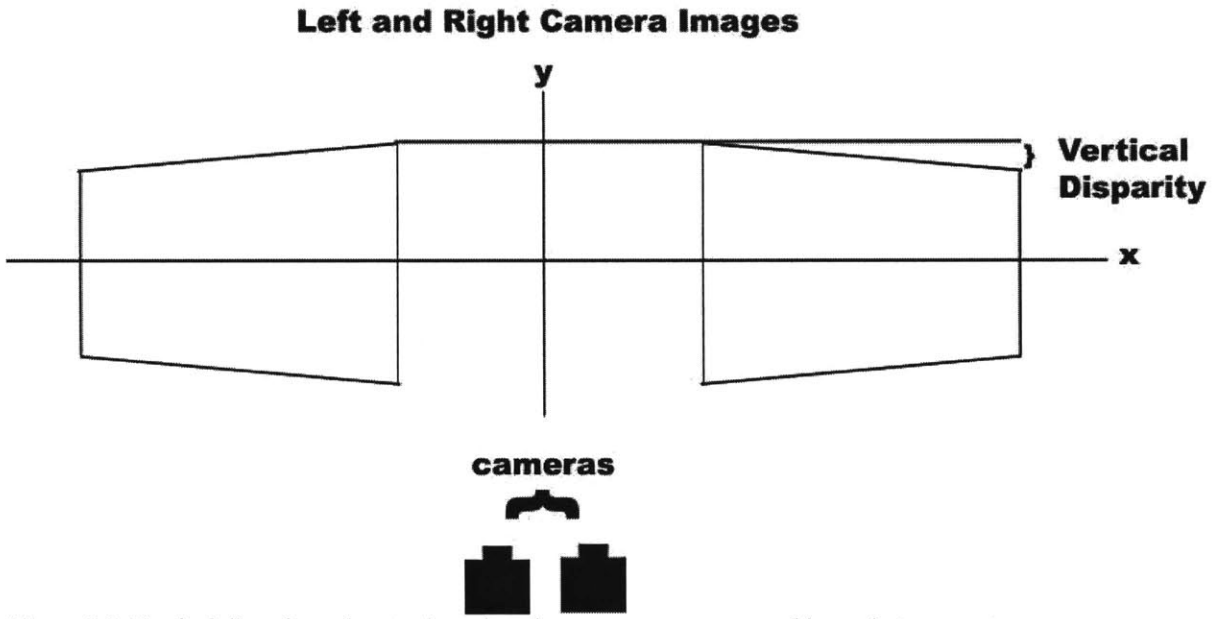
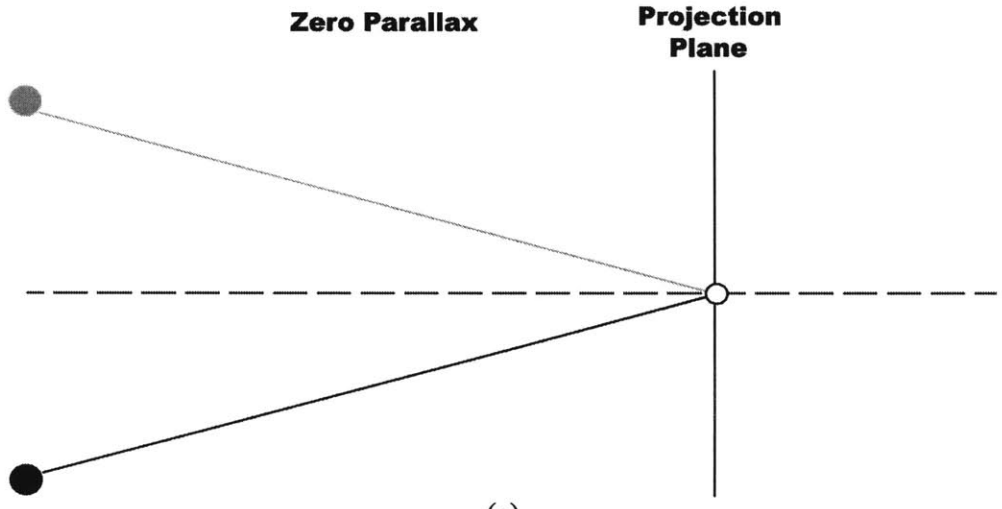


Figure 2.2. Vertical disparity or keystoning when the cameras are converged inward at one spot.

2.3 Display Parameters

2.3.1 Parallax

Parallax is the distance between corresponding left and right image points measured at the screen or image sensors. In Figure 2.3(a), the left and right image points have a screen parallax of zero because the left and right eyes converge on the plane of the screen for which they are focused. Figure 2.3(b) shows the case of positive (uncrossed) screen parallax in which the image points will appear behind the plane of the screen. Figure 2.3(c) shows image points that appear to be in front of the plane of the screen in theater space and have a negative screen parallax.



(a)

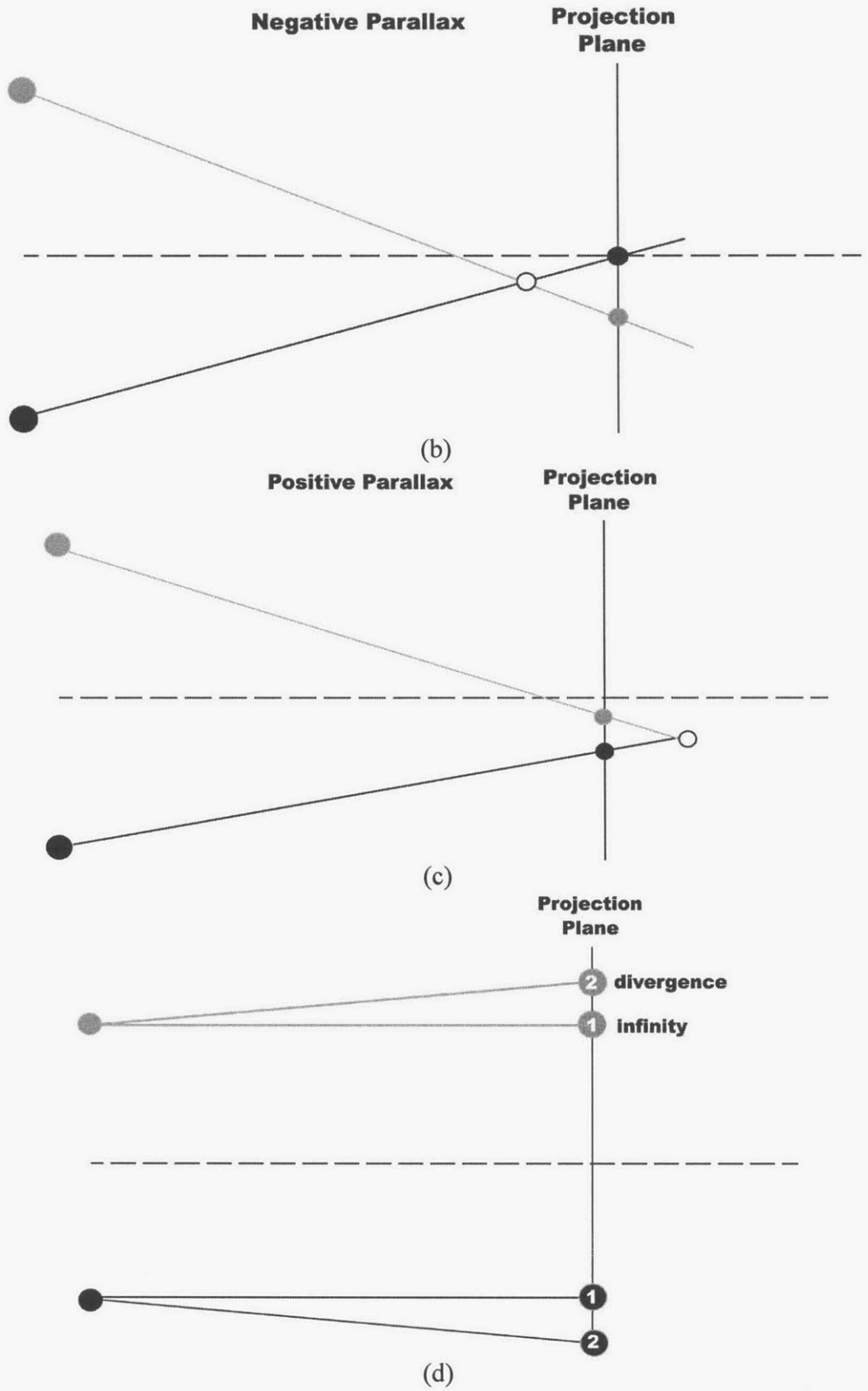


Figure 2.3. Parallax

Figure 2.3(d) shows corresponding left and right points, marked 1, equal to the interocular and considered to be at stereoscopic infinity. In this case the optical axes of the eye will be parallel when viewing the image points. This correlates with the human visual system when viewing very distant objects. For the market points labeled 2 in Figure 2.3(d) the image points are now farther apart and require the eye to diverge, or angle outward, in order to fuse such image points. If the divergence increases, some degree of fatigue and discomfort will occur to the viewer.

2.3.2 Divergence

Figure 2.4 gives a more general approach as discussed by Lipton in discussing homogenous image points. Zone 1 shows the region of large values of negative screen parallax, which makes it hard for fusion. Zone 2 is divided into parts *a* and *b*, with screen parallaxes between the two points between plus and minus the average interocular distance t_e. The negative zone *(a)* will produce images in theater space that are generally fusible to the viewer. The positive portion of the Zone, (*b*), is the region where nearly all photography should take place, and it lies in screen space. Lastly, there is the region of divergence in Zone 3, which is also divided into parts (*a*) and (*b*). In this zone the image separation will have screen parallax larger than t_e. Region (*a*) of Zone 3 will allow for divergence up to 1 degree. Several authors have proposed that 1 degree of divergence is still acceptable for fusion without strain. Region (*b*) will extend beyond the maximum acceptable divergence of 1 degree (Lipton, 1982).

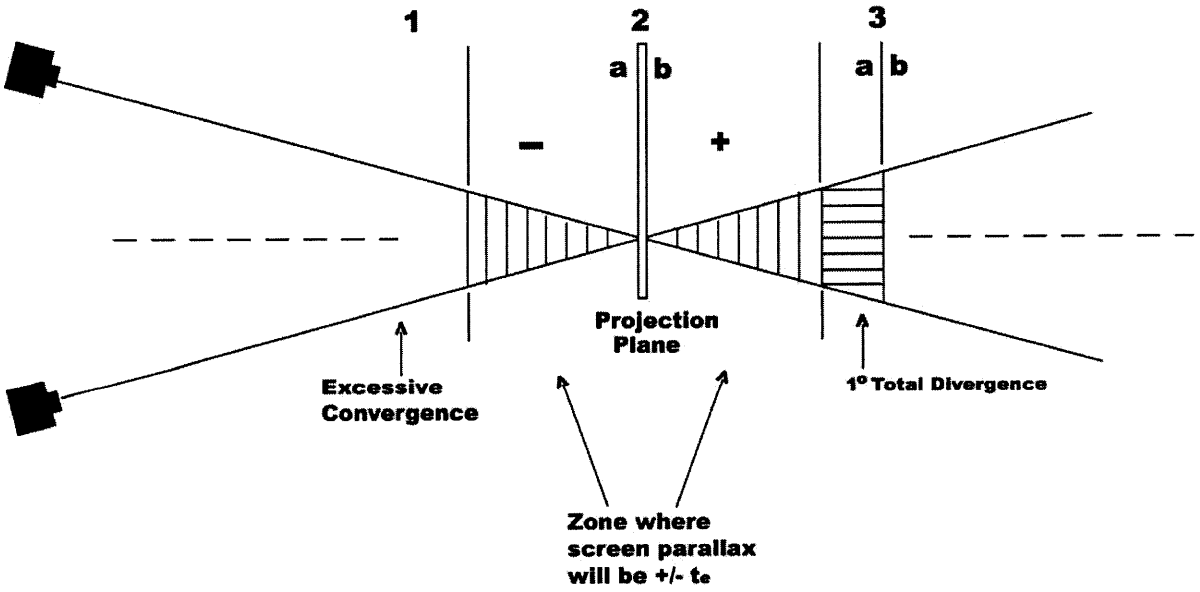


Figure 2.4. Divergence and parallax zones.

2.3.3 Magnification and Orthoscopy

Frame magnification is an important step in back calculating the stereo views needed by the cameras and also in the ability to maintain orthoscopy, which is the particular combination of shooting and projection variables that produce a stereoscopic image appearing exactly like the real object or scene at the time of photography. The frame magnification, as seen in equation 2.2

is a ratio of the screen height or width of the projected image to the frame height/width. The frame is the projection device used in the three-dimensional display, which in this study involved a DLP for a view sequential display and two LCDs in a stereo head-mounted display.

$$M = \frac{H_s}{H_f} = \frac{W_s}{W_f} \tag{2.2}$$

To achieve the orthoscopic condition would require that the subject imaged subtends the same angle in space, or covers the same portion of the retina during projection as it would for the observer at the scene.

$$V = Mf \tag{2.3}$$

The following expression states that the viewing distance during projection is proportional to the product of linear magnification and camera focal length. However, in an effort to achieve pleasing images, as in most photographic situations, requires following other paths to create the best images. For the realm of stereo-capture, adhering to orthoscopic capture may or may not happen.

2.4. Depth Range

Both Rule (1938) and Lipton (1982) give a basic depth range-equation where P_m is the maximum screen parallax; M is frame magnification, f_c the camera focal length, and t_c the interaxial. D_o and D_m, are the distance from the camera to the convergence plane and far plane respectively.

$$P_m = Mf_c t_c \left(\frac{1}{D_o} - \frac{1}{D_m} \right) \tag{2.4}$$

An object at distance D_o from the camera will have zero screen parallax while an object at D_m will have the maximum positive parallax. The goal is to keep the screen parallaxes as small as possible to offset the breakdown of accommodation/convergence while avoiding large parallax values resulting in excessive divergence.

Rule, Spottiswoode, and Norling, believed that divergence should not be allowed and that the maximum screen parallax between homogenous points of distant objects should be kept equal to or less than the interocular, t_e. Their argument is that subjects at photographic infinity, without important foreground compositional elements, look just as deep if the value of P_m is less than the average t_e.

However, MacAdam, Gunzburg, Hill, Levonian, Ovsjannikova and Slabova have preferred a total divergence of 1 degree allowing for acceptable fusion for a majority of people and extending the depth range. A leeway of 1 degree is greatly permissible when the background is darker in comparison to the foreground.

To accommodate for both cases P_m is transformed for the divergent case P_d and P_e for the non-divergent case. For the divergent case, d is the excess parallax of homologous points; V is the

distance from the viewer to the screen, and θ_d the angle of divergence. P_e is equal to t_e, which is 65mm.

$$d = \tan\theta_d * V \tag{2.5}$$

$$P_d = d + t_e \tag{2.6}$$

$$P_e = 65mm \tag{2.7}$$

P_d is also affected by the viewer in relation to the display screen. As the viewer recedes from the screen the angle of divergence decreases. Maximum screen parallaxes for divergent homologous points can appear to be very large, even several times t_e, but when analyzed from the point of view of angular measure, divergence can still be held to tolerable limits.

Now the stereoscopic constant K can be defined for both the divergent and non-divergent case using either P_d or P_e.

$$K = \frac{P_d}{M} \tag{2.8}$$

The basic depth-range equation now is transformed to the following manner and easily transformed to equations 2.9 and 2.10.

$$K = f_c t_c \left(\frac{1`}{D_0} - \frac{1}{D_m} \right) \tag{2.9}$$

$$\frac{1}{D_m} = \frac{1}{D_0} - \frac{K}{f_c t_c} \tag{2.10}$$

When the distance to the far plane D_m is at stereoscopic infinity, $1/ D_m$ equals zero, the following equation results.

$$\frac{1}{D_h} = \frac{K}{f_c t_c} \rightarrow D_h = \frac{f_c t_c}{K} \tag{2.11}$$

This relationship defines the hyperconvergence distance or the distance for which the lens axes must be converged in order to produce acceptable screen parallax for very distant objects. The following formula is very similar to the lens-maker equation found in geometrical optics.

$$\frac{1}{D_m} = \frac{1}{D_o} - \frac{1}{D_h} \tag{2.12}$$

Another important parameter is observing objects photographed in front of the plane of

convergence, with negative screen parallax and appearing in theater space. Large values of parallax, quite a bit greater than negative t_e, can be comfortably fused by many people. Rule and Spottiswoode have derived the following relationship where D_2 is the near plane.

$$\frac{1}{D_2} - \frac{1}{D_m} = \frac{2P}{Mf_c t_c} \tag{2.13}$$

Recalling that the hyperconvergence distance is defined by equation 2.13 and solving for $1/D_2$ we get equation 2.14.

$$\frac{1}{D_2} = \frac{2}{D_h} - \frac{1}{D_m} \tag{2.14}$$

Thus by solving for the near-plane distance in terms of the hyperconvergence distance and the far-plane distance, we have expressed what we call the near-range equation.

2.5 Keystoning

Rotating the cameras inward does not provide an orthostereoscopic reconstruction. Saunders (1968), Rule (1938), and Woods (1993) have shown the mathematical effect of translating the image sensor versus rotating them inward. A linear perspective causes foreshortening of objects as the distance from the viewer (camera) increases and is commonly called keystoning as seen in Figure 2.1. The vertical disparity or vertical parallax occurs when homologous points in a stereo pair do not lie on the same horizontal scan line. Lipton (1982) recommends that the vertical parallax between homologous points should be kept with 10 minutes of arc. Fender and Julesz (1977), testing with random-dot stereograms, found that experimental subjects could fuse two views with vertical parallax of six minutes of arc.

Vertical parallax V between a left and right image is equal to $y_l - y_r$. McAllister (1993), using perspective transformation, calculated y for the left and right images with the following equations.

$$V = \frac{2dx_0 y_0 \sin(\phi/2)}{[(z_0 - R)\cos(\phi/2) + R]^2 - x_0^2 \sin^2(\phi/2)} \tag{2.15}$$

Where y_l and y_r are the coordinates of point P_o, described by position (x_o,y_o,z_o) relative to both the left and right eyes respectively. The center of projection has coordinates (0,0,0), which can be considered the film plane, and d is the distance from the center of projection to the projection plane. For a camera we can assume that d is the focal length of the camera lens. R is the center of rotation that we can assume it is the nodal point of the camera. Lastly, $\phi/2$ is the half angle between the left and right eye views. This is shown in Figure 2.5.

Ariyaeeinia (1992) used a model with two rotated converging cameras where the x axes on the image planes become non-collinear. The vertical disparity is calculated from the angle between the two planes where φ is the rotation angle of the cameras from the parallel position and β is the angle between the optical axis and vertical object position.

$$\theta = \tan^{-1}(2\sin(\varphi/2)\tan\beta) \tag{2.16}$$

In the left image plane, if we rotate the x_1y_1 co-ordinate system through an angle ϕ, a new co-ordinate system, $x'_1y'_1$, is obtained whose horizontal axis is parallel with the right image. The relationship between the two co-ordinate systems can be defined by a rotation matrix that leads to the following vertical parallax equation.

$$y_1 - y_2 = 2(x_1 + x_2)\sin(\phi/2)\tan\beta \tag{2.17}$$

The above equation indicates that vertical disparity increases with the rotation angle of the cameras from the parallel position. The images of an object point have zero vertical parallax when $x_1 = -x_2$, where the image points are $y = 0$ such that $\beta = 0$, and lastly, when $\phi = 0$ because the axes of the cameras are parallel.

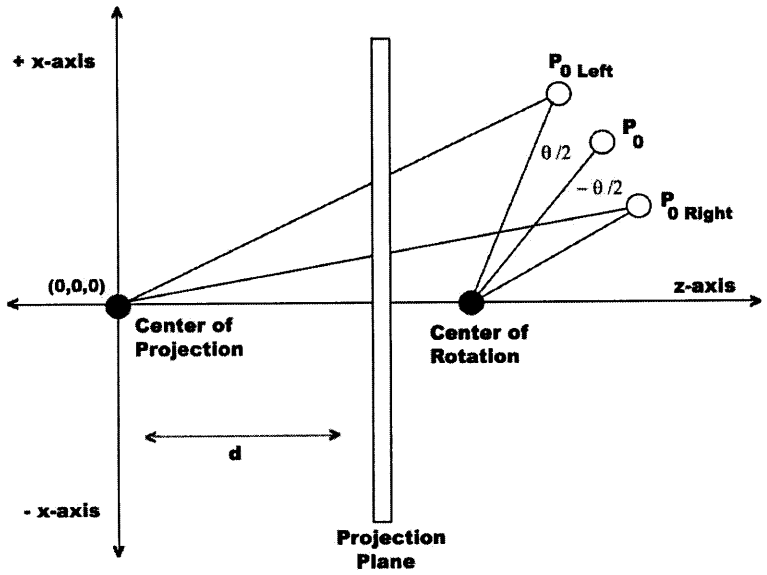


Figure 2.5. McAllister's description of keystoning in perspective projection.

2.5.1 Keystone Correction

Bourke (2000) does a general analysis of keystoning by describing the scaling effect induced by the perspective projection. There is a scaling in the x and y direction about the origins such that the modified coordinates x' and y' are described by the following equations. If S_x and S_y are not equal the result is a stretching along the axis of the larger scale factor.

$$x' = S_x x \tag{2.18}$$

$$y' = S_y y \tag{2.19}$$

McAllister (1994) derived a method to look at the distortion as a function of the distance e of the viewer to the screen, the amount $[u, v]$ of the translation of the center of the image I and the height h and width w of a point in I. Based on his system the maximum vertical disparity for a viewer is found by the following equation. For a majority of viewers, at 6 degrees of arc, the total maximum of vertical disparity is 0.001745e.

$$Vertical\_Disparity = 2 * e * \tan\left(\frac{arc\min ute}{2}\right) \tag{2.20}$$

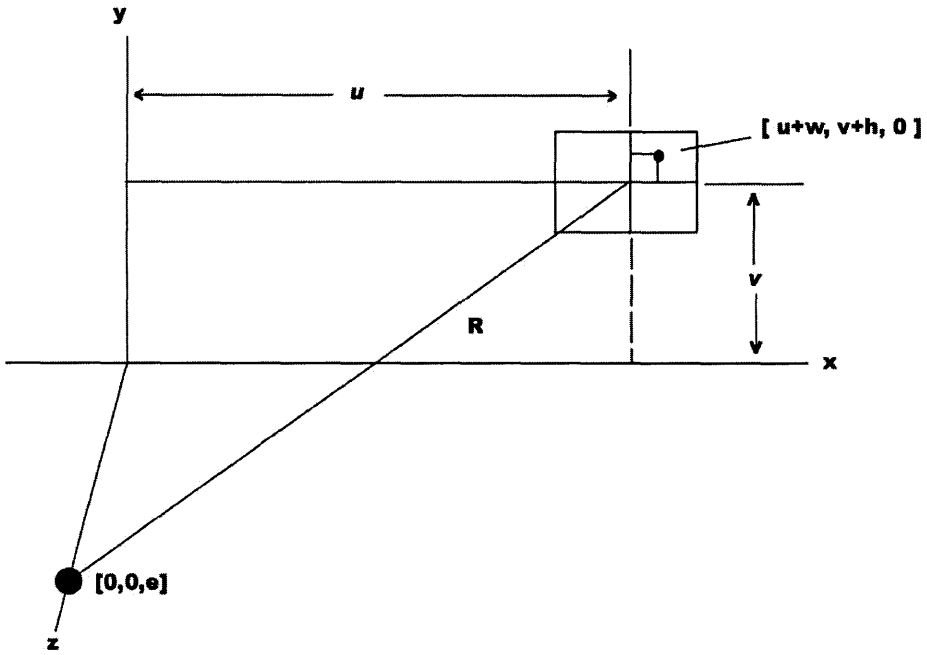


Figure 2.6. McAllister's keystone projective warp.

McAllister calculated 2.21 and 2.22 which determine the coordinates where the ray intersects the image plane and are the forward projection map.

$$w'(u,v,w,h) = \frac{we^2\sqrt{u^2+v^2+e^2}}{\sqrt{u^2+e^2}\,(u^2+v^2+e^2+vh+uw)} \tag{2.21}$$

$$h'(u,v,w,h) = \frac{e((u^2+e^2)h-uvw)}{\sqrt{u^2+e^2}\,(u^2+v^2+e^2+vh+uw)} \tag{2.22}$$

If we do this for all points in the translated image we obtain the image *I'* (centered at *O*) which normally the viewer would see but in this case it is what the camera, seen at *e*, records. For the special cases when μ or ν is zero we can reduce equations 2.21 and 2.22 to the following equations.

$$w'(u,0,w,h) = \frac{we^2}{(u^2+e^2+uw)} \tag{2.23}$$

$$h'(u,0,w,h) = \frac{e\sqrt{u^2+e^2}h}{(u^2+e^2+uw)} \tag{2.24}$$

$$w'(0,v,w,h) = \frac{we\sqrt{v^2+e^2}}{(v^2+e^2+vw)} \tag{2.25}$$

$$h'(0,v,w,h) = \frac{e^2h}{(v^2+e^2+vw)} \tag{2.26}$$

We want to deform the image *I* so that the viewer sees it without keystoning after translation. We must find functions g_1 and g_2 for which $w = g_1(w',h')$ and $h = g_2(w',h')$. The inverse of a projective warp is a projective warp; the functions g_1 and g_2 have a form similar to the previous equations.

$$w = g1(w',h') = \frac{(e^2+u^2)^{3/2}(e^2+u^2+v^2)w'}{u\sqrt{e^2+u^2}(v^2-e^2-u^2)w'+ev\sqrt{e^2+u^2}\sqrt{e^2+u^2+v^2h'}+(e^4+e^2u^2)\sqrt{e^2+u^2+v^2}} \tag{2.27}$$

$$h = g1(w',h') = \frac{\sqrt{e^2+u^2}(e^2uv+u^3v+uv^3)w'-e(e^2+u^2+v^2)^{3/2}h'}{u\sqrt{e^2+u^2}(v^2-e^2-u\ )w'+ev\sqrt{e^2+u^2}\sqrt{e^2+u^2+v^2h'}+(e^4+e^2u^2)\sqrt{e^2+u^2+v^2}} \tag{2.28}$$

For the special case $v = 0$ we get the expression

$$w = \frac{(e^2+u^2)w'}{e^2-uw'} \tag{2.29}$$

$$h = \frac{e\sqrt{e^2+u^2}h'}{e^2-uw'} \tag{2.30}$$

and for $u = 0$ the expression reduces to

$$w = \frac{e^3\sqrt{e^2 + v^2}\,w'}{e^3 - evh'} \tag{2.31}$$

$$h = \frac{e(e^2 + v^2)h'}{e^3 - evh'} \tag{2.32}$$

2.6 Lens Distortion

Lens distortion, for the purposes of this study, can be broken down into tangential and radial distortion. Tangential distortion is caused by imperfect centering of the lens components and other manufacturing defects in a compound lens (Lin, 2000).

Radial distortion, often called pin-cushion or barrel distortion, is another source of image distortion and induced vertical parallax. This distortion is caused by the use of spherical lens elements, resulting in the lens having different focal lengths at various radial distances from the center of the lens. Increasing focal length from the center of the lens is called pin-cushion distortion while a decreasing focal length from the center is barrel distortion.

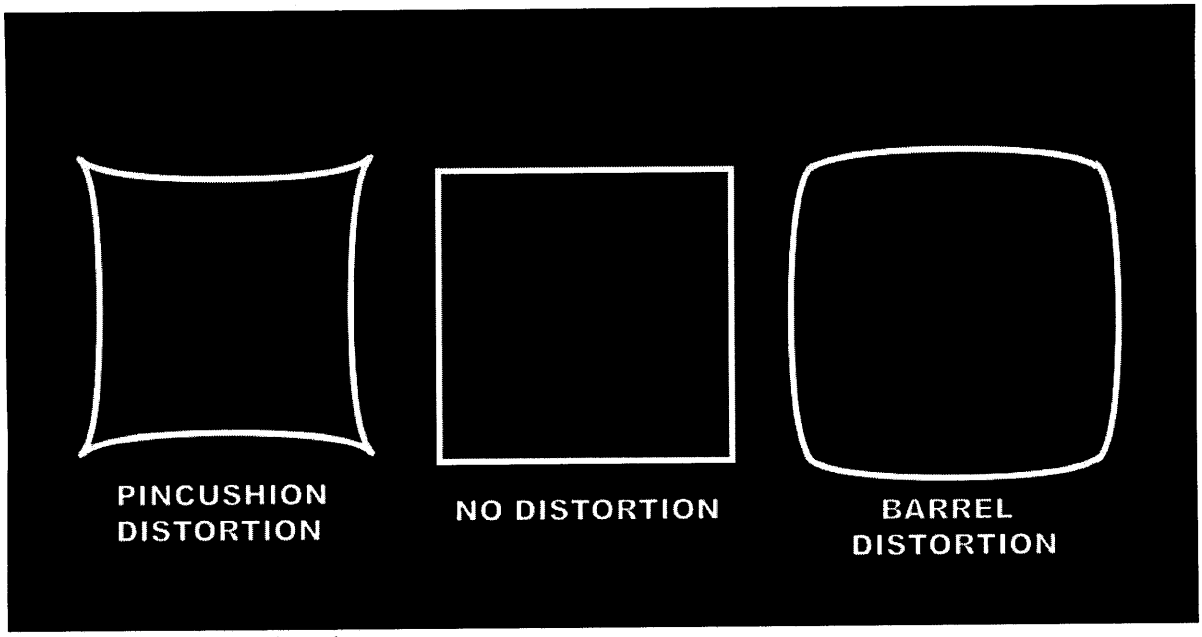


Figure 2.7 Pin-cushion distortion.

The equation that corrects for the curvature of an idealized lens is shown in equations 2.33 and 2.34. For many projections a_x and a_y will be similar or related by the image width to height ratio. The more lens curvature the greater the constants a_x and a_y will be, where the value 0 is for no correction and 0.1 would be associated for a typical wide angle lens. The "||" notation indicates the modulus of a vector (Bourke, 2004).

$$P'_x = P_x(1 - a_x\|P\|^2) \quad 2.33$$

$$P'_y = P_y(1 - a_y\|P\|^2) \quad 2.34$$

It is assumed that the image coordinates are normalized in both axes such that $-1 < x < 1$ and $-1 < y < 1$. The reverse transform that turns a perspective image into one with lens curvature is given by the following equations.

$$P_x = \frac{P'_x}{1 - a_x\left\|P/(1 - a_x\|P\|^2)\right\|^2} \quad 2.35$$

$$P_x = \frac{P'_y}{1 - a_y\left\|P/(1 - a_y\|P\|^2)\right\|^2} \quad 2.36$$

When correcting a lens distorted image it is necessary to use the reverse transform. The reason is that one doesn't normally transform the source pixels to the destination image but rather one wants to find the corresponding pixel in the source image for each pixel in the destination image. This method guarantees that all output pixels are calculated and that there are no "holes" in the final image.

2.7 Stereoscopic Tools

The following material showed some of the main properties that arise and determine the correct stereo capture and display. Several 3D capture/display systems built with these properties in mind will be reviewed in Chapter 3. Well-established stereoscopic imaging principles will be used to make the Spatial Imaging Scalable Camera Array robust for a multitude of environments as described in Chapter 4.

CHAPTER 3. CAMERA CAPTURE FOR 3D

Stereo capturing began in the mid-1830s with Sir Charles Wheatstone's design and use of the stereoscope to understand stereopsis. The stereoscope essentially was the foundation for stereophotography. The first photographs were made with a single camera that was moved through a distance of about 65mm, the distance of the average human interoccular, for a one second exposure (Bowers, 2001). Later in 1849 Sir David Brewster coined the idea of the twin-lens stereo camera (Brewster, 1856). Since then many inventions have been developed in both the creation and display for viewing active and passive stereo pairs.

3.1 Stereoscopic Cameras

Throughout the history of stereo cameras, their design can be broken into two classes, multi-lens group systems and single-lens group systems. In both group systems there is an even further distinction of single sensor and multiple sensor subclasses. Each class and subclass has its advantages and disadvantages.

3.1.1 Single-lens, Single-sensor Designs

The typical single camera stereoscopic system essentially follows Wheatstone's use of one camera except now the camera is fixed to a slide bar. One image is taken at one point, the camera moved along the bar to a fixed separation then the second image is taken. The advantages of such a design allow for clear alignment and matching is much easier.

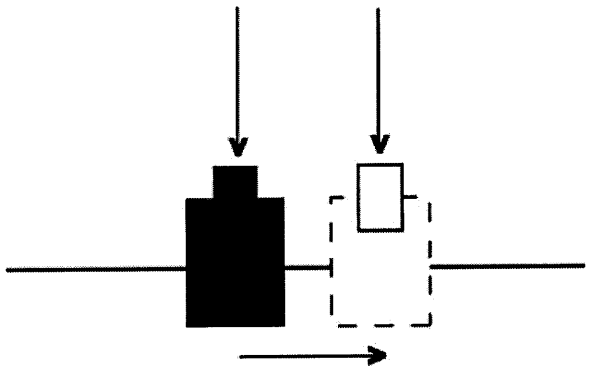


Figure 3.1. Single camera moved on slide bar.

Another common technique, emphasized in the Stereo 70 system in stereo cinema by NIKFL, the Soviet Motion Picture and Photography Scientific Institute, split the 70mm band of film into two 35mm frames (Ovsjannikova, 1975).

Paul (1997) uses a video camera rotating about an axis parallel with its own, with some oscillation to allow convergence. The rotation speed is an exact fraction of the frame rate to give stereoscopic image pairs. Goshtasby (1993) created a single camera with mirrors and a mechanically switching mirror timed with the video frame rate, and a system using

a refractive block rotating at an angle in time with the video frame rate. The refraction difference between the two angles of the block allows a stereo pair to be formed.

Lo's (1995) system uses a switchable lens aperture where the light focused through different parts of the lens forms a disparity for objects not in the focal plane of the system. There are several elements in the system that switch and allow a variable separation. A similar technique has been used by Watanable and Mayhew and Costales (Montgomery, 2002).

Many others such as Koung, Burke, Sudeo, Tashiro, and Toh have used adapters that can be placed on the front of ordinary video cameras providing alternate frame stereo images to be taken. These systems use a mirror and beamsplitter combination and require that there is some switching mechanism that is timed with the video frame rate (Montgomery, 2002).

Additionally, Ahmed (1999) showed a system that used a special 3D Ring Lens instead of traditional lenses. It captures two images for every point in the object field of the new system, except in occlusion, with a fixed geometric constraint between the correspondence points.

Rohaly and Hart (2000) created a high resolution, ultra fast 3D imaging system based on projecting a speckle pattern onto an object and imaging the resulting pattern from multiple angles. The system uses a special lens camera system with a rotating off-axis aperture. The rotating aperture allows adjustable non-equilateral spacing between defocused images to achieve greater depth of field and higher sub pixel displacement accuracy. Additional processing is done for correlating the views and reduces image disparity.

3.1.2. Single-lens, Multi-sensor Designs

To avoid the problem of synchronization several designs use a single-lens system with multiple sensors. In the early stereo-cinema field the Norling camera was essentially a 70mm camera that used two 35mm rolls of stock, immediately adjacent to each other, exposed through two side-by-side apertures. Lenses of various focal lengths featured continuously variable interaxial separation using a periscope-type device that maintained image orientation through a range of rotation (Norling, 1953).

Mochizuki (1992) and Shimizu (1999) use mirror type adapters to image a scene into two separate homologous images on two places on an image sensor. This requires adding components that need to be aligned and rely on minimal lens aberrations on the boundaries of the lens surface. Another problem is cross talk between images since the resolution of the images is less than half of the sensor resolution.

For medical applications, specifically close endoscopic work, McKinley (1998) created a compact single-lens system using a single primary lens with small secondary lenses. The secondary lenses image the light from two separate directions from the primary onto two sensors.

3.1.3. Multi-lens, Single-sensor Designs

Multi-lens, single-sensor designs reduce all the image sensors to a single sensor. Merging all of the sensors makes it easier for matching and synchronization although it does make it harder for the system to be adaptable and versatile. One particular example of this type of system is seen in the Mars Polar Lander stereo imager. This system uses a single CCD with two lenses and mirrors for control (Smith, 1998). A similar design was also introduced by Bove (1989) for recording a long- and short depth-of-field image at the same time side-by-side on a single frame of film.

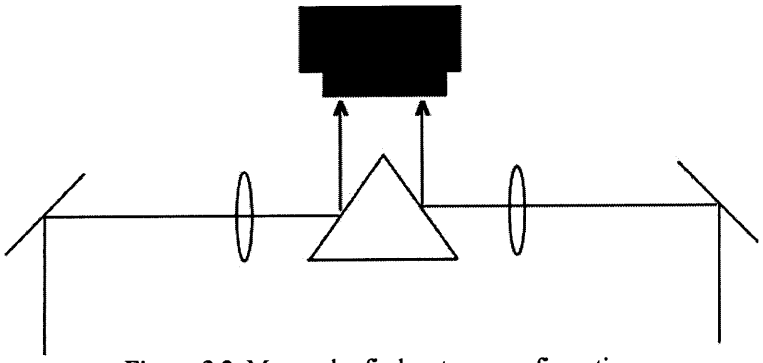


Figure 3.2. Mars polar finder stereo configuration.

3.1.4 Multi-lens, Multi-sensor Designs

This is perhaps the most common and most studied category of stereo cameras as it largely consists of two or more separate cameras in various orientations. Most rigs have similar schemes using two studio cameras mounted side by side on a sliding bar. This type of system became the commercial norm in the stereo-cinema industry.

The slide bar allows for the cameras to be easily translated and converged. The lenses can be controlled independently to allow zooming. The disadvantage is that the system is considered bulky. The advantage is that the control of the system is versatile, adaptable, and easy to take both 2D and 3D shots. The alignment process has to be precise to reduce disparity and is time intensive.

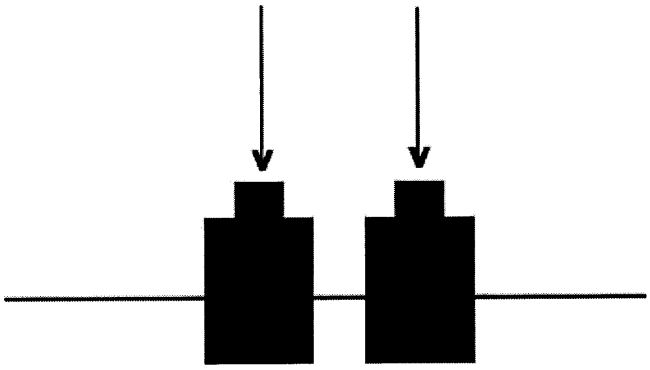


Figure 3.3. Two camera mount.

To ease alignment and reduce disparity separation, Gunzburg (1953) used a setup with cameras facing each other shooting into mirrors. A stereo cinematic camera called the Fox unit used two cameras set at right angles, one camera shooting the subject's image reflected by a semi-silvered mirror, with the other shooting through a mirror (Lipton, 1982). However the Fox rig added extra alignment parameters and ghosting due to reflection from the semi-silvered mirror as shown below.

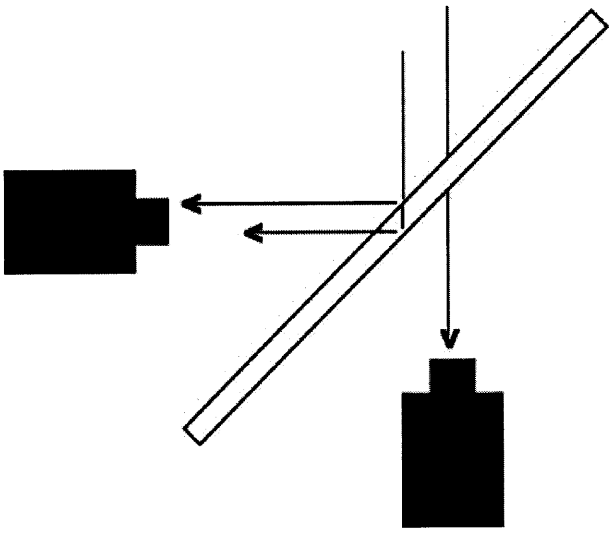


Figure 3.4. Fox unit.

Other mirrored systems include Iwasaki's (1999) four-mirror system allowing for changes in effective baseline separation while Sekida (1993) had a similar system with changing convergence. Matsunaga (1998) also describes a four-camera system and mirrors, which allows for a high-resolution image center while using a change of field to adjust disparity.

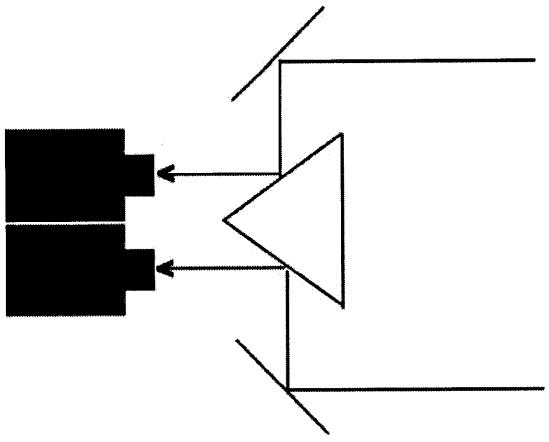


Figure 3.5. Iwaskai's four mirror system.

Some additional early cinematic multi-lens, multi-sensor designs involved the Universal

stereo camera, again, used two side-by-side machines, but with one upside down, for capturing short interaxial distances. There was also an early stereo-camera that used a triptych process using a triple array of cameras and projectors for presentations on large, curved screens (Lipton, 1982).

3.2 Image Based Modeling and Rendering

There is a significant trend in stereo cameras or multi-view camera arrays being used for creating three-dimensional environments from one or more images. Instead of making environments from the bottom up, once can use images, which provides naturally photo-realistic rendering . Using images allows for the creation of immersive 3D environments for real places thus expanding the applications of entertainment, virtual tourism, telemedicine, telecollaboration, and teleoperation (Levoy, 1996). Multiple images can also be used to relight scenes, synthesize images from original viewpoints, and derive geometric models (McMillan, 2004). Below is a list of camera rigs that are being used for depth information and image based rendering.

Shao (2002) has mixed rendering techniques with motion tracking algorithms to create multi virtual camera views. Although process intensive, combining these equations with new photo-realistic rendering have allowed for unique imaging solutions.

Cull et al (1997) created a streaming 3D video using a Beowulf-style distributed computer with 64 processors and 64 video camera/capture pairs. The system was a test-bed for comparing sensor spaced modeling and reconstruction algorithms. They used tomographic and stereo triangulation algorithms on this space and consider mappings from the sensor space to associated display spaces.

Kawakita (2000) created the Axi-vision camera that can acquire both color and distance information of objects. An intensity-modulated light illuminates objects and the camera with an ultra-fast shutter that captures the light reflected from the scene. The distance information is obtained from the two images of the same scene taken under linearly increasing and decreasing illuminations.

Nyland's (2001) image-based model approach combined images with depth taken from several places around an environment. The range data or depth is acquired using a scanning laser rangefinder. Each scan is approximately 10 million range samples at a resolution of 25 samples per degree and an accuracy of approximately 6 mm. The images are taken with a high-resolution camera registered to the laser rangefinder scans.

Naemura et al (2002) developed a system using densely arranged cameras that can perform processing in real time from image pickup to interactive display, using video sequences instead of static images, at 10 frames per second. Their camera array consists of 16 cameras that can be arranged in a lattice or connected in a row. The cameras can also be arranged sparsely by inserting empty units between camera head units. Also, the lenses can be changed to capture light rays from several viewing angles.

Their method of image capture involves a quad processor that combines the video

sequences from four cameras and outputs a video sequence divided into four screens. Video from the 16 cameras therefore consists of four-screen sequences obtained through the use of four quad-processor units. A fifth quad processor combines these four sequences so that the video from 16 cameras becomes a single 16-screen sequence. In this regard, when connecting quad-processor units in an L-level cascade, one video board can accommodate 4L cameras worth of video in the computer. All image alignment is done by translation, a rather quick and non-precise method but still provided acceptable image capture.

3.3 Three-Dimensional Displays

This summary list is not intended to be a complete survey of the 3D display field, but rather to describe the range of potential technologies upon which such displays depend. The list is separated into stereocopic displays and autostereoscopic displays.

3.3.1 Stereoscopic Displays

Stereoscopic displays, also known as aided viewing systems require users to wear special devices to separate two different perspective views, one each going to the left and right eyes quasi-simultaneously. Several multiplexing methods exist to send the appropriate optical signals to each eye. In all but some location-multiplex displays, all stereoscopic techniques force the observer to focus on a fixed image plane.

3.3.1.1 Color-multiplexed (anaglyph) Displays

Anaglyph displays present the left and right eyes images that are filtered with near-complementary colors such as red and green. The observer wears respective color-filter glasses for separation of the images.

3.3.1.2 Polarization-multiplexed Displays

The basic arrangement consists of two monitors or projectors polarized 90 degree with respect to each other using orthogonally oriented linear or circular polarized filter sheets. The two views are combined by a beam-splitter and the observer wears polarized glasses. Although, in theaters, the two projection systems are overlapped onto the same screen while the observers wears polarized glasses.

3.3.1.3 Time-multiplexed Displays

The human visual system has a memory effect in which it is capable of merging the constituents of a stereo pair across a time lag of up to 50 ms. Time-multiplexed displays exploit this feature by showing the left and right eye views in rapid alternation and synchronized with an LC-shutter, which opens in turns for one eye, while occluding the other eye.

3.3.1.4 Time-sequentially Controlled Polarization

There are displays that have merged the time and polarization-multiplex techniques. A monitor's faceplate is covered with a modulator, consisting of a linear polarizer, a liquid crystal π-cell and a quarter-wave retardation sheet to turn linear polarization into circular polarization. The π-cell switches polarization in synchronism with the changes of the left and right eye views. Circular polarizing glasses serve for de-multiplexing.

3.3.1.5 Location-multiplexed Displays

In these types of displays the two views are created at separate places and relayed to the appropriate eye through separate channels by means of lenses, mirrors, and fiber optics. The most recognized of these displays is the helmet-mounted display or head mounted display (HMD). With HMDs, the perceived images subtend a large viewing angle, typically up to 120 degrees horizontally by 80 degrees vertically. Usually, the outside environment is occluded from the viewer by a visor, allowing them to be totally immersed in the scene environment (Melzer, 1997).

3.3.2 Autostereoscopic Displays

In general, the goal is to move away from aided viewing devices to free viewing systems or autostereoscopic displays. In autostereo displays the ability to address each eye is integrated into the display itself knowing that each eye is occupying different points in space. Direction multiplex is the only way to channel the information of the left and right views into the appropriate eyes. Compared to stereoscopic techniques, it is often possible to multiplex more than two views at a time. For this reason individual perspective views can be delivered to different observers. Volumetric and electro-holographic approaches produce 3D images where the effective origin of the waves entering the observer's eye match with the apparent spatial position of the corresponding image points.

3.3.2.1 Electro-holography

Holographic techniques can record and reproduce the properties of light waves in terms of amplitude, wavelength, and phase differences almost to perfection, which makes it the ideal autostereoscopic viewing 3D technique. Recording requires coherent light to illuminate both the scene and the camera target. For replay, the recorded interference pattern is again illuminated with coherent light. Diffraction or phase modulation will create an exact reproduction of the original wavefront.

Electro-holography or Holovideo, created at the MIT Media Laboratory Spatial Imaging Group is a truly (natural holographic projection) three-dimensional real-time digital imaging medium. Holovideo demonstrates that the two crucial technologies, computation and optical modulation, can be scaled up to produce larger, interactive, color holographic images. Synthetic images and images based on real-world scenes are quickly converted into holographic fringe patterns using diffraction-specific computational algorithms (Plesniak, 2003). To diffract light to form an image in real time employs a scanned, time-multiplexed acousto-optic modulator, and utilizes parallelism at all stages (Lucente, 1994).

3.3.2.2 Volumetric Displays

A volumetric display will project image points to definite loci in a physical volume or space where they appear either on a real surface, or in translucent aerial images forming a stack of distinct depth planes. If using a real surface a self-luminous or light reflecting medium is used which either occupies the volume permanently or sweeps it out periodically. Systems like these are produced by Actuality Systems.

Translucent systems create aerial images in free space which the observer perceives as cross sections of a scene lined-up one behind the other. The images belonging to

different depth layers are written time-sequentially. These type of multiplanar displays are well known by the BBN SpaceGraph and the TI varifocal mirror display. Another example includes a color spatial display developed by Kenneth Carson at MIT using a raster frame buffer and varifocal mirror (Carson, 1984). The mirror rapidly changes the focal length in synchronization with the refresh rate of a 2D monitor, which subsequently varies the depth of a projected scene. The mirror was simply a bass drum with an aluminized mylar drum head on one side and a hifi woofer on the other side to provide a driving force for the vibration.

A slightly different and novel type of 3D volumetric display was created by Elizabeth Downing using scanners from surplus optical-disc players to trace two infrared laser beams through a transparent cube that contains light-emitting impurities. Where the two beams intersect, their combined energy causes the impurities to emit a burst of red, blue, or green light. The eye sees the illusion of a color image.

3.3.2.3 Direction-multiplexed Displays

These types of displays use optical effects like diffraction, refraction, reflection and occlusion in order to direct the light emitted by pixels of different perspective views exclusively to the appropriate eye.

3.3.2.3.1 Diffraction

Diffraction-based approaches use diffractive optical elements (DOE) or holographic optical elements (HOE) to diffract light into different directions creating multiple overlapping scenes for the viewer as they move their head.

3.3.2.3.2 Refraction

Refractive optical elements use picture sized large lenses or small lenslets to address the observer's eyes. One method, *integral imaging*, creates a spatial image composed of multiple tiny 2D images of the same scene, captured with a very large number of small convex lenslets. Each lenslet captures the scene from a slightly different perspective. A lens sheet of the same kind is used for display. As the image plane is positioned in to the focal plane of the lenslets, the light from each image point is emitted into the viewing zone as a beam of parallel rays at a specific direction. Therefore, the observer perceives different compositions of image points at different points of view.

Another refraction-based display uses lenticular techniques using an array of vertically oriented cylindrical lenslets. The light from each image point is emitted at a specific direction in the horizontal plane, but non-selectively in the vertical plane. Therefore, changes of perspective in accordance with vertical head movements cannot be achieved by optical means. Based on lenticular techniques, direct-view and projection-type 3D displays have been implemented.

The third type of refraction approach uses a field-lens placed at the locus of a real image in order to collimate the rays of light passing through that image, without affecting its geometrical properties. Various 3D display concepts use a field lens to project the exit pupils of the left and right image illumination systems into the appropriate eyes of the

observer. The effect is that the right view image appears dark to the left eye and vice versa.

3.3.2.3.3 Reflection and Occlusion

A common reflection-based approach uses a retro-reflective screen for direction multiplexing by reflecting the incident rays only into their original direction.

Occlusion-based approaches take advantage of parallax effects such as parts of an image that are hidden from one eye but visible for the other eye. These types of displays differ in the number of viewing slits ranging from a dense grid to a single vertical slit. They also differ in presentation mode in regards to being time-sequential versus stationary and in whether the opaque barriers are placed in front of or behind the image screen, parallax barrier versus parallax illumination techniques. Most systems are classified as barrier-grid, parallax-illuminated, or moving-slit displays (Pastoor, 1997).

3.4 Display Applications

The number of three-dimensional displays is numerous all with their specialties plus their inherent advantages and disadvantages. The goal however is to be able to use the SISCA with any number of these displays. It would be ideal that a viewer/user can choose within a software program the type of 3D display of their choice. Ultimately, the SISCA would not only capture and present the correct views to the display and viewer but also understand the specialties of each 3D display to create pleasing images.

CHAPTER 4. SPATIAL IMAGING SCALABLE CAMERA ARRAY

This chapter discusses previous work that led up to the design, testing, and analysis of the Spatial Imaging Scalable Camera Array (SISCA). Design issues will deal with the physical apparatus and the constraints of the system based on the cameras, servos, software, and the final three-dimensional output displays. Testing will involve the limitations of the system under three varying environments; unmanned aerial vehicles, studio/facial recognition, and surgical applications. Part of this testing will also include methods for camera calibration and correcting for geometric distortions for the 3D output. Lastly, a qualitative analysis of SISCA will be given

4.1 Previous System

Previously, a multiple camera array using a shearing geometry was shown to work with one computer for connection and processing in real time. In these recent experiments an array of four cameras were combined into one USB Hub and run at 5 frames per second to capture four views in a portrait type mode with a viewer's face taking up most of the frame. The four frames were stitched together to make one image that was then inkjet printed onto photographic quality print paper. Above the paper rested a 60 lens per inch (lpi) lenticular sheet to create the sensation of three-dimensionality to the stitched content. The number of lenses and the size of the images determined the camera spacing, object (viewer) distance, and the particular columns taken from each image for the best stitching and three-dimensional effect.

4.2 Design of SISCA

The lenticular application was a great idea and worked well with static scenes. However, the possibility of streaming multiple cameras at one time provided inspiration to tackle active scene acquisition and look at a much different image capture paradigm. This led to the design of the SISCA with a purpose to correctly pick stereo or multiple views from an infinite array of inexpensive compact image sensors and optimize the output data for specific user-defined 3D displays. Overall, the SISCA is meant to benefit the user in getting rich spatial information for their application while the underlying camera array handles the difficult task of being the virtual stereo-cinematographer. The biggest changes from the previous system was switching SISCA from a shearing geometry to a crossed-lens geometry and increasing the number of cameras from four to six.

4.2.1 Experimental Parameters

Six basic parameters that uniquely characterize a stereoscopic camera and display system had to be considered when designing the SISCA. The camera system configuration is determined by (1) the distance between the cameras (t_c), (2) the convergence distance (the distance away from the cameras at which the optical axes of the cameras intersect), (3) the resolving power of the cameras. The display system is determined by (1) the viewing distance of the observer from the display, (2) the size of the display (measured by its horizontal width) and (3) the distance between the viewer's eyes. Each of these parameters will directly influence the limits and quality of the stereo-camera output.

A major factor in determining these parameters is the environments/applications that the SISCA

is being used for. The three varying environments chosen for gathering experimental parameters were unmanned aerial vehicles, studio/room recording with specificity on portrait shots for facial recognition, and lastly hypostereo surgical procedures. Testing the scalable camera array under extremely different environments will ideally indicate its possible performance in other applications.

4.2.1.1 Unmanned Aerial Vehicle

Three-dimensional capture for unmanned aerial vehicles is important for both navigation and surveillance purposes. When looking at far off objects, as usually encountered in an aerial environment, those objects appear at stereo infinity. To capture any depth information or hyperstereo requires the two stereo views and hence cameras to be considerably far apart. This distance can sometimes be much larger than the wingspan of the aerial vehicle, which on average is 50 feet.

If t_c is larger than the wingspan, the UAV will have to move in a horizontal sweeping motion to capture both stereo views. This opens up several possibilities for image capture. A single camera can be used for capture as the plane moves from one position to the next. If a second camera is used then the aerial vehicle only has to move a shorter distance. An alternative is to come up with a way to use multiple cameras on the wing but allow for some form of motion-tracking algorithms to create interpolated scenes as the aerial vehicle sweeps in the horizontal direction from side to side and as it is moving forward.

For an aerial environment it was calculated that the shallowest convergence point D_o would be no less than 1000m and the maximum altitude reached by any UAV is approximately 19,500m. With the current camera system, SISCA has a stereo acuity up to 3256m with a t_c of 50 ft. This easily covers objects at 1000m. To record objects at 19,500m requires a stereo baseline of 299.4ft (≈ 300ft).

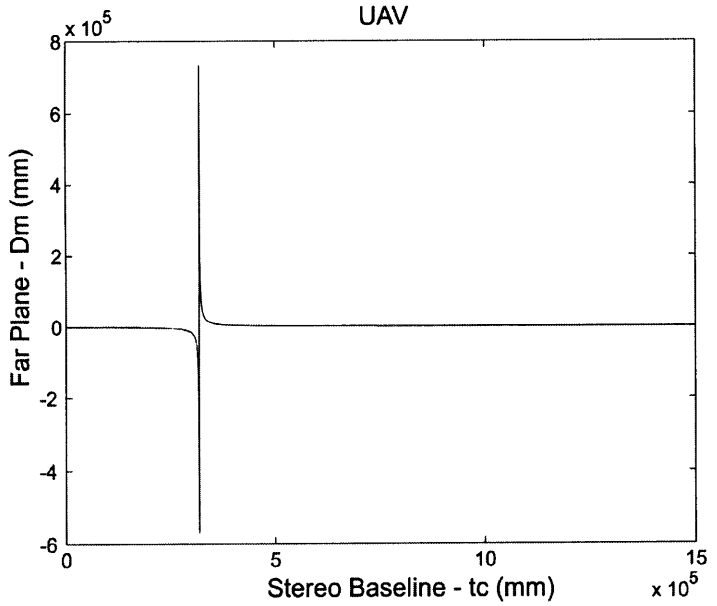


Figure 4.1. Maximum depth as the baseline separation is increased.

Ideally, reducing the stereo baseline is optimal for less travel distance and using more cameras on the wing for capture. Decreasing the stereo baseline while maintaining the current depth range involves increasing the focal length of the cameras, such that it has a narrower angle of view, or increasing the resolving power of the entire imaging system.

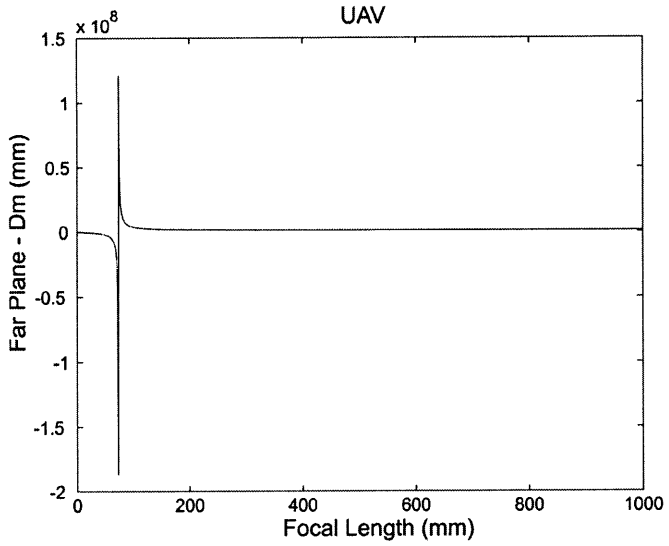


Figure 4.2. Maximum far plane depth versus changing focal length and a stationary camera separation of 50ft.

With the current imaging system, a baseline of 50ft, and a minimum convergence distance, D_o, of 1000m, the maximum far plane distance for acceptable viewing with a head-mounted stereo system with 1 degree of divergence using equation 2.12 gives a result of –50m. Without divergence the result is –79m. For the view sequential display the result is –13m with divergence and –18m for the non-convergent case. A negative D_m in this case is a result of a depth inversion and will not be pleasant to view. It also means that a much larger baseline separation or focal length camera is needed for proper scene recording.

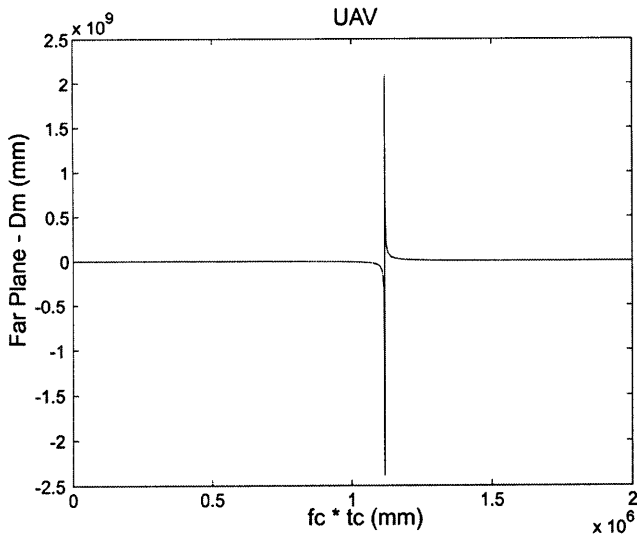


Figure 4.3. Maximum far plane depth versus a change in combination of both focal length and camera separation.

As the baseline separation increases the D_m decreases considerably as seen in Figure 4.1. This is even true for higher altitudes. Changing the focal length produces the following results in Figure 4.2. There is a definite range where any distinguishable depth can only be seen beyond the convergence point. Once a maximum is reached the far plane distance falls off rapidly. Figure 4.3 gives a combination of a changing focal length or changing stereo baseline on the far plane distance.

In general for any depth beyond the convergence plane requires a system with a hyperconvergence distance D_h as described by equation 2.11 where the focal length, the camera separation, and the stereo constant combine to be larger than the convergence distance D_o. If not, the far plane is inverted. In the current setup with a 50 ft camera separation and a fixed focal length of 3.5mm at D_o = 1000m would require a stereo constant of 5.4×10^8. This is for one meter of distance beyond the convergence plane. Even as the far plane depth increases the stereo constant must stay within 0.05, which is hard to achieve with most standard 3D displays.

It is also necessary to exclude nearby objects from the images. This isn't much of an issue when working in an aerial environment. If this isn't the case there will be excessive on-film divergence in the images. Overall, some general limitations have been described for an aerial environment, especially for acquiring depth at stereo infinity.

4.2.1.2 Studio/Facial Recognition

Studio three-dimensional capture is important for a multitude of applications from entertainment, to laboratories, to corporate boardrooms. More importantly it is beneficial to facial recognition in that it can capture multiple viewpoints of a face. This allows for precision feature matching from previously recorded images at varying angles.

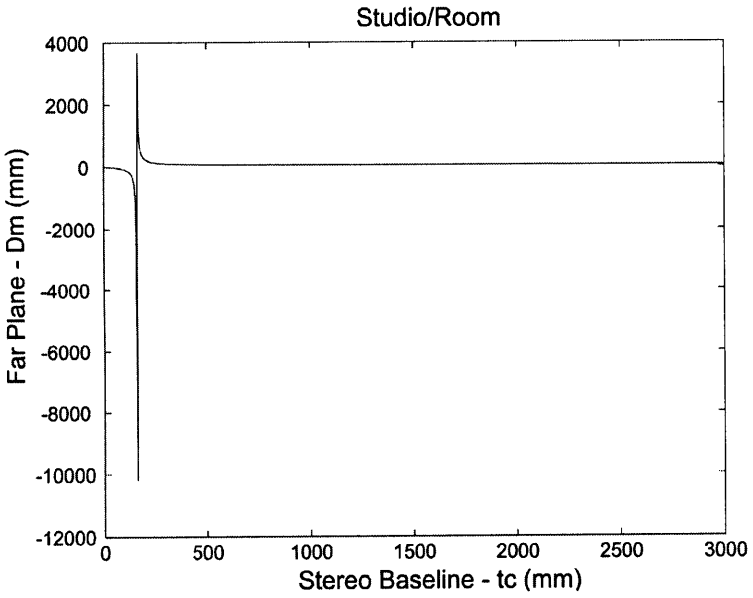


Figure 4.4. Maximum depth as the baseline separation is increased.

As a note, faces tend to lack enough surface detail which makes it hard to correlate stereo images. Horn (1989) and others try using shape-from-shading techniques and photometric stereo

to handle this correlation. Most likely, the best solution is to extract 3D image based models from the stereo or multi-view images. In this case the SISCA only captured a multitude of images but did not perform the facial recognition or any form of image based modeling or rendering.

Compared to the aerial vehicle application, the facial recognition is less constrained in that the camera rig can be as long as your room is wide. However, that isn't nearly an ideal situation but since the object is much closer to the cameras the whole array can be scaled down and kept still.

For this application the convergence point D_o will be no closer than 500mm. Since there is more room we can decide the camera separation more carefully along with the far plane D_m. With the head-mounted stereo display in the divergent case the minimum camera separation is 160mm and for the non-divergent case 104mm. If the view sequential display is used the camera separation must be 580mm and 434mm. Figure 4.4 shows the far plane as t_c is changed with a constant convergence plane at 500mm.

A trend in Figures 4.1-4.4 indicate that the maximum far plane distance is reached when D_h, composed of the focal length, camera separation, and stereo constant equal $\Delta D/2$, where ΔD is the difference between the convergence plane and far plane distances. Changing either one of the hyperconvergence parameters will determine the maximum far plane. Obviously the user will want to maximize depth or find particular depth regions depending upon the artistic intentions of the viewer and application. A useful guide is that half the maximum far plane distance will be achieved when $\Delta D/2$ is divided by the focal length of the camera system.

4.2.1.3 Surgical Applications

Teaching surgical procedures from multiple angles while expressing spatial relationships between extremely small anatomical features is important for future medical doctors along with the future of robotic telemedicine.

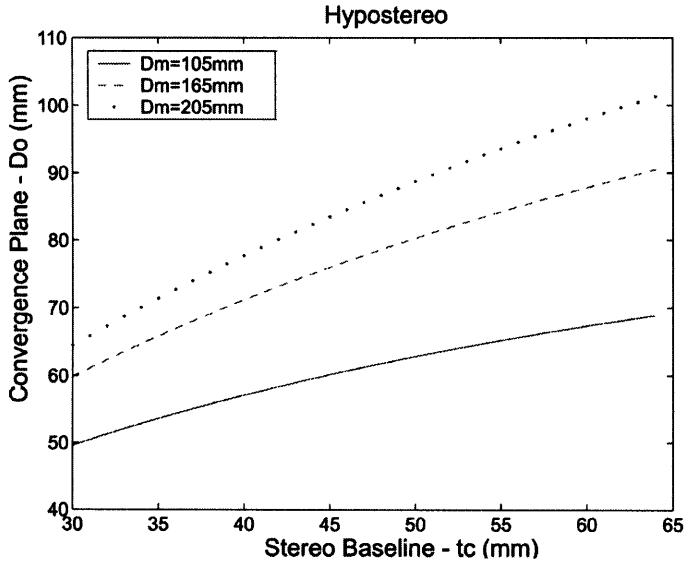


Figure 4.5. Convergence plane distance as the baseline separation is increased from 30mm to 65mm.

This application requires an extremely small apparatus to avoid hindering the procedure. The separation of the cameras is extremely close and should be as compact as possible. In this environment the small focal length lens is advantageous for macro stereo.

In the following scenario to create a sense of largeness involves decreasing the distance between the left and right eyes to show stereoscopic detail on small items. Thus the camera separation is under 65mm and thus for any far plane depth requires that the hyperconvergence distance is larger than the convergence distance. Keep in mind that when the cameras are stacked next to each other they have a separation of 30mm.

In a surgical scene that the background is very shallow such that the far plane distance is no more than 50mm to 100mm. Figure 4.5 shows the convergence distance with a varying t_c from 30mm to 65mm. This is for the divergent case using the head-mounted stereo display.

The amount of depth needed in a scene will decide where the cameras should be placed on the stereo baseline for proper focus on the convergence plane. Most likely, the further the camera is away from the physician the less obtrusive the device.

4.2.1.4 Review of Parameters

The previous sections on the three major environments/applications thoroughly discussed the capabilities and parameters needed for the correct capture and final display needed for either a head mounted stereo goggle system or view-sequential display. Table 4.1 below summarizes these parameters.

Table 4.1. Experimental parameters of SISCA for three applications.

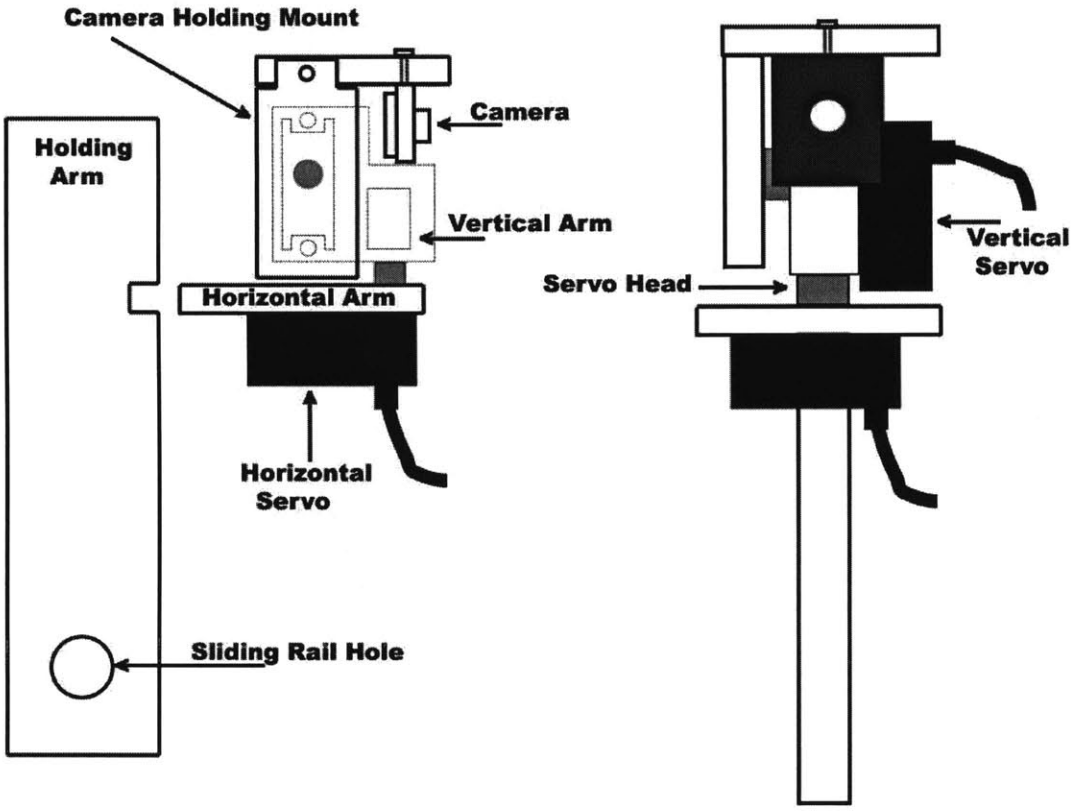
e : no divergence *d* : 1 degree divergence f_c : 3.5mm	**THREE-DIMENSIONAL DISPLAY**			
APPLICATION	***Head Mounted Stereo***		***View-Sequential***	
	e	*d*	*e*	*d*
Aerial Vehicle (m)	t_c =		t_c =	
Min D_o =1000	204	317	866	1287
Max D_o = 19500	4050	6232	16898	25095
Facial Recognition (mm)	t_c =		t_c =	
Min D_o = 500	104	160	434	580
Surgical Applications (mm)	t_c = 30-65 D_o =		t_c = 30-65 D_o =	
Min D_m = 50	37-43	33-40	16-25	21-30
Max D_m =100	59-76	48-67	26-43	19-34

All numbers in Table 4.1 rely on the focal length of the camera system being fixed at 3.5mm. The columns defined by *e* and *d* is for stereo with no convergence and 1 degree of convergence respectively. The final numbers depend on the parameters of the two displays and their interaction with the camera imaging system. A more detailed analysis of the interaction of these two displays and the current camera system are described later in the chapter. All the numbers would change if any of the systems or parameters were modified.

For the aerial vehicle and facial recognition applications the table gives the minimum camera separation t_c for the minimal amount of depth beyond the convergence plane. Increasing t_c will obviously increase the depth plane up to a limit. Since the aerial application has an altitude range there is a minimum and maximum convergence plane. The facial recognition application only depends on a minimum convergence plane while the maximum could be dependant upon the user.

In the surgical application a maximum far plane distance is given based on the depth needed for studying anatomical features. For hyposterео, the camera separation is kept under 65mm and the closest the cameras can lay next to each other is 30mm. In this case finding the convergence plane for these parameters are found. These limits and ranges help in the design and experimentation of the camera array.

4.2.2 Physical Apparatus



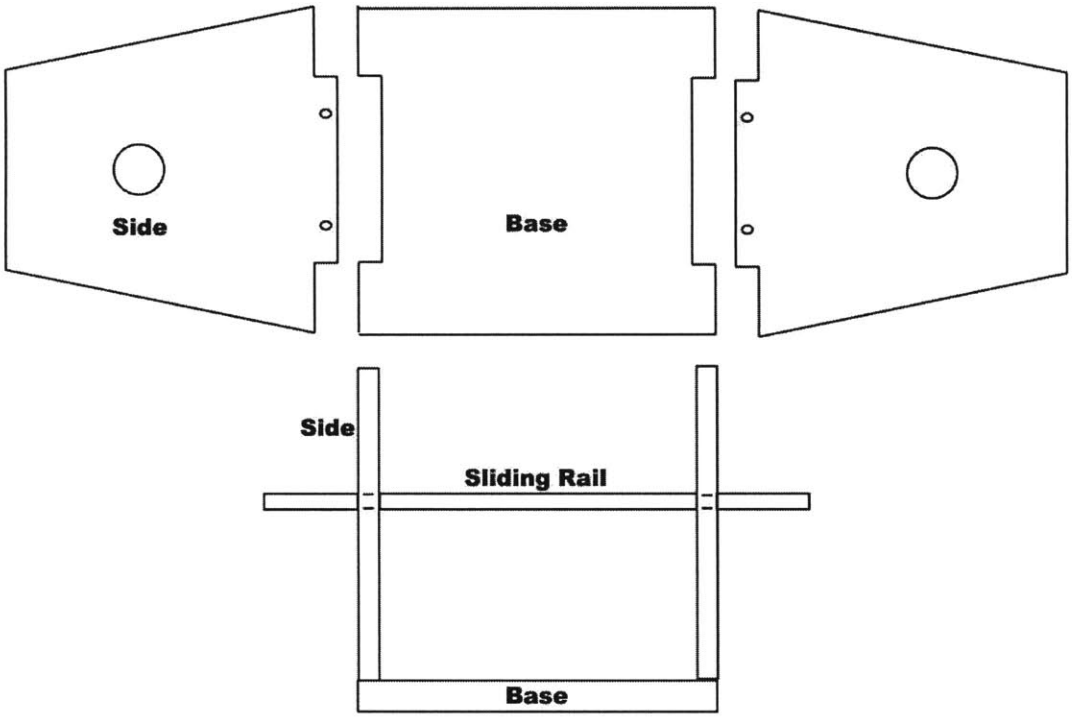


Figure 4.6. Spatial Imaging Scalable Camera Array.

The top diagram in Figure 4.6 begins with the schematic diagram of the SISCA. There is a side and front view of the camera holding and servo movement features. Each one of these camera mounts is placed on the sliding rail, which fits into the side and base fixture shown in the second diagram in Figure 4.6. The basic components of the camera mount starts with the holding arm that is made to support the weight of the outstretched camera and servos. The holding arm has a circular hole along with setscrews to allow for easy movement and locking capabilities upon the circular sliding rail. Additionally, the holding arm contains a notched area for the horizontal arm where the horizontal servo rests. Each servo has a rotating servo head as distinguished by the

gray knobs seen in the top figure. Upon the horizontal servo lies the vertical arm, which contains the vertical servo. Attached to the vertical servo is the camera holding mount. The camera holding mount has a screw to hold the camera stationary above the entire system. The bottom figure is the final assembled product of six camera mounts on the sliding rail.

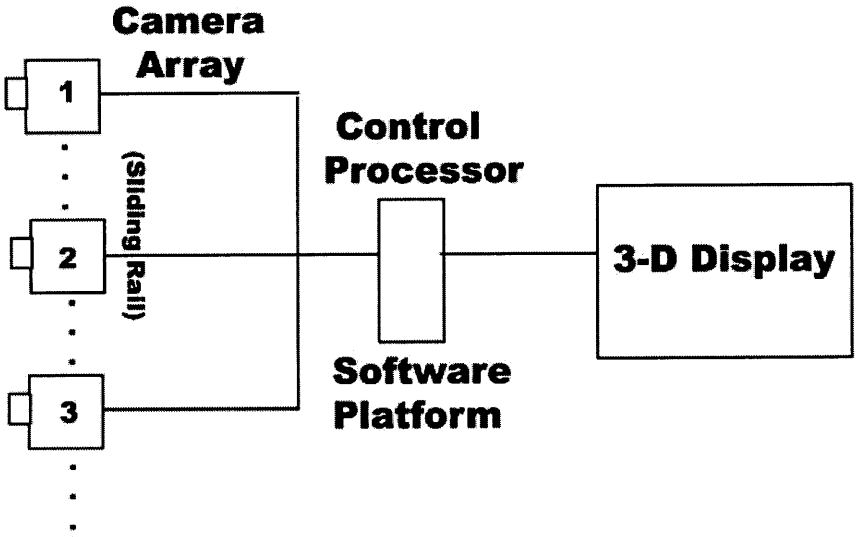


Figure 4.7. SISCA image processing chain.

The camera array uses USB connections that can be directly attached to the USB slots on a computer/control processor/software platform directly or through a USB hub. For SISCA two Belkin 4-port USB hubs were used to connect six cameras to the control processor. Figure 4.7 shows the generic interaction of the SISCA system. The camera array sends video images to the control processor; the software platform allows the user to interact with the camera array. This effectively changes how new video images are sent to the control processor. Lastly, the software platform allows for capture and proper modification of the video images to correctly match the final output to a 3D display.

4.2.2.1 Camera

The SISCA is using the same equipment - cameras, servos, software, as is being used over all testing environments. Obviously more sophisticated lenses with a wider degree of zooming ability could be used but at a much larger increase in cost. In this research the cameras have a fixed 3.5mm focal length. Since the focal length f_c is fixed, it will directly influence the near, far, and convergence points along with the baseline position of the cameras. Table 4.2 gives an overview of the current camera's systems properties.

The external and internal properties of the camera are constant but the system must have the stereo acuity to match the camera separation t_c and far plane point D_m for the testing environments. Using equation 2.1 the parameter t_e will be replaced by t_c. To calculate $\Delta\alpha$ involves dividing the angle of view of the lens by the resolving power of the entire system. In this case the angle of view is 85 degrees and the resolving power is 317 vertical lines horizontally across the entire CMOS chip. Changing either the camera separation, angle of view, or resolving power of the system will create varying degrees of available depth in a scene.

Table 4.2. SISCA Camera Properties

Camera Model:Logitech QuickCam Pro 3000	
Average Focal Length	3.5mm
Angle of View	85 degrees
Sensor Size	6.35mm x 6.35mm
Resolving Power	50 lines per mm

Each camera is attached to a horizontal and vertical servo such that the look around of the camera relies on the range of movement of the servos. Table 4.3 gives the range of movements for the Tower Hobbies System 3000 T-5 High Speed Nano Servo. Each servo can make incremental step changes at either 0.72 or 0.36 degrees. For higher precision the SISCA used a stepping rate of 0.36 degrees.

Table 4.3. SISCA Servo Properties

Servo Model: Tower Hobbies System 3000 T-5 High Speed Nano Servo	
Precision Movement per step	0.36 degrees
Step Range	245
Angular Range	88.2 degrees
Input	Serial

4.2.2.2 Display

In all testing cases the output has to be coordinated with either a stereo head mounted system or a 16-view sequential display. Table 4.4 lists the essential parameters for both of these three-dimensional displays.

Table 4.4. Three-Dimensional Display Properties

I-Visor Head Mounted Display	
Resolution	800x600x3
Field of View	31° diagonal
Image Size	44 in. horizontal
Viewing Distance	2 m
MIT/University of Cambridge 16-View Sequential Display	
Resolution	800x600x3
Field of View	30° diagonal
Image Size	30 cm horizontal
Viewing Distance	1.25 m

Some additional parameters required by the software platform are calculated with the camera and three-dimensional display properties. These parameters are listed in table 4.5 where the stereo constant K (Equation 2.8), the maximum screen parallax, P (Equation 2.6), and magnification, M (Equation 2.2) are listed. The subscript d is for the 1-degree divergent case and e is the non-divergent case.

Table 4.5. Display/Camera Parameters

View-Sequential Display		Head Mounted Stereo Display	
P_d	86.8750	P_d	100
P_e	65	P_e	65
M	21.4286	M	89.4080
K_d	4.5042	K_d	1.1185
K_e	3.0333	K_e	0.7270

4.2.2.3 Software Platform/User Interface

Controlling the camera array and the output to the three-dimensional display is done through the SISCA GUI user interface. The interface consists of a view of each one of the cameras as shown in the top portion of Figure 4.8. The bottom portion is where the user has control over the environment that will be chosen such that the scene input will be correctly matched to the 3D display. Currently, the environments that can be chosen are 'Aerial' for an unmanned aerial vehicle, 'Facial' for facial recognition in a studio, and 'Hypostereo' for recording surgical operations.

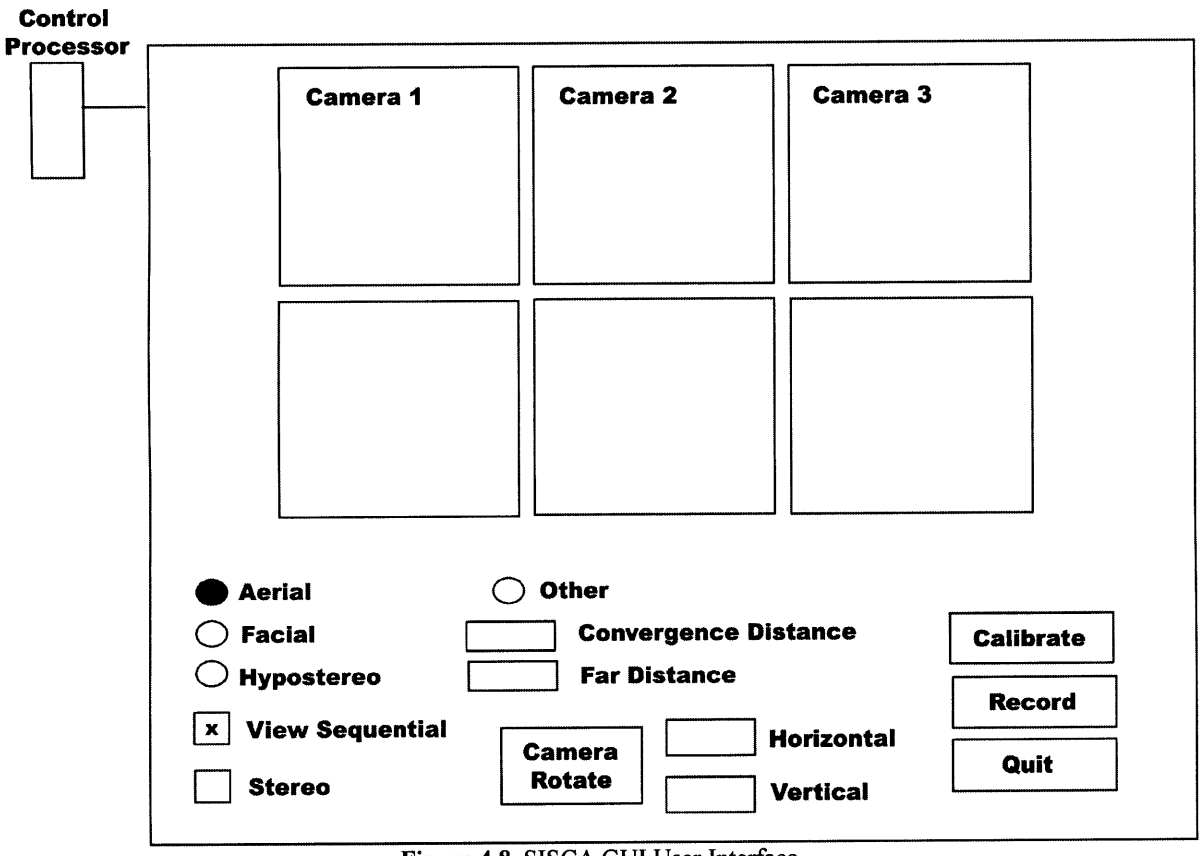


Figure 4.8. SISCA GUI User Interface.

The parameters such as the convergence plane and far plane distance are set to constant values as taken from Table 4.1. Using the constants in Table 4.5 and those in Table 4.1 along with the depth range equations, the camera separation t_c can be computed. If t_c is known then the convergence plane is kept constant and the far plane distance can be computed.

The other factor that will change t_c is the 3D display the user chooses which is either the view-sequential device or the stereo head mount system. Two other features included in the GUI is the 'Other' button such that the user can enter a convergence and/or far plane distance. Also, if the user has a specific object to track the cameras can be rotated directly in the 'Camera Rotate' box by entering the amount of movement in degrees the user wants to move horizontally and/or vertically. In this mode all of the cameras will move the same amount.

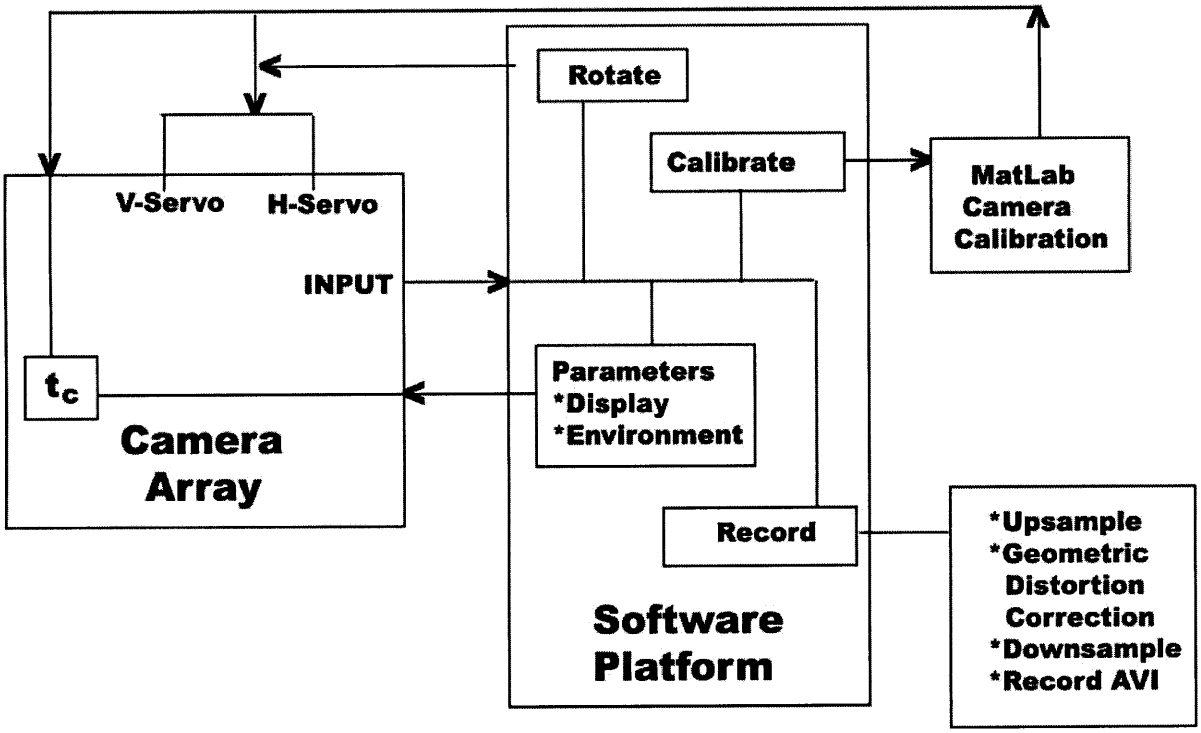


Figure 4.9. Software flow.

The buttons on the right side allow for calibration, recording, and quitting the system. The calibration process will be explained in section 4.3. Calibration requires 8-12 images of a test target. After the target has been placed the calibrate button can be pressed and each of the cameras records a snapshot of the target. Each time the calibration button is pressed the new snapshots are sequentially numbered and separated into folders for each camera. The record button will record AVI files from each camera and will upsample the scene, perform geometric distortion corrections such as keystoning and lens distortion and then downsample the images into the final AVI file. These files are then outputted to the 3D display. A flow of the entire SISCA GUI is shown in Figure 4.9.

4.3 Testing Procedures

4.3.1 Camera Calibration

Camera calibration has been an important feature in determining the geometrical and optical characteristic properties, the intrinsic parameters, of the image capturing devices while also providing the 3D position and orientation of the camera frame relative to a certain world coordinate system, the extrinsic parameters. This is crucial when making estimations about depth parameters in a scene with one or more cameras without prior knowledge of actual distances or sizes. Many attempts and similar techniques have been employed for camera calibration.

There are four categories in which most calibration techniques fall into. The first category involves full-scale nonlinear optimization. In this category the most common method is direct linear transformation (DLT) originally created by Abdel-Azis and Karara (1971). The DLT method uses a set of control points whose object space/plane coordinates are already known. The control points are normally fixed to a rigid calibration frame. The flexibility of the DLT-based calibration often depends on how easy it is to handle the calibration frame.

The DLT avoids the necessity for a large-scale nonlinear search by solving for a set of parameters or coefficients of a homogeneous transformation matrix with linear equations. By ignoring the dependency between the parameters results in a situation with the number of unknowns greater than the number of degree of freedoms. Tsai (1987) improved upon this by finding a constraint or equation that is only a function of a subset of the calibration parameters to reduce the dimensionality of the unknown parameter space, also known as the radial alignment constraint. Tsai's two-stage technique is well known and frequently used in the robotic and computer vision community.

Zhang (2002) has had a significant impact on camera calibration and has come up with a technique that requires a single camera to observe a planar pattern shown at least two different orientations. In Zhang's method radial lens distortion is modeled and the calibration consists of a closed-form solution, followed by a nonlinear refinement based on the maximum likelihood criterion. This technique has been incorporated into both an Intel OpenCV application and Jean-Yves Bouguet's comprehensive MATLAB camera calibration program. For calibrating SISCA Bouguet's MATLAB program was used.

The following program uses a pinhole model of 3D-2D perspective projection with 1st order radial lens distortion. The model has 11 parameters.

The five intrinsic properties are listed below.

- Focal length: Usually calculated in pixels and given in both the x and y direction.
- Principal Point: For small cameras is also the nodal point of the system.
- Skew Coefficient: The coefficient defines the angle between the x and y pixel axes.
- Distortions: Image distortion coefficients (radial and tangential distortions).
- Pixel error: The error between homologous pixel points

The six extrinsic properties are listed below.

- R_x, R_y, R_z – rotation angles for the transform between the world and camera coordinate frames.
- T_x, T_y, T_z – translational components for the transform between the world and camera coordinate frames.

In Bouguet's MATLAB program an extra extrinsic parameter is given, a rotation vector, which is calculated from the rotation matrix produced by Rodrigues rotation formula. Most calibration programs either guess or have user input into the number of sensor elements (pixels) in the x and y direction and roughly the size of the pixels in both dimensions.

The standard calibration device used with Bouguet's program is a standard checkerboard pattern as seen in Figure 4.10. Each square is 30mm by 30mm. For calibration the cameras don't need to be aligned with each other since any mismatch will be corrected after the calibration using the precision servos. The checkerboard pattern is shown at different orientations, not parallel to the image plane, for 8 to 12 calibration images for each camera. However, to make a correlation between cameras requires that the numbered images from each camera recorded the same target orientation.

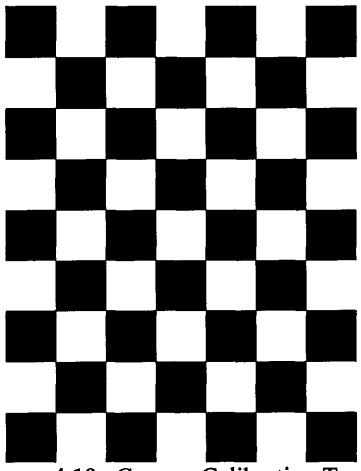


Figure 4.10. Camera Calibration Target

Next, each camera is calibrated separately using Bouguet's MATLAB GUI. All the calibration images for a particular camera are loaded into a database. The default size of the patches on the checkerboard are taken as 30mm. Regardless of orientation, a single corner on the checkerboard must be chosen as the initial corner and that same location must be used as the initial corner for every calibration image from all the cameras. Then the three other corners are picked on the checkerboard pattern. All the corners for each patch are highlighted as seen in Figure 4.11.

Once this has been done for all the images a calibration is done on all the images. The calibration is an iterative process. The result is a list of intrinsic camera parameters. It also gives the extrinsic properties of the camera in regards to the calibration target in regards to a rotation matrix and translation matrix. Figure 4.12 shows a typical output response of the orientations of

the target to the camera. On the figure in part (a), the frame (O_c,X_c,Y_c,Z_c) is the camera reference frame. The pyramid corresponds to the effective field of view of the camera defined by the image plane. There is an option to switch from a camera-centered view to a world-centered view as shown in part (b). On this new figure every camera position and orientation is represented by a pyramid.

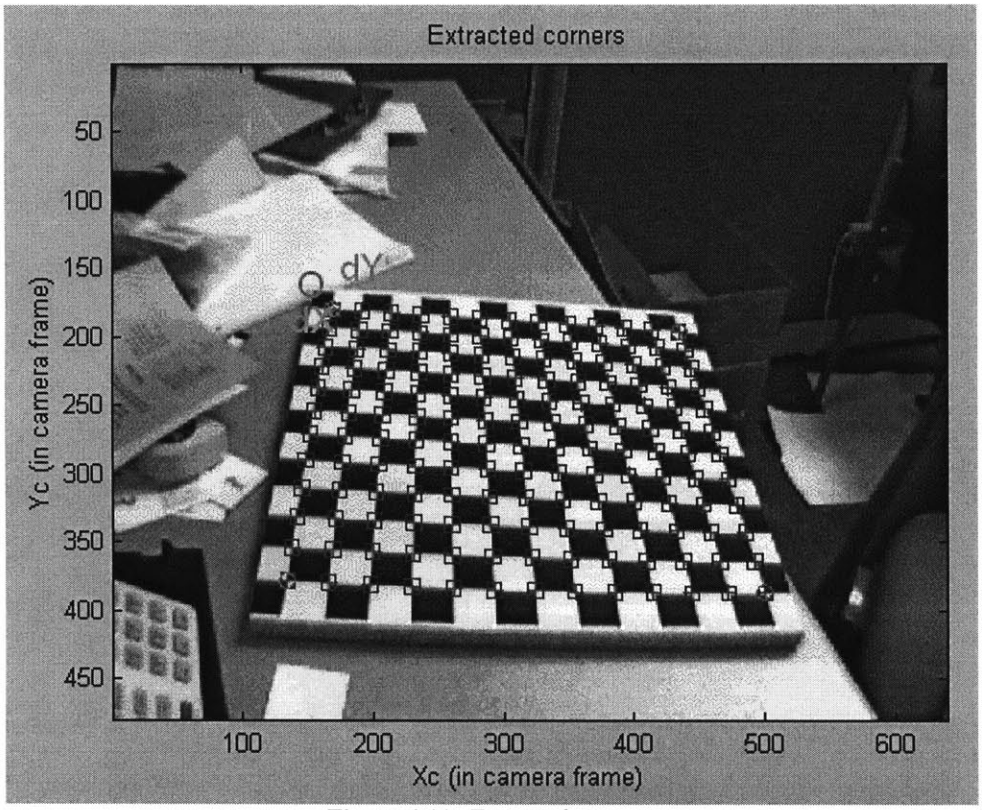
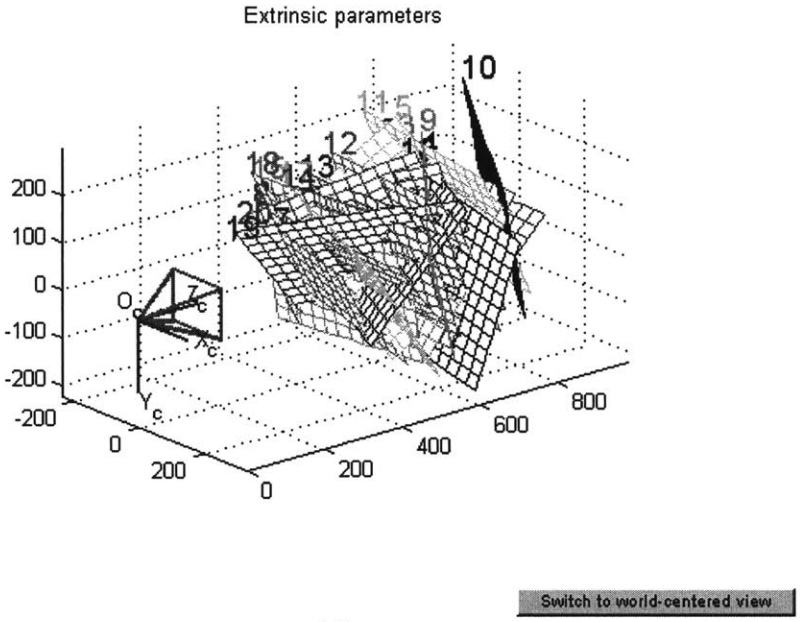
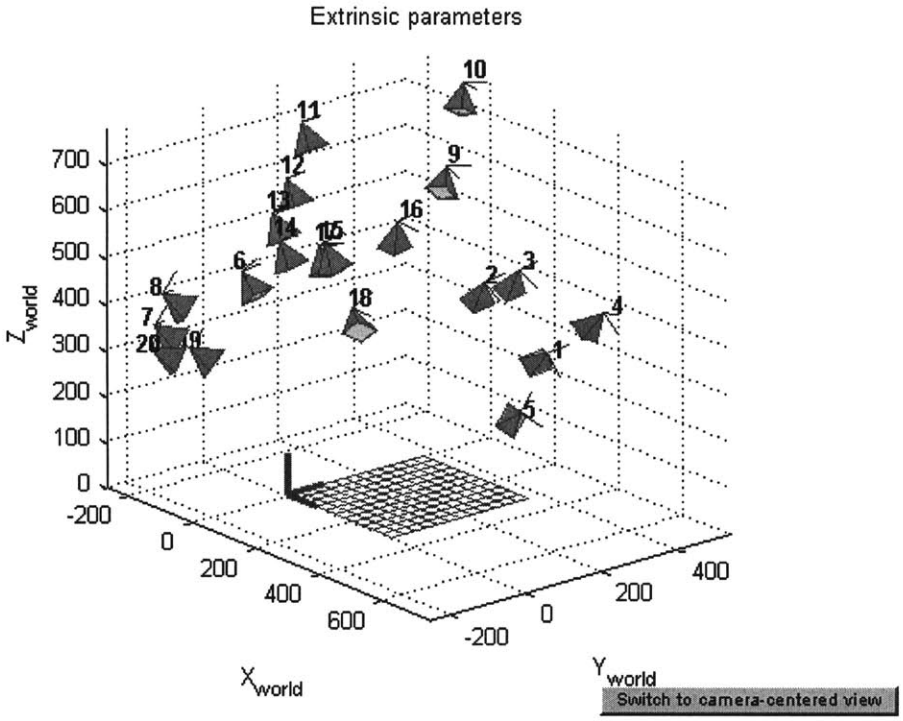


Figure 4.11. Extracted corners.

Also, with each camera a lens distortion model can be generated as seen in Figure 4.13. Part (a) shows the combined effect of radial and tangential distortion on each pixel of the image. Each arrow represents the effective displacement of a pixel induced by the lens distortion. Observe that points at the corners of the image are displaced by as much as 25 pixels. Part (b) shows the impact of the tangential component of distortion. On this plot, the maximum induced displacement is 0.14 pixel (at the upper left corner of the image). The third plot (c) shows the impact of the radial component of distortion. This plot is very similar to the full distortion plot, showing the tangential component could very well be discarded in the complete distortion model. On the three plots, the cross indicates the center of the image, and the circle the location of the principal point.



(a)

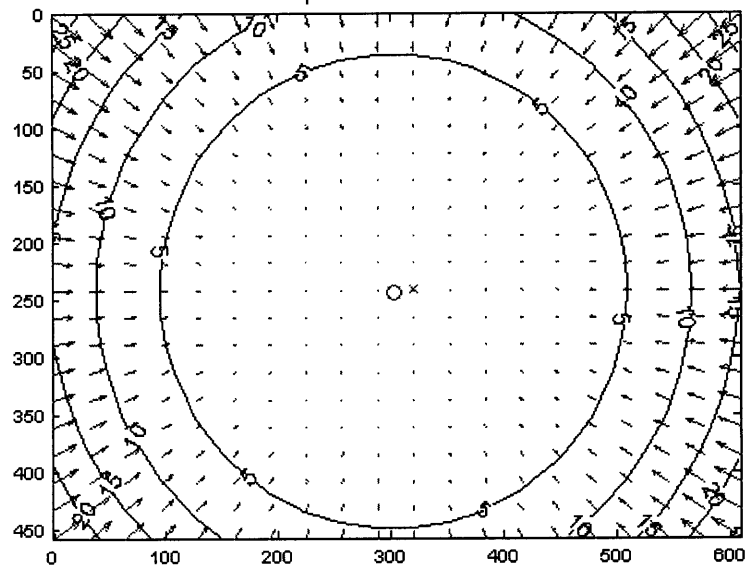


(b)

Figure 4.12. Visualization of camera and target positions.

(a)

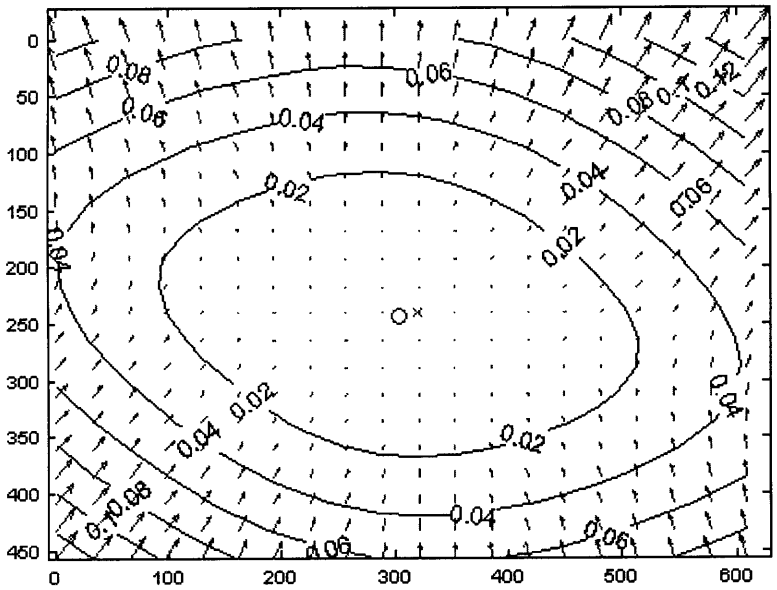
Complete Distortion Model



Pixel error	= [0.1174, 0.1159]	
Focal Length	= (657.303, 657.744)	+/- [0.2849, 0.2894]
Principal Point	= (302.717, 242.334)	+/- [0.5912, 0.5571]
Skew	= 0.0004198	+/- 0.0001905
Radial coefficients	= (-0.2535, 0.1187, 0)	+/- [0.00231, 0.009418, 0]
Tangential coefficients	= (-0.0002789, 5.174e-005)	+/- [0.0001217, 0.0001208]

(b)

Tangential Component of the Distortion Model



Pixel error	= [0.1174, 0.1159]	
Focal Length	= (657.303, 657.744)	+/- [0.2849, 0.2894]
Principal Point	= (302.717, 242.334)	+/- [0.5912, 0.5571]
Skew	= 0.0004198	+/- 0.0001905
Radial coefficients	= (-0.2535, 0.1187, 0)	+/- [0.00231, 0.009418, 0]
Tangential coefficients	= (-0.0002789, 5.174e-005)	+/- [0.0001217, 0.0001208]

(c)

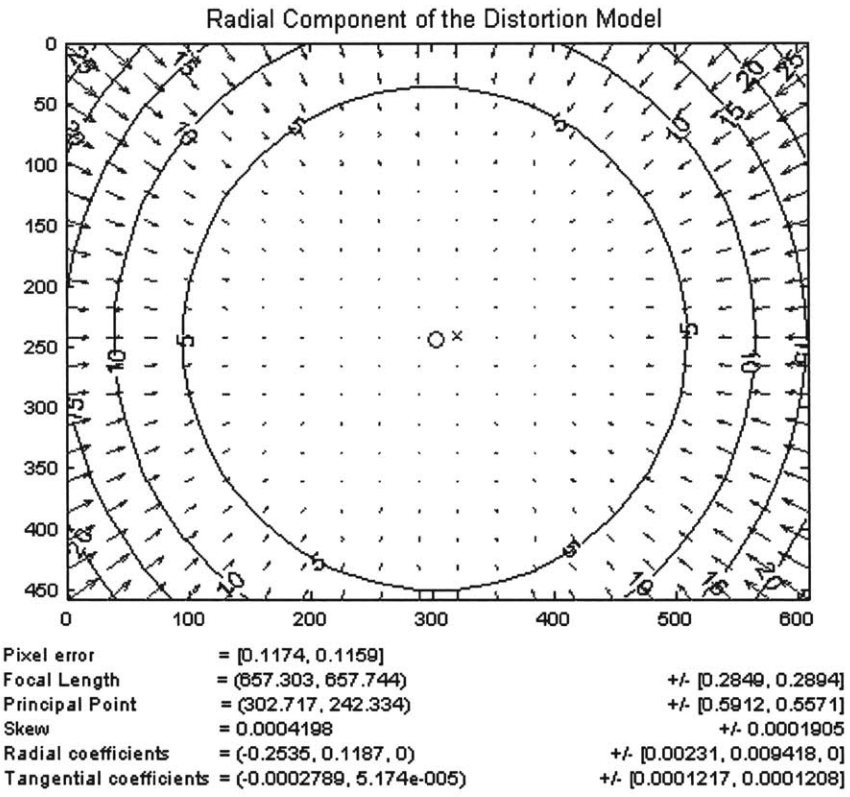


Figure 4.13. Radial and tangential lens distortion.

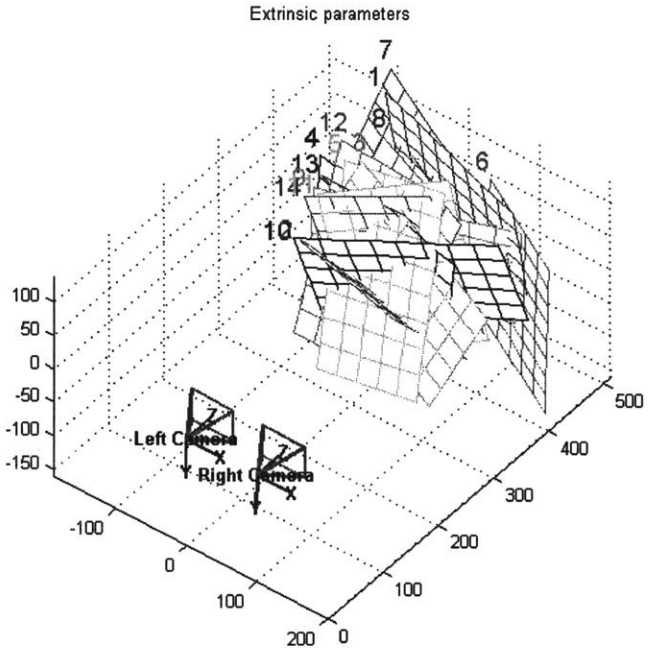


Figure 4.14. Visualization of stereo camera and target location.

After all the cameras are calibrated then a stereo-camera calibration can take place using the output files with the intrinsic and extrinsic profiles of each camera to each point in the calibration image. The output files for each camera are loaded into Bouguet's stero GUI and each point in correlating images is compared and used to create extrinsic parameters of the relative rotation and translation position to each other. Again, a correlation of the cameras can be seen in Figure 4.14.

To choose the two cameras requires picking a center camera to compare all other cameras in relative location. Since our camera array is small it is considered ideal to choose the camera in the middle. If the array was infinitely large and calibration could be performed much quickly then the center camera would be chosen that had a parallel view of the most important feature on an object or scene. The extrinsic parameters between the two cameras are recorded into a data file that is accessible to the SISCA's program.

4.3.2 Camera Alignment

The next step in camera calibration is to align all the cameras to the same height by zeroing the T_y translation vector in relation to the center camera. This is done for all cameras and done with precision servos. Once the cameras are aligned vertically, the user will pick the scene to be captured. For now the choices are limited to aerial, facial recognition, or surgical scenes. In each case a convergence point D_o and far point D_m are chosen. After the user chooses one of the two 3D displays, t_c can be calculated. Then the T_x translation vector for each camera in relation to the center camera can be compared and adjusted along the baseline to match t_c. The final step in camera alignment is converging all the cameras but the centered camera on the object of interest. The convergence point D_o is known so is t_c so the angle of rotation is calculated by simple trigonometric equations.

4.3.3 Recording

Placing the cameras in the correct position and orientation was the first crucial step. This is the last stage in which scene data is recorded from one or more cameras. In the recording phase the output must correct for geometric distortions, image rotations, and minor enhancements for the 3D display.

To preserve space, especially for hypostereo scenes the cameras were rotated vertically and to get the proper orientation requires an inverse rotation in the image processing stage. The second part of the recording phase involves correcting for lens distortion. Using image-processing techniques in MATLAB the corrections are made easily on the AVI image matrix using equations 2.33 and 2.34. The final step involves the correction for keystone distortion, which can be solved by multiple methods as discussed below. In all cases the image is up-sampled by four times using a bilinear interpolation. The final image is down-sampled back to the original image width and height.

4.3.3.1 McAllister

In this case the coordinate system as shown in Figure 2.6 shows *e*, which is not the viewer but the object. While the *u* and *v* axes are the cameras separated from the center camera converged towards the object. Each pixel in each camera is represented by the width *w* and height *h*. The reverse projective warp can be found using equation 2.21-2.22 for the case when *v* equals zero.

This assumes that in the camera alignment stage all the cameras are at approximately the same height on the v-axis allowing that height to be normalized to zero. The new coordinates are located in the up-sampled image by multiplying the new coordinates by the up-sampled rate of 4.

4.3.3.2 Projective Transformation

This transformation is a creation used by MATLAB when the scene appears tilted. Straight lines remain straight, but parallel lines converge toward vanishing points. To use this feature 4 homologous points that can be found in both a base image and the test image. In this case we can use the image from the center camera as the base and all the images from the other cameras as the test images. Rather than use extensive object recognition processing to find 4 homologous points, one can use the McAllister method previously but instead of calculating the projective warp for all points it can be reduced to 4 points. The four most extreme points are the image four corners of both images.

4.3.3.3 General Method 1

In this method one can determine the amount of rotation incurred from the cameras surrounding the center camera and use this convergence angle to determine the spread of keystone distortion across the image. Depending on the direction of scaling one can use equations 2.18 and 2.19 to compensate for the scaling effect. Another way to calculate the shear in both directions is take the 4 most extreme points as in the projective transformation model.

4.3.3.4 General Method 2

Another method again uses the convergence angle from the cameras or the 4 most extreme points and uses this information to determine a mask that cuts off the extreme portions of any keystone effects usually found on the edges of most images. This is all done on the up-sampled image such that much of the original image is preserved as it is down sampled. This is also the quickest method and provided if not as good or better results than the other three-keystone correction methods.

4.3.3.5 OpenGL

The more graphical way to approach the keystoning problem is using a software library like OpenGL. With OpenGL the keystone distorted image is texture-mapped onto a rectangular polygon. Then the polygon is twisted appropriately to match the capture toe-in angle, the result is an undistorted image on the polygon. Depending on the program, the image might be cropped by the keystone-shaped frame of the polygon. Additional hardware support such as MIP maps and anisotropic filtering can minimize or eliminate aliasing that might occur through the distortion process.

4.4 Vertical Disparity Limits

The vertical disparity permissible before correction for the head mounted stereo display is 3.5mm with a view distance of 2m and for the view sequential display the maximum disparity is 2.2mm with a view distance of 1.25.

The vertical disparity between two cameras in each of the three environments is shown in Figure 4.15. The figures are based on one camera maintaining a parallel fixation while the camera to the right of the object rotates. If the camera were to the left the figures would be similar since

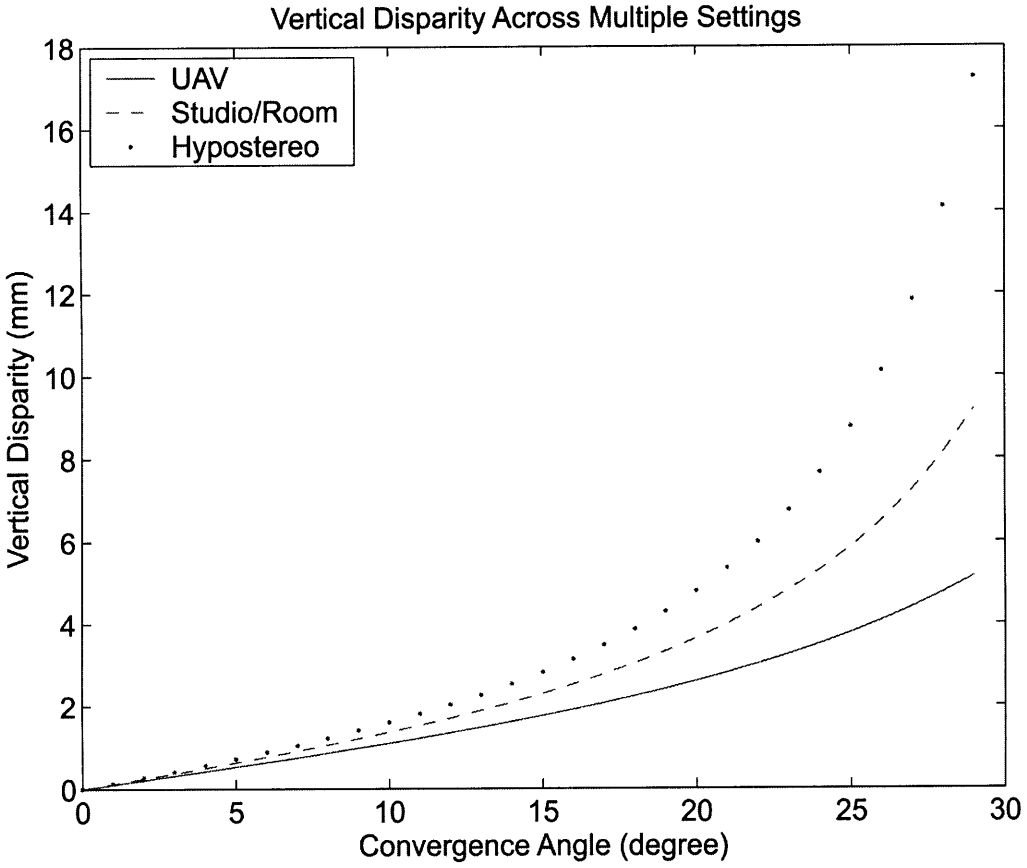


Figure 4.15. Vertical Disparity Limit.

the x-axis would be negative and would thus be reflected in the equation 2.17 used to calculate the disparity. Since one camera is fixed the angle is zero. Plus, the lenses in the current system have an angle of view of 85 degrees or a half angle of view (HOV) of 42.5. The maximum vertical parallax occurs at the edges. The object point $P(x_o,y_o,z_o)$ is place on the farthest edge of the centered camera for the maximum disparity. The parameters t_c and z_o are based on each of the environments as seen in Table 4.6.

Table 4.6. Parameters for calculating vertical disparity (units in mm).

	UAV	Studio/Room	Hypostereo/Surgical
t_c	15240 mm (50ft)	160 mm	45 mm
z_o	1,000,000 mm	500 mm	80 mm
Head Mounted Stereo Angle of Acceptable Parallax	24.0 °	19.5 °	17.0 °
View Sequential Display Angle of Acceptable Parallax	17.6 °	14.5 °	12.6 °

Table 4.6 also shows the angle of acceptable parallax before fusion breakdown based on both three-dimensional displays and the three environments. The aerial environment has fewer problems with vertical disparity since objects are much further away. The vertical parallax is more pronounced as the baseline and convergence point are reduced and also on the view sequential display in comparison to the head mounted display.

4.5 Performance

Stereoscopic distortions are ways in which a stereoscopic image of a scene differs from actually viewing the scene directly. These distortions, depending on how badly they are presented and dependent on each individual viewer, will ultimately decide how good the entire capture/display system works.

The first inherent problem starts with depth plane curvature. Woods (1993) shows a good comparison of image/display projection between a parallel camera configuration and a toed-in configuration. The toed-in effect will cause a curvature in the depth plane such that objects at the corners of the image appear further away from the viewer than objects at the center of the image. Depth plane curvature is closely linked with keystone distortion. This is definitely a problem in all three testing environments since the cameras are toed-in.

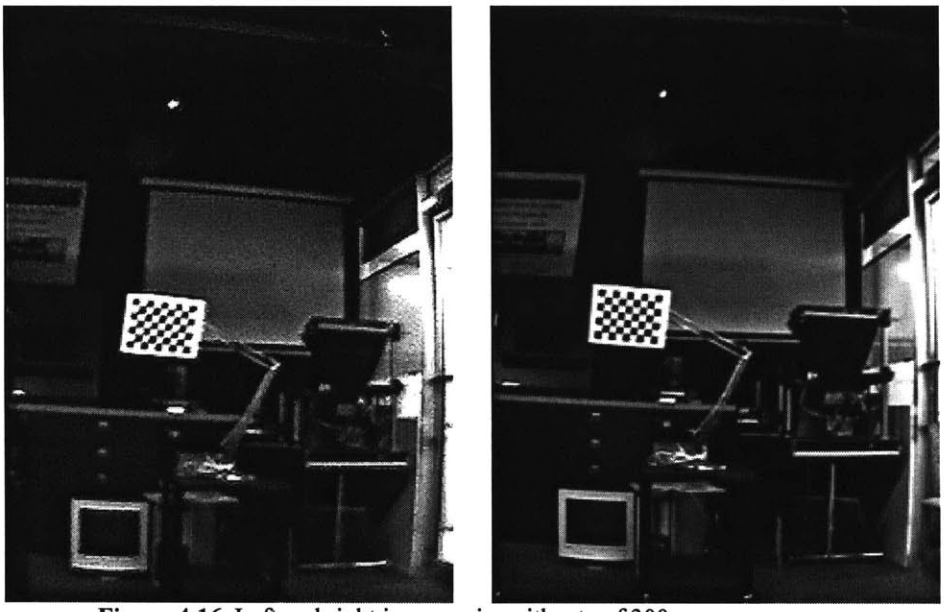


Figure 4.16. Left and right image pairs with a tc of 300mm.

The second problem is a non-linearity in depth since it leads to wrongly perceived depth. The depth will be stretched between the viewer and the monitor and compressed between the monitor and infinity. If the camera system were in motion as on the UAV then it could possibly lead to false estimations of velocity. For example, an object moving closer to a moving camera rig will appear to be moving faster. Another depth illusion that may cause problems is cardboarding in which a shift in that each flat image is shifted left or right such that they appear at different depths. The only way to create a linear relationship between image depth and object depth can only be obtained by configuring the stereoscopic video system such that object infinity is displayed at image infinity on the stereoscopic display.

With distorted images on a view-sequential display, as the viewer changes viewing location the image appears to follow the viewer or the object appears to be rotating or stretching. A sideways movement of the observer leads to a shearing distortion. Images out of the monitor appear to shear in the direction of the observer and images behind the surface of the monitor shear in the opposite direction. The end effect is that wrongly perceived relative object distances are a result from improper matching of image data and display characteristics.

A viewer's motion will also lead to the false perception of motion in the image. This effect is however less noticeable as the viewer moves away from the display. One minor effect is that when there is a non-linear relationship between image an object depth a mismatch between the depth and size magnification can lead to an image appearing flat or conversely stretched. In many cases where the viewer is an expert in their particular study/environment they will be able to compensate for non-linear relationships if they have a previous understanding of the object.

The biggest distortion problem encountered by cameras and displays is keystone distortion causing vertical parallax. The amount of vertical parallax is greatest in the corners of the image and increases with increased camera separation, decreased convergence distance and decreased focal length. In most cases there is also horizontal parallax, which is essentially the depth plane curvature.

General Method 2, which created a virtual mask to cut the borders of an upsampled image, proved to be the best all-around method for correcting keystone distortion in terms of speed, acceptable quality, and for small convergence angles. More precision, especially at large convergence angles, requires McAllister's method or the use of the MATLAB projective transformation. One problem that caused much of the parallax was aligning the cameras. This persisted even after camera calibration and knowing the precise extrinsic relationships of all the cameras to each other and the environment. The reason for the problem was that the smallest spatial movement by the servos was 0.36 degrees and only became more difficult if the cameras were not attached to the servos directly over the nodal point of the system or center of rotation.

Each testing environment experienced alignment issues especially as t_c increases. Thus the corners in aerial image were the most drastically effected. The hypostereo environment becomes increasingly worse as the convergence distance gets shorter. The best environment for recording and display were the studio/room environment where the cameras separation and convergence distance are in a mid-range between the extreme far and short distances needed for hyper and hypo stereo. Essentially, as the cameras move closer to the natural interaxial distance of the human eye is the output naturally displayed and viewed.

Artifacts from aliasing and sampling are also present during recording and after post-processing. The camera image sensor samples at the NTSC level of 640 x 480 resolution and will inherently result in aliasing at lower frequencies, much more than if the camera had a finer resolution as found on high-end cameras. The aliasing as seen most commonly by jagged edges only becomes worse as the image is upsampled with a bilinear kernel to remove the keystoning. Usually, the General Method 2 is used for the keystoning which crops the image vertically and occasionally

horizonatally. The downsampling is performed from the cropped image resulting in improved image quality and anti-aliasing.

Another artifact worth mentioning was chromatic differences between the cameras which can be associated with different spectral properties of each of the silicon detectors among many other factors including how the light is getting to the camera, especially from overhead fluorescent lighting and competing light sources of varying spectral characteristics. Due to processing time the chromatic differences weren't corrected for in post-processing. The result to the viewer looking at the images imposed little discomfort although the mixing of various hues was unnatural.

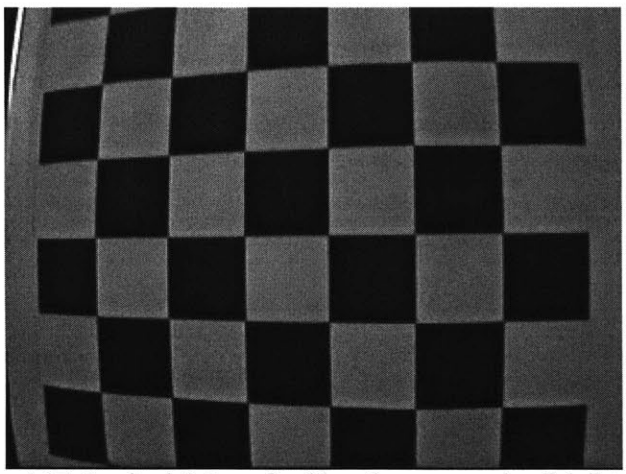
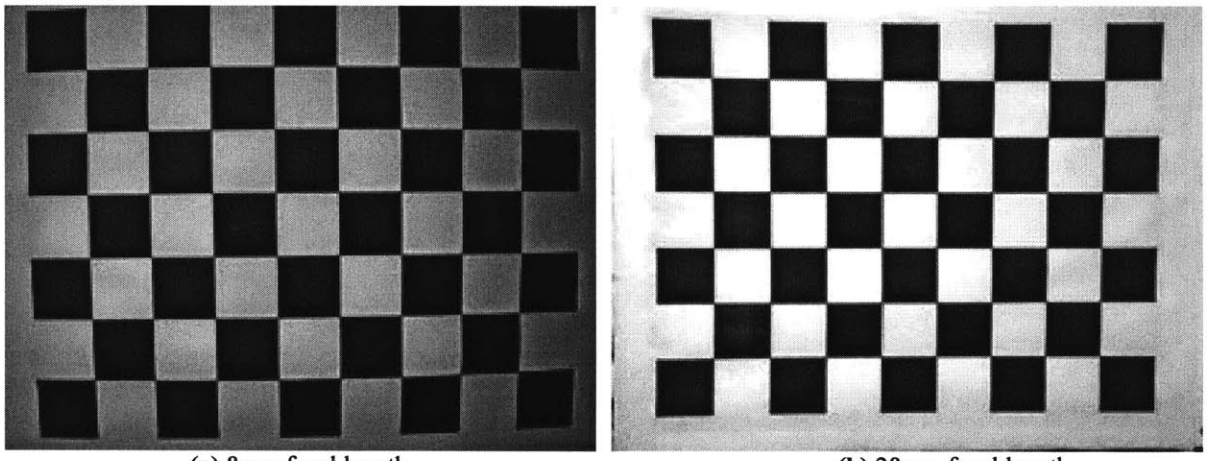


Figure 4.17. Logitech 3.5mm focal length camera and radial distortion.



(a) 8mm focal length **(b)** 28mm focal length

Figure 4.18. Nikon CoolPix 990 at varying focal lengths and radial distortion.

Lastly the other source of vertical parallax can occur from lens distortion. The amount of vertical parallax by a lens will depend upon the radial distance from the center of the lens, the amount of horizontal parallax the image possesses and the properties of the lens. Unfortunately,

the radial distortion is worst for short focal length lenses. This can be seen holding a grid pattern in front of the Logitech 3.5mm camera, Figure 4.17, and a Nikon CoolPix zoom lens at 8mm and 28mm, Figure 4.18.

If the camera array were being used for static scenes in one single environment then it would be ideal to use the parallel camera configuration. This eliminates the keystone distortion and depth plane curvature. However, under scaling conditions in real-time environments, the very precise and small (under 1mm) shifts of the sensor from the camera lens is impractical. It would be attractive to use shearing if it could be increased in speed and maintain accuracy possibly using hydraulic mechanisms with fast-switching microfluidics.

4.6 Summary

A scalable camera array for three-dimensional image capture and display was assembled from off-the-shelf hardware with the following results.

- A software controller was built which could calculate translation as well as rotate a set of cameras vertically and horizontally under programmatic and precise control.
- A GUI was built in which the user could specify the environment (Aerial, Studio, or Macro), the type of final 3D display (Stereo Headmount or View-Sequential). When a particular environment was chosen the convergence and far plane distances were kept constant. For each 3D display the parameters such as frame width, image size, and viewing distance were also kept constant.
- The GUI permitted a user to specify their own convergence and far plane distances along with vertical and horizontal control over the servos for focusing on particular objects in a scene.
- The system was used to capture scenes at three spatial scales in a mock aerial situation, a studio application that involved multiple view facial capture, and a mock surgical environment.
- Keystone (vertical disparity) correction algorithms and software was implemented before final scene output to address the major source of geometric distortion caused by a crossed lens imaging system.
- Camera calibration of the SISCA was performed using Bouguet's MATLAB calibration tools for finding the intrinsic and extrinsic properties of each camera in relation to all other cameras.
- Camera alignment was implemented with the calibration data working in conjunction with the software GUI and the user's scene selection and display selection to properly position the cameras using the horizontal and vertical servos.
- The performance of the system was analyzed by observing how the final image output looked based on image alignment and selection, keystoning, lens distortion, aliasing, and chromatic differences between sensors.

CHAPTER 5. FUTURE APPLICATIONS

This chapter will discuss the Spatial Imaging Scalable Camera Array in future applications, improvements in the array, and the additional features it will contribute with other rising imaging technologies.

5.1 Program Environment

Currently the software environment is constrained to parameters pre-input for an unmanned aerial vehicle, facial recognition in a studio, and macro surgical procedures. Chapter 4 gives a list of the constrained parameters for the convergence and far plane distances. There is an option for the user to enter in their own near and far plane distances. The user also has the ability, regardless of the scene to specify between a stereo display and multi-view display with the screen size, frame size, and viewing distance already specified. Additional features allow the user to rotate the cameras in the horizontal and vertical direction. The degree of movement is saved for future keystone correction and monitoring the camera locations since calibration.

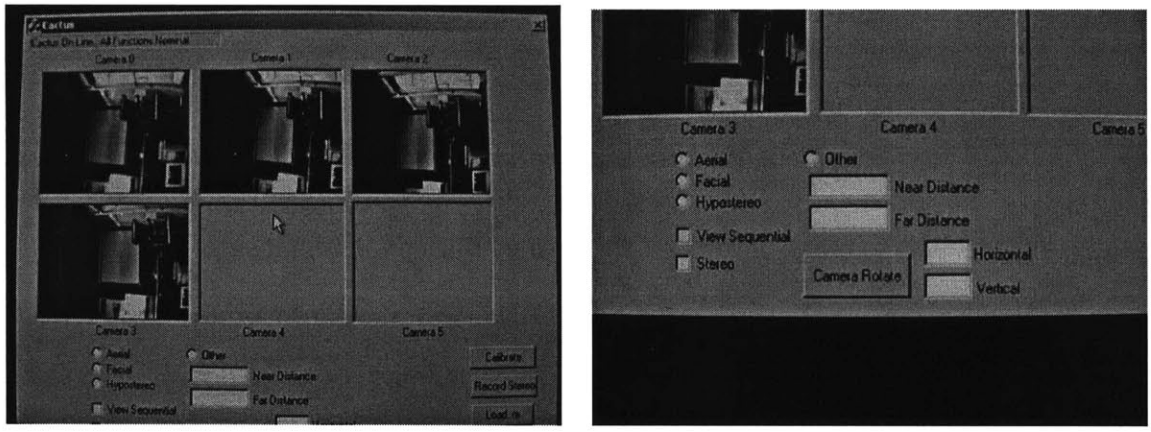


Figure 5.1. Images of the current SISCA software program.

The ideal case would entail future improvements that would allow the system to be completely autonomous unless the user wants direct control over the entire system. If this could be achieved then the program in autonomous mode would heavily rely on object recognition and tracking capabilities. Theoretically, the system would have the ability to distinguish between multiple environments and recognize pertinent objects of interest. Although difficult, the cameras would also be able to maintain focus on an object as it moves or the camera rig moves. A more realistic approach would allow user control by the use of a mouse to point directly to objects in the scene. The boundaries of these objects can be isolated and tracked with prediction-based movement algorithms.

When tracking objects in a scene, the software could estimate the near, convergence, and far planes of important objects. The cameras could additionally be on a micron based track system

for extremely precise camera separation thus allowing the program to coordinate all parameters for proper recording. The last piece of the program would allow the user to enter a database to specify their three-dimensional display. The user wouldn't have to enter any data since the database will have the necessary dimensions of all the systems.

Another area of extreme importance is the speed of capturing data from multiple cameras on one computer, which can only record from one camera at a time. The first step is to capture one frame from each camera's streaming frame buffer as quickly as possible. Ideally, this rate on a non-interlaced system should be 1 / (30 x Number of Cameras) frames per second. The second step is to implement a method that can compress stereo information in order to reduce redundant information temporally and spatially.

Lastly, creative, robust, and quickly applied calibration techniques are needed not only for reducing the efforts expended by a user but also for increasingly challenging object based rendering techniques for scene creation. Many of these object based applications are starting to use a large degree of randomness in terms of using multiple cameras that are not the same model, lie in unique positions from each other and from objects of interest, and record to different output devices.

5.2 Object Detection and Tracking

One of the key features of the future software program is object detection and tracking. It's relatively easy to have the user distinguish objects of interest in images but what about building vision systems that can tell where the object is and what type of object it is looking at as it moves through the world? The ideal situation is to have the software choose and recognize the most important object(s) for object tracking.

The problem is difficult and largely unsolved and tends to take a top down approach to exploiting visual content of a whole image. First, the scene is identified in which additional parsing allows for the identification of regions and if done correctly, for recognizable objects. The amount of image processing in these methods limits the accuracy when trying to work in real-world real time applications.

Object detection and tracking will require a combination of methods that work particularly well in a narrow range of environments. Combining these methods covers a larger range of difficulties that arise such as noisy backgrounds, a moving camera or observer, bad shooting conditions, and object occlusion. Much of the success of these consolidated solutions will be based on reliable assumptions and the ever-increasing computing power available (Azarbayejani, 1996).

Essentially, vision-based techniques can't be solely relied upon for tracking. My analysis of future object tracking involves radio frequency identification (RFID) tags which will allow users to acquire real-world location coordinates of objects carrying a low power microchip. Currently the microchips powered locally or globally can emit with their antennas data about the product that they are attached to up to several meters from a tag reader. That way in a studio scene one could locate and identify objects that have tags based on triangulation. A camera array could

potentially lock on to the real-world location and maintain a fixation on the product as the camera rig or object moves. This can occur even with occlusion by other objects.

One of the greatest areas to benefit from object tracking with multiple view image capture is for surgical procedures, especially telemedicine applications. Placing RFID tags on the surgical instruments would allow the cameras to focus on specific areas of operation and allow students viewing the procedures to view these areas from multiple views. This system in robotic telemedicine applications could allow distant surgeons to specifically see the problem with greater efficiency including any crucial and precise movement of the medical tools (Ladd, 2002).

5.3 Stereo Image Compression

MPEG is a successful compression scheme for motion video that exploits the high correlation between temporally adjacent frames. Several proposed methods have been created to exploit the high correlation between spatially or angularly adjacent still frames, left-right 3D stereoscopic image pairs for a number of different three-dimensional displays. If left-right pairs are selected from 3D motion streams at different times but appear to have similar perspective-induced disparity and motion-induced disparity then a high correlation will exist between these image pairs. This high degree of correlation allows for some form of compression for stereo video streams (Siegel, 1994).

In holographic displays the goal is to reduce the space bandwidth product by reducing vertical parallax. In stereo and multi-view displays the desired outcome is a compression method analogous to the one used to code the color in conventional television signals. In single image broadcast only a small fraction of the bandwidth suffices to graft a small but adequate amount of chromaticity information onto the luminosity. The goal is to figure out how to graft a small but adequate amount of disparity information onto a monocular video channel (Siegel, 1997).

There are four fundamental algorithmic components of a stereoscopic or multi-view compression scheme. The first algorithm is a conventional algorithm for coding one of the main views. The second is for constructing the disparity map or for constructing a function that predicts at every image point the disparity vector between it and the corresponding point in another image taken from a specified perspective offset from the current perspective. Thirdly, an algorithm is needed for coding and decoding such that the transmitted disparity map is more compact than the independent conventional coding of whatever subset of all the images are actually needed at the receiver. The last algorithm should compactly represent residual error between the predicted and original views. It is advantageous to integrate motion compensation with the disparity calculation and coding. Algorithms 2 and 3 are only unique to stereoscopic coding versus single image video coding.

The suppression theory suggests that a viewer who is given a blurry image for one eye and a sharp image for the other eye will still perceive a three-dimensional scene, although blurry. The first algorithm in coding the main view or mainstream uses a conventional format like MPEG. This stream is transmitted to one eye. A secondary stream called the auxiliary scene transmits a low-resolution disparity map that has been constructed at the transmitter from the left and right perspective images. The second eye's perspective is not transmitted. Instead the perspective is estimated by the receiver, from the resolution of the disparity map and by distorting the main

stream according to the directions encapsulated in the disparity map. Although this low resolution is unpleasant as a single image, it binocular fusion with the sharp mainstream image stimulates the stereopsis perception.

In stereo compression it is assumed that every world point is visible and can be identified from at least two perspectives. Yet occlusion and aliasing have the problem of no corresponding points and too many corresponding points respectively. Both occlusion and aliasing must be corrected with expensive ad hoc heuristic detection and repair algorithms that attempt to replicate the "reasonable assumption generating" machinery that the human brain invokes when faced with too little or too much data. As stated in Chapter 2, occlusion is considered to be the most important depth cue in the human visual system such that it is necessary to find a way to depict it when using compression schemes.

There has also been some relevant work in object-based compression. The idea is to locate specific features in a scene and based on their inherent behavior under certain geometry and dynamics will conform to a known model of movement or shape. If this can be predicted in reasonably quick manner compression could be reduced considerably spatially and temporally. This technique or idea is more suited for entertainment where absolute prediction is not nearly as critical as in other applications.

For the following two cases, I believe some of the setbacks can be reduced using RFID tags. Having RFID tags with real world coordinates allows for detecting occluded objects and interpolating their positions, allowing for a repair mechanism in the compression software. It's debatable if the RFID tag could help in aliasing. For object-based compression the RFID tags will one day be able to code information about an object's characteristics such as physical parameters or appearance that would easily relate to scene modeling. By quickly grabbing this information one could quickly compile a model for compression.

5.4 Image Based Rendering

Image based rendering (IBR) and the techniques to improve the output quality of computer-generated scenes is a growing area of research especially since the dramatic effect it has brought to cinematic special effects. IBR techniques are growing into areas of modeling based on random scenes from random cameras with little or no form of calibration. The purpose of such as system is to take multiple two-dimensional images and create a "fly through" presentation of a scene on demand.

Much of IBR involves the creation of new scenes by interpolation of real camera imagery. It appears that this new scene view comes from a virtual camera. Essentially, the software developer has a virtual camera array based on a few real cameras that is not only infinite in the horizontal and vertical directions but in the third dimension. Effectively, the idea of a virtual camera array opens up possibilities of scalability. However, such techniques require further calibration techniques and correlating virtual cameras which could essentially be made to be different in their characteristics (focal length, field of view, image sensor size) to one another.

The SISCA should not only be considered useful for three-dimensional displays but as a small subset of real image sensors in a much larger virtual camera array. Essentially it becomes its

own IBR device. One of the goals of the SISCA is to show scalability not only between multiple environments and varying three-dimensional displays but also in terms of IBR applications. Secondly, the SISCA can be considered a generalized multi-image collector such that an array of imaging devices, not necessarily cameras can be used for three-dimensional capture.

For example, IBR is becoming hugely popular in non-invasive surgery by creating panoramic "fly through" scenes of particular anatomy. One particular type of device being used to scan and analyze very small areas in the human body is an endoscopic imaging device that uses fiber bundles. Imagine using one or more single fiber bundles as their own individual image sensors or essentially as individual cameras located spatially in the endoscopic device to create three-dimensional IBR scenes without the panning and rotation. Essentially, the system is a mini-SISCA for image based rendering output.

5.5 Integral Imaging

As mentioned in Chapter 3.3.2.3.2, *integral imaging* is a refraction-based approach for direction-multiplexed displays. Both the capture and display aspect of integral imaging rely on a lenslet sheet for 3D deconstruction and reconstruction respectively. Although, the display aspect is important, there is a greater interest in how the integral imaging capture can be associated with the SISCA.

Integral imaging has several disadvantages compared to holography or stereo techniques in that the resolution, viewing angle, or depth is considerably lower. The resolution in integral imaging is determined by the lenslet size, the resolution of the CCD and the display device, lenslet aberrations and misalignment. Additionally, resolution is limited by the pitch of the lenslet array, which effectively determines the spatial sampling rate of the ray information in the spatial domain.

One method to increase the resolution or spatial sampling rate involves time multiplexing. The idea is to rapidly vibrate in synchronization the lenslets in a lateral direction during both scene capture and display. Further analysis has recently shown that it may be preferable for circular movement of the array. The only stipulation when using either movement is that the vibration occurs within the flicker-fusion frequency of the human eye.

In a typical integral imaging system the resolution and image depth must be balanced, improving one lowers the quality of the other. A possible solution to maintain a high spatial resolution with and increase the depth of focus is to use an array of lenslets with varying focal lengths and aperture sizes in both the capture and display of an object. A system that uses this technique along with a time multiplexing movement is discussed by Jang and Javidi (2004).

Essentially, the SISCA in its prior form was a much larger integral imaging system for capturing static scenes for lenticular display. The main difference to its current form is that to keep the object of attention centered on each camera involved shearing the lenses from the image sensors. Then in software specific columns from each camera are stitched together into one image. When a vertical lenslet array (lenticular sheet) is placed over the image the effect is a three dimensional image to the viewer.

Hypothetically the SISCA in its current form could be a much more interesting scaled integral imaging system. Previously, I mentioned the SISCA being scaled downward into a multi-image collector for endoscopic three-dimensional imaging. Now I'm proposing scaling upward from a large array of small lenslets used in integral imaging to using a smaller number of larger lenses with a much finer sampling resolution behind each lens using a CMOS sensor with 640x320 pixels. In general, this increase in physical lens size and resolution would possibly eliminate the need for any type of lateral or circular movement.

Although the lenses will not be directly next to each other as in a typical integral imaging system, image interpolation and stitching can be used to capture over a much larger area, especially the larger lenses will have a wider field of view (FOV). If the lenses/image sensors must be moved they do have the advantage of being on vertical and horizontal servos. However, sending continual servo commands could reduce image capture speeds.

Also, the lenses could be modified to zoom lenses in that the focal length can easily be changed across cameras, thus keeping resolution and increasing depth of field. Another benefit of having zoom lenses is that the focal length can be changed at different camera locations rather than being static as in Jang and Javidi's scenario. Lastly, it should be noted that there has also been research conducted on compression of full parallax integral 3D-TV image data (Forman, 1997), which could be used to reduce the redundant information found from wide field of view camera overlap found on the SISCA.

5.6 Object Based Media

Within the MIT Media Lab is the Object Based Media Group working on individual electronic wall tiles that can easily be placed next to each other into a large horizontal and vertical array. Each tile has a LCD screen, a camera, a microphone, and an IR sensor. Although the cameras are equally spaced from each other they can still be made to think in a pattern wise fashion to capture stereo information from their current separation or from interpolated views between two tiles.

The most interesting part about the cameras on all the tiles is calibrating them separately and in relation to all other tiles and the communication between all the tiles to accomplish the calibration. Most likely the calibration would involve triangulation between nearest neighbors with this process continuing from row to row and column to column. If the tiles know their spatial location to the nearest tile neighbor as in left, right, up, or down, then the calibration can be performed much easier and perhaps without user assistance. Future forms of calibration of the SISCA could benefit from implementing a similar method of communicating with cameras embedded in an environment with some knowledge of location.

5.7 Conclusion and Future Work

An overview of the research in the design, testing, and analysis of the SISCA was presented in this paper. The following is a list of the accomplishments of the SISCA.

- A camera array rig holding 6 USB cameras was designed and implemented for the recording of active scenes across a multitude of environments which could be displayed on several different three-dimensional displays.

- A GUI was built for a user to control calibration, alignment, and rotation of the cameras using precision control horizontal and vertical servos.
- Image processing techniques were applied to provide the correct scene output from the cameras to either a Stereo Headmount or View-Sequential 3D display. Several distortions such as keystoning and lens distortion were corrected for in software.
- The camera array was tested and analyzed in three applications ranging from a mock aerial situation, multiple view facial capture in a studio, and a mock surgical environment.

This last chapter on future applications is meant for improving the functionality of SISCA, as well as the multi-image capture paradigm, and extending its operation to other applications beyond UAVs, facial recognition, and surgical applications. Below is a list of future directions.

- Further research and development of object recognition and tracking with possibly the use of RFID technology.
- Improved calibration techniques for multiple cameras with varying characteristics in known and unknown locations.
- Extending the SISCA capture device to several other applications such as robotic telemedicine and new image based rendering methods.
- Replicating the SISCA in other imaging devices for large-scale integral imaging and medical applications such as endoscopic devices.

REFERENCES

Abdel-Azis, Y.I, Karara, H.M. (1971). Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Proceedings of the American Society of Photogrammetry Symposium on Close-Range Photogrammetry, 1-18.

Ahmed, Y.A., and H. Afifi (1999). New stereoscopic system. *Proceedings of SPIE Vol. 3639.*

Ariyaeeinia, A. M. (1992). Distortions in stereoscopic displays. *Proceedings of SPIE Vol. 1669*: 2-9.

Azarbayejani, A. and A. Pentland (1996). Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. *In Proceedings of the 13th ICPR, Vienna, Austria: IEEE Computer Society Press.*

Baker, J. (1987). Generating images for a time-multiplexed stereoscopic computer graphics system. *Proceedings of SPIE: True 3D imaging techniques and display technologies,* 761:44-52.

Bakos, N., C. Jarvman, and M. Ollila (2003). Simplistic dynamic image based rendering. *Association for Computing Machinery, Inc.* NVIS, Linkoping University, Sweden.

Becker, S. C. (1997). *Vision-assisted modeling for model-based video representations.* Massachusetts Institute of Technology Media Laboratory PhD thesis

Bone, E. and C. Bolkcom (2003). Unmanned aerial vehicles: Background and Issues for Congress. Washington/Congressional Research Service. RL31872.

Bourke, Paul (http://astronomy.swin.edu.au/~pbourke/).

Bowers, B. (2001). *Sir Charles Wheatstone FRS: 1802-1875.* Institution of Electrical Engineers in association with the Science Museum, London.

Brewster, Sir David (1856). The stereoscope: its history, theory, and construction, with its application to the fine and useful arts and to education. J. Murray, London.

Bulthoff, Isabelle, Heinrish Bulthoff, and Pawan Sinha (1998). Top-down influence on stereoscopic depth-perception. *Nature Neuroscience*, 1(3): 254-257.

Carson, K. (1984). *A color spatial display based on a raster framebuffer and varifocal mirror.* Massachusetts Institute of Technology Visual Studies MS thesis.

Chang, K.I, K.W. Bowyer, P.J. Flynn (2003). *Face recognition using 2D and 3D facial data.* Workshop in Multimodal User Authentication, 25-32.

Cossairt, O. (2003). *A view sequential display*. M.S. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Cull, E.C., D.P. Kowalski, J.B. Brchett, S.D. Feller, D.J. Brady (2002). Three-dimensional imaging with the Argus sensor array. Proceedings of SPIE, Vol. 4864: 211-222.

Ezzat, T. and T. Poggio (1996). Facial analysis and synthesis using image-based models. *In Proceedings 2nd International Conference on Automatic Face and Gesture Recognition.* .

Forman, M.C. and A. Agoun (1997). Compression of full parallax integral 3D-TV image data. *Proceedings of SPIE*, Vol. 3012: 222-226.

Fuchs, H., G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S.W. Lee, H. Farid, and T. Kanade (1994). Virtual space teleconferencing using a sea of cameras. *In Proceedings of the First International Symposium on Medical Robotics and Computer Assisted Surgery*, Pittsburgh, PA.

Glausier, C.A. (1992). Desert storm UAV lessons learned database. Booz, Allen & Hamilton, Inc, Washington, D.C.

Goshtasby, A. and W.A. Gruver (1993). Design of a single-lens stereo camera system. *Pattern Recognition* 26(6), 923-927.

Gunzberg, J. (1953). The story of natural vision. *American Cinematographer*, 34:534-35, 554-56, 612-16.

Halle, M.W. (1997). *Multiple viewpoint rendering for 3-Dimensional displays*. Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Bove, V.M. Jr. (1989). *Synthetic movies derived from multi-dimensional image sensors*. Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Heseltine, T., N. Pears, J. Austin, and Z. Chen (2003). Face recognition: A comparison of appearance-based approaches. *Proc. VIIth Digital Image Computing Techniques and Applications.*

Horn, B., and M.J. Brooks, Eds. (1989). Shape from shading. MIT Press, Boston.

Howard, I.P., and B.J. Rogers (1995). *Binocular vision and stereopsis*. Oxford University Press, New York, NY.

Iwasaki, M. (1999). 3D image pick-up apparatus. JP Patent 2000227332. 23rd July 1999.

Jang, J., and B. Javidi (2004). Time-muliplexed integral imaging for 3D sensing and display. *Optics & Photonics News* 15(4): 36-43.

Jia, Y., Y. Xu, W. Liu, C. Yang, Y. Zhu, X. Zhang, and L. An (2003). *A miniture stereo vision mahine for real-time dense depth mapping*. Springer-Verlag Berlin Heidelberg.

Julesz, B. (1977). Recent results with dynamic random-dot stereograms. *SPIE*, Vol. 120, Three-dimensional imaging.

Kanade, T., A. Yoshida, K. Oda, H. Kano, and M. Tanaka (1996). *A stereo machine for video-rate dense depth mapping and its new applications.* In IEEE Conference of. Computer Vision and Pattern Recognition, 196-202.

Kawakita, M., K. Iizuka, T. Aida, H. Kikuchi, H. Fujikake, J. Yonai, and K. Takizawa (2000). Axi-vision camera: a three-dimension camera. *Proceedings of SPIE*, Vol. 3958: 61-70.

Kroeker, K.L. (2002). Graphics and security: Exploring visual biometrics. IEEE Computer Graphics & Applications 22(4): 16-21.

Ladd, A.M., K.E. Bekris, A. Rudys, G. Marceau, L.E. Kavraki, and D.S. Wallach (2002). *Robotics-based location sensing using wireless ethernet. MOBICOM '02*, Atlanta, GA.

Levoy, M., and P. Hanrahan (August 1996). Light field rendering. Proc. ACM Conference on Computer Graphics (SIGGRAPH'96), New Orleans, 31-42.

Lin, Y. and C. Fuh (2000). Correcting distortion for digital cameras. *Proc. Natl. Sci. Counc. ROC(A), 24(2)*: 115-119.

Lipton, L (1982). *Foundations of the stereoscopic cinema.* Van Nostrand Reinhold Company Inc. New York, NY.

Lo, A.K.W. and K. Lao (1995). Single-lens, multi aperture camera. WO Patent 95/30928. 5th May 1995.

Lucente, M. (1994). *Diffraction-specific fringe computation for electro-holography.* Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Matsunaga, K. (1998). New stereoscopic video and monitor system with central high resolution. *Proceedings of SPIE*, Vol. 3295: 164-170.

McAllister, D.F. (1993). *Stereo computer graphics and other true 3D technologies.* Princeton University Press: Princeton, NJ.

McAllister, D.F. (1994). Digital correction of keystoning caused by image translation. *Proceeding of SPIE*, Vol. 2177: 97-107.

McMillan, L (2004). *Image-based rendering using image-warping – motivation and background.* LCS Computer Graphics Group, MIT.

Melzer, J.E. and K. Moffitt (1997). *Head mounted displays designing for the user.* McGraw-Hill, New York, NY.

Mochizuki, A. (1992). Image pick up device. JP Patent 6197379. 28th May 1992.

Montgomery, D., C. K. Jones, J.N. Stewart, and A. Smith (2002). Stereoscopic camera design. Proceedings of SPIE, Vol. 4660: 26-37.

Naemura, T., J. Tago, and H. Harashima (2002). Real-time video-based modeling and rendering of 3D scenes. *IEEE Computer Graphics and Applications*, 22(2).

Naemura, T. and H. Harashima (2000). Ray-based approach to integrated 3D visual communication. *SPIE*, Vol. CR76.

NASA Unmanned Aerial Vehicles, Wallops Flight Facility, Goddard Space Flight Center (http://uav.wff.nasa.gov/).

Norling, J.A. (1953). Stereoscopic Camera. 2,753,774. Ap. Feb. 12, 1953; Pat. Jul. 10, 1956.

Nyland, L., A. Lastra, D.K. McAllister, V. Popescu, C. McCue (2001). Capturing, Processing and Rendering Real-World Scenes. Proceedings of SPIE, Vol. 4309.

Ovsjannikova, N.A., and Slabova, A.E. (1975). Technical and technological principals "Stereo 70". *Teknika Kino I Televidenia*, 16-26.

Pastoor, S. and M. Wopking (1997). 3-D displays: A review of current technologies. *Displays* 17: 100-110.

Plesniak, W. (2003). Incremental computing of display holograms. *Journal of Optical Engineering* 42(6) 1560-1571.

Priyantha, N.B., Anit Chakrabory, and Hari Balakrishnn (2000). The cricket location-support system. *6th ACM International Conference on Mobile Computing and Networking*, Boston, MA.

Paul, E. (1997). Method and apparatus for producing stereoscopic images with single-sensor. US Patent 5883695. 19th Sept 1997.

Rohaly, J. and D.P. Hart (2000). High resolution, ultra fast 3D imaging. Three-Dimesnional Image Captureand Applications III. *Proceedings of SPIE*, Vol. 3958, 2-10.

Rule, J.T. (August 1938). Stereoscopic Drawings. *Journal of Optical Society of America*, 28: 313-322.

Saunders, B.G. (1968). Stereoscopic drawing by computer – Is it orthoscopic? Applied Optics, 7(8): 1499-1504.

Sekida, M. (1993). Three-dimensional image pick-up system in photographic or video camera. JP Paten 7152096. 30th Nov 1993.

Shao, Juliang (2002). Generation of temporally consistent multiple virtual camera views form stereoscopic image sequences. *International Journal of Computer Vision*, 47(1/2/3): 171-180.

Shimizu, E. (1999). Stereoscopic image system. JP Patent 2000298320. 13th April 1999.

Siegel, M., P. Gunatilake, S. Sethuranman, A. Jordan (1994). Compression of stereo image pairs and streams. *Proceedings of SPIE*, Vol. 2177, 258-268.

Siegel, M., S. Sethuranman, J.S. McVeigh, and A. Jordan (1997). Compression and interpolation of 3D-stereoscopic and multi-view video. *Proceedings of SPIE*, Vol. 3012, 227-238.

Siegel, M., Yoshikazu Tobinaga, and Takeo Akiya (1998). Kinder Gentler Stereo. *IS&T/SPIE '98 3639A-03*.

Smith, P.H. (1998). Imager for the Mars pathfinder (IMP): a multispectral stereo imaging system. *Proceedings of SPIE*, Vol. 3295, 4-9.

St.-Hilaire, P (1994). *Scalable optical architectures for electronic holography*. Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Tam, W.J., L.B. Stelmach, and P. Corriveau (1998). Psychovisual aspects of viewing stereoscopic video sequences. Proceeding of SPIE, Vol. 3295, 226-237.

Telemedicine for the Medicare Population. Summary, Evidence Report/Technology Assessment: Number 24. AHRQ Publication Number 01-E011, February 2001. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/clinic/epcsums/telemedsum.htm

Thompson, R.L, I.D. Red, L.A. Munoz, D.W. Murray (2001). Providing synthetic view for teleoperation using visual pose tracking in multiple cameras. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Hu*mans, 31(1): 43-53.

Tsai, R.Y. (1987). A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Journal of Robotics and Automation, RA-3 (4): 323-344.

Wolf, W., B. Ozer, and T. Lv (2003). Architectures for distributed smart cameras. *Proceedings of the International Conference of Multimedia and Expo*, July 6-9, Vol. 2.

Woods, A., T. Docherty, and R. Koch (1993). Image distortions in stereoscopic video systems. *Proceedings of SPIE*, Vol. 1915, 36-48.

Zhang, Z. (2002). *A flexible new technique for camera calibration*. Technical Report MSR-TR-98-71, Microsoft Research, Redmond, WA.