Digital Stereo Video: display, compression and transmission

Rhys Hawkins

A thesis submitted for the degree of
Master of Philosophy at
The Australian National University

February 2002

Typeset in Palatino by TEX and LATEX 2$_\varepsilon$.

Except where otherwise indicated, this thesis is my own original work.

Rhys Hawkins
22 February 2002

Acknowledgments

I would like to thank Professor Rod Boswell and Dr Henry Gardner for giving me the opportunity to change direction in my life. Both have made a significant contribution to my project and have my deepest thanks.

I would also like to thank Peter Alexander for his invaluable help in his many fields of expertise.

Finally, I would like to thank all those at RSPHYSSE who provided such a pleasant work environment.

Abstract

Digital stereo video is a digitised stream of stereo image pairs which, when observed by a human with binocular vision, conveys a great deal of information about a particular scene. The main benefit over traditional digital video is that with stereo, depth in an image can be perceived with greater accuracy.

With the growth of the Internet in the last decade, we are beginning to have enough bandwidth for the transmission of high quality digital video across the world in near real-time. This growth is set to accelerate, and the quality and feature set of digital video will subsequently increase. It is inevitable that the transmission of digital stereo video across the Internet will be commonplace in the near future.

This dissertation presents several techniques that can be applied now to facilitate the production, transmission and display of digital stereo video. The results from different implementations indicate that many goals can be achieved cheaply with commodity components, and more importantly, that high quality digital stereo video can be transmitted over the Internet with bandwidth and processing power not much greater than available today.

Contents

Chapter 1

Introduction

Background material and various definitions are given to help the reader understand subsequent chapters. The motivation and objectives are given in detail.

1.1 Overview

The intent of this thesis is to investigate various aspects of the display, compression and transmission of digital stereo video. The findings of these investigations are then used in the construction of the Robot Command Station (RCS), a stereo virtual reality system for the remote control of a mobile robot.

Stereo video contains time synchronised video, from two slightly displaced viewpoints, which emulates human binocular vision. By using stereo video, as opposed to mono video, it is hoped to give an observer a measure of the depth in the scene and an impression that they are actually present where the mobile robot is located. Such systems are commonly called tele-presence systems, and are the subject of much research at present. It has been claimed that operator performance is improved in such systems [23, 58, 49, 52, 51].

The virtual reality system was originally planned to be based on the "Wedge" two-screen virtual reality technology. This technology had been developed locally at the Australian National University using personal computer technology rather than high-end graphics workstations. As the Wedge is based on consumer technology, it is substantially cheaper than conventional virtual reality systems and has resulted in several Wedge installations over the past three years.

Hence, the initial stage in this project was the investigation into the transmission of stereo video, over the Internet, to a virtual reality display system with the express requirement that this stereo video should allow an operator to control a remote mobile robot. Subsequently, complex datasets, audio, and other related data can be added to enable collaborative research.

It was also hoped that, eventually, the various technologies developed in this thesis could contribute to the linking of virtual reality stations, via the Internet, for collaborative analysis of experimental data. Such a system would enable greater collaboration between distant research institutions where complex data sets are involved.

1.2 Background

This project has many different aspects to it, so before stating the objectives and motivations, a brief coverage of the concepts is necessary. The following sections have been kept brief as many of the points are expanded in the following chapters or described adequately in the references.

1.2.1 Computer Graphics

Soon after the introduction of computers, graphical displays were created using hard-copy printers and later, cathode ray tubes. In modern day computing, computer graphics is an incredibly broad topic, but we limit the discussion here to a small number of fundamental concepts.

Two dimensional computer graphics can be broken into two distinct categories, raster and vector graphics. In vector graphics, the image is described in terms of the objects in it and their respective parameters. Vector graphics can facilitate very high quality images which can be scaled without loss of information, and are often used in typesetting where high quality is required for diagrams. One disadvantage of vector graphics is that it is often difficult and inefficient to describe real-world scenery in this manner.

Raster graphics on the other hand specifies an image by a series of picture elements (pixels) arranged into a grid. A raster image is a sampled image, with its resolution limited by the number of pixels in the image. When a raster image is scaled, the scaled image must be interpolated which results in a loss of information. All computer displays use raster images to display the items currently on the computer screen, so that even vector graphics must be rendered into a raster image before they can be displayed.

Each pixel in a raster display is represented by a specified number of bits which may or may not contain colour information. Some common pixel representations are shown in Table 1.1. Most colour representations use the red, green, and blue tristimulus colours to specify each pixel. This is primarily because the most common colour display system is the cathode ray tube (CRT) which uses red, green and blue phosphor to illuminate each pixel.

For example, the True-Colour representation uses 8 bits for each tristimulus colour giving a total of 24 bits. Each of the 8 bits is used to give a relative intensity for each of the colours, ranging from 0 to 255. On the display system, these relative intensities are mixed to give a particular colour, in much the same way that different quantities of paint can be mixed to achieve different colours on a canvas except that displays use transmission rather than reflection to create colour.

A slightly different approach is used for Pseudo-Colour images. Instead of the 8 bits per pixel being split into a number of bits for the relative intensities of red, green, and blue, the 8 bits represents a colour number. This colour number is an index into a list of colour definitions, which give a high fidelity definition of the 256 colours. In images with low

Bits per Pixel	Palette Size	Colour	Common Name
1	2	···	Black and White
8	256	···	Monochrome or Greyscale
8	256	●	Pseudo-Colour
16	65536	●	Hi-Colour
24	16777216	●	True-Colour

Table 1.1: Common Pixel Representations

number of colours, Pseudo-Colour images can be quite useful at reducing the storage size for the image. However, for many complex real-world scenes, Pseudo-Colour images must use colour matching or dithering which degrade the quality of the image.

1.2.2 Motion Video Concepts

When motion video first became viable, a photograph was taken several times a second on a strip of light sensitive film. Through a chemical process, a film was made which could be subsequently displayed by projecting light through the film onto a projection screen. It was found very early, that for continuous motion to be perceived, approximately 24 photographs or frames per second were required. This is due to the persistence of vision of the human visual system which can be regarded as the rate at which the human visual system and brain can process visual information. The minimisation of the number of frames per second gives a considerable cost saving for the making of motion pictures.

When a motion film is displayed at 24 frames per second a high degree of flickering can be perceived. The flickering is caused by the period of darkness between successive frames of the film. A trivial solution to this is to display each frame twice at double the rate. In the human visual system, a darker environment reduces the eyes ability to detect flicker. In theatres with low light levels, the increased display frequency results in a much smaller amount of perceived flicker. Modern theatres operate at higher frame rates to reduce the flickering further.

With the advent of television, many similar problems were encountered. It was desired to minimise the number of frames per second, in order to reduce the amount of bandwidth required for the transmission of television signals. The number of frames per second in conventional television signals around the world varies between 25 and 30 frames per second depending on the particular standard. High Definition Television (HDTV) doubles these conventional frame rates through the use of digital transmission and compression.

Television uses a cathode-ray tube for the display of images. This operates by scanning an electron beam (or three beams for colour displays) across and down a phosphor screen. When the phosphor is struck with the electron beam, it emits light for a short period of time. Colour displays use three differently coloured phosphors in close proximity, which are perceived as a single colour at sufficient distance. As the brightness of the phosphor decreases quite rapidly (order of milliseconds), at 25 to 30 frames per second there is a substantial amount of flickering.

The method employed to reduce flickering on television displays is interlacing. When a television signal is received, each frame has a defined number of scan-lines and samples or pixels per scan-line. Interlaced television signals have special signaling that transmits all the even scan-lines followed by all the odd scan-lines. Hence, two separate images are displayed for each frame, and these images are commonly called the even and odd fields. So for normal television signals, the display rate is actually between 50 and 60 Hz, in which little flicker can be perceived.

One of the more practical reasons for having an interlaced CRT display is that it requires less bandwidth for higher refresh rates. This is why early computer monitors used interlacing, as this made various components, and therefore the entire monitor, cheaper and easier to manufacture. The disadvantage of interlacing is that the image is often of lower quality because a small amount of blank space can often be seen in between scan-lines on the display. This can be quite disconcerting, especially on large monitors that are viewed from relatively close range.

1.2.3 Projector Technology

Of the different projectors that are available, there are three different types which are distinguished by the method that they project an image onto a projection screen: cathode ray tube (CRT), liquid crystal display (LCD), and digital light processing (DLP) projectors. All of these different types of projectors have their own advantages and disadvantages.

CRT projectors have three separate cathode ray tubes for each of the tristimulus colours. An image is formed on each of the CRTs and is subsequently projected through a lens onto a projection screen. The main problem with CRT projectors having three separate CRTs is that these tubes have to be aligned to each other so that the image formed by the different colours coincide on the projection screen. This means that if the projector is moved then the three CRTs must be realigned. Whilst this is a tedious operation, it does mean that the image formed by the projector is highly adjustable allowing greater flexibility in the configuration of the projection system. Another side-effect of having three CRTs is that the overall size, weight and cost of the projector increases, although this does allow the image to be brighter.

LCD projectors use reflection off or transmission through a liquid crystal display to form an image. Since it is only necessary for one projection lens to be used, these devices are much smaller and more portable than CRT projectors. A disadvantage of LCD projectors is that they use polarisation of the light to control the relative intensity, so that the light emitted from LCD projectors is linearly polarised which in turn reduces the brightness of the projected image.

DLP projectors use reflection of light from a digital micro-mirror device (DMD), which is an array of small mirrors, each of which can be oriented at plus or minus a small angle (usually about 10^o). These mirrors are used to increase and decrease the amount of light that is projected out of the projector lens, the rest being absorbed by the projector's body. The relative intensities of the individual picture elements is controlled by the amount of time light is reflected through the lens. A rotating colour wheel, synchronised to the DMD, is used to provide colour to the projected image.

Both LCD and DLP projectors are designed to be portable and are quite compact. They are usually used for desktop presentations so that the image they project is higher than the projector. The image shape cannot be manipulated directly as it is formed in a solid state device, whereas CRT projectors illuminate a region of phosphor. This severely limits the positioning of LCD and DLP projectors with respect to the projection screen when distortion-free projection is required. So while LCD and DLP projectors are cheaper and easier to setup, they are not as flexible as CRT projectors.

1.2.4 Computer Networks

The defacto standard in computer networks is the Internet Protocol (IP)[1] and its derivatives. This protocol has existed in its present form since the early 1980s. In using the Internet as most people do, the basic Internet Protocol is never directly encountered. This is because this protocol is the foundation of several higher level protocols used on the Internet.

The Internet Protocol uses a 32 bit address to uniquely identify each computer connected to the network. This is often written as four bytes, e.g. `150.203.205.12`, where the numbers are often called the domain, sub-domain, network, and host respectively. Most people are more familiar with specifying a computer on the Internet by name, e.g. `www.anu.edu.au`, and this is converted to the numerical equivalent using the Domain Name Service (DNS) application which uses many servers that store databases mapping Internet addresses to names.

The most common way to connect computers to facilitate communication between them and sharing of information is via Ethernet. There are several forms of Ethernet with different connectors, cabling and speeds. The most commonly used Ethernet system is 10-base-T which can transmit data at 10 million bits per second, or 10 Mb/s [2].

Ethernet uses a Carrier Sense, Multiple Access, Collision Detect (CSMA/CD) protocol at hardware level. This means that multiple computers share the same network (Multiple Access) without a master controlling device. When a computer wishes to send data, it must check if the network is in use (Carrier Sense). It is possible for two computers to attempt to send a message at the same time, in which case the messages will collide. When this occurs (Collision Detect), both computers wait a random length of time before attempting to send their data again.

Each computer is connected to the network by an Ethernet adaptor which has a hardware address consisting of 6 bytes. This address is called a medium access control (MAC) address and is unique for each adaptor and therefore is used to uniquely identify each computer on

[1] More correctly, the Internet Protocol version 4 (IPv4)

[2] The convention used in this thesis is lower case b represents bits and upper case B represents bytes, where 8 bits equals 1 byte.

the network. The Address Resolution Protocol (ARP) maps an IP address to the hardware address, and the Reverse Address Resolution Protocol (RARP) does the reverse. Routers are used to control the flow of messages between different networks, and can often provide a wide variety of functions such as fire-walling for security. The Internet Control Message Protocol (ICMP) is used for communication between a host and a router and vice-versa. While this protocol is rarely used directly (the exception is when networks are being tested for connectivity) it forms an important part of the Internet Protocol suite.

Most applications that use the Internet Protocol do not use it directly. Several higher level protocols operate using the Internet Protocol, the most common of which are the Transmission Control Protocol (TCP), and the User Datagram Protocol (UDP). There are several examples of applications that use these protocols, the most popular of these, the World Wide Web, uses TCP for all transmission of data.

1.2.5 Human Binocular Vision

Humans, like many other animals, have two eyes through which they perceive objects in the world. Several reasons have been proposed as to why humans have two eyes, such as to increase our chance of survival by having a redundant eye, or to enable greater refinement of hand-eye coordination. However, the most generally accepted reason for humans having two eyes is that it provides an additional depth cue that facilitates spatial mapping of the external environment.

Recent physiological experiments [53] have shown that a significant portion of the brain is solely dedicated to the processing of binocular information. The main benefit of having a second eye, closely located to its fellow eye, is that it allows us to see two separate views of the world from slightly disparate positions, as shown in Figure 1.1. The human visual system automatically verges the two eyes on an object of interest. Since the object of interest will be in the centre of vision of both eyes, from a geometrical point of view it will have zero disparity. Objects behind the vergence point will have a positive angular disparity or un-crossed disparity, whereas objects in front have a negative angular disparity or crossed disparity. Thus, the disparity that can be measured by the brain for different objects in a scene can be used to gauge the relative distances of these objects from the observer. The mechanism for the measurement of the disparity is not fully understood. While we can show theoretically that relative distance can be discerned from the disparity as seen from the two eyes, it is also believed that the vergence of the eyes on an object of interest gives a measure of relative distance.

We are, in most cases, completely oblivious of the two separate images our eyes sense. This is because the brain fuses the separate images into a single image in a process called stereoscopic fusion. When stereoscopic fusion fails, one of two things can happen: two ghost images are seen, or the image from the left and right eye alternately dominates, and these scenarios are known as diplopia and eye dominance respectively. The limits of angular disparity that can be stereoscopically fused is not well defined because of the variability between humans, but a useful rule of thumb is that 2 minutes of arc is the maximum disparity

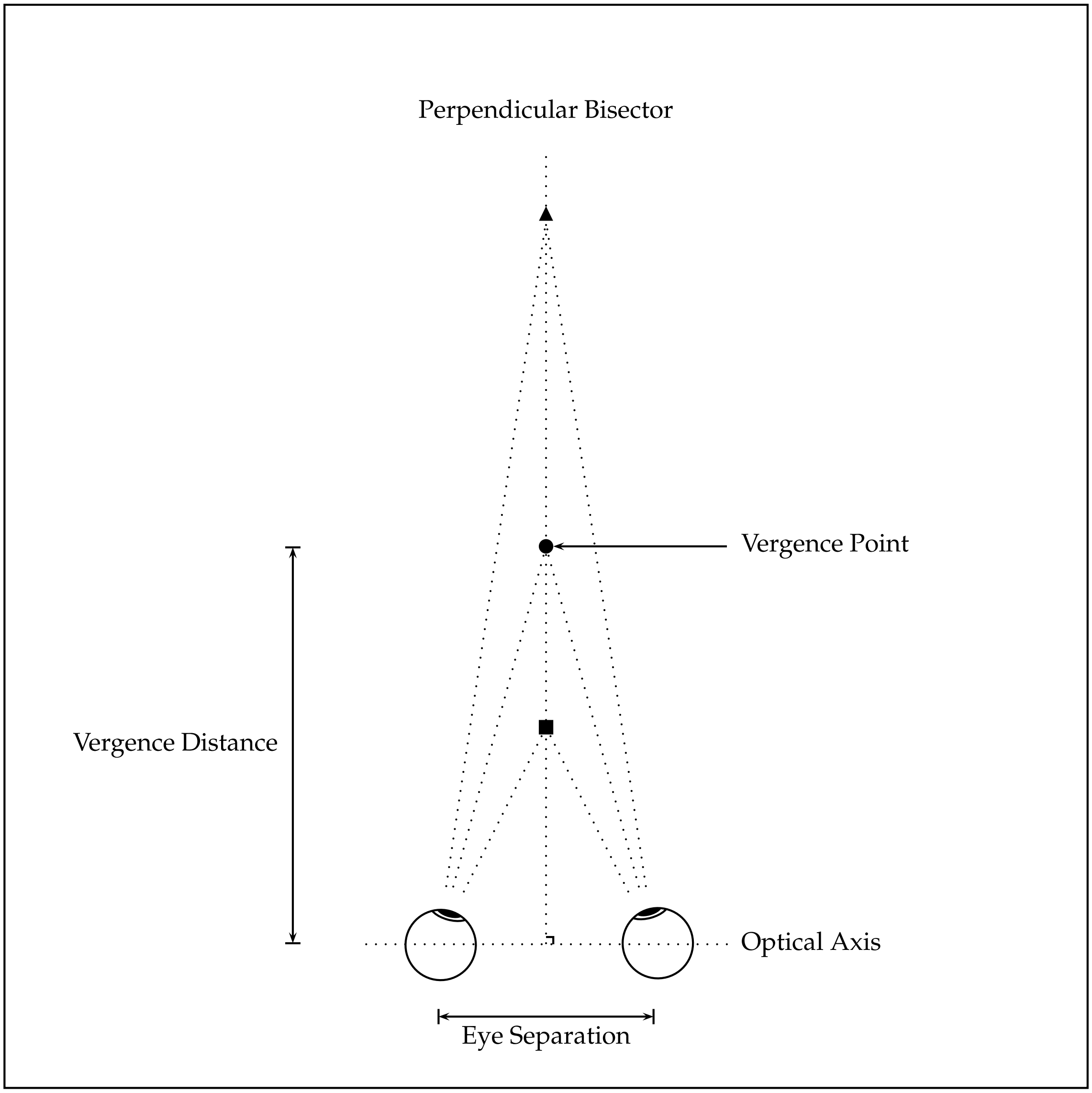


Figure 1.1: The geometry of viewing

that can be fused without adjustment of the vergence point.

It should be noted that several depth cues can be obtained using monocular vision and it is for this reason that two dimensional displays using projector or monitor technology can be so effective. However, two dimensional images or displays are prone to illusions and depth ambiguities, the best example of these being the artwork of Escher. For this reason, it has been found that stereoscopic displays are capable of giving an observer a better impression of the environment.

1.2.6 Stereo Video

Stereo video is created using two synchronised video cameras generally located as in Figure 1.2. These cameras are used to emulate the human eyes, and have similar properties such as separation, vergence and field of view. However, analogies with the human visual system are a little misleading as there are several distinct differences.

The control of the vergence in the human visual system is very fast and, except for extremely close objects, we are not aware of the exertion and relaxation of the muscles which control this movement. With a stereo camera setup, automatic adjustment of the vergence is generally very expensive as it requires specialised hardware. For this reason, the stereo cameras are more commonly fixed in place at a specified separation and vergence. The effect of fixing the vergence is to limit the range of depths that the stereo effect can be perceived.

A critical facet in the production of stereo video is the synchronisation of the video cameras to each other. This is particularly important when objects are moving in the scene or the video cameras are in motion. If the cameras are out of synchronisation then objects in motion will have distorted vertical and horizontal disparities, and this will degrade the stereo effect.

Similarly, the exposure, zoom, and focus controls for cameras, all of which have analogies with the human visual system, are deficient when compared to the human visual system. This is also very important for stereo video production because small differences in these settings between video cameras are quite detrimental to the perception of stereo.

1.2.7 Overview of the Wedge system

The development of the Wedge system[11] began in December 1996 and was an effort to construct a visualisation system that was similar in principle to the CAVE[19] type systems, but using cheaper components and not requiring special edge blending to eliminate corner effects. One result of the Wedge development was the Wedgeorama theatre which has two screens perpendicular to each other that are 2.2m high and 4.0m wide, each of which is rear-projected by CRT projectors. The projectors are driven by an Intergraph computer with 256 MB of RAM and dual Pentium CPU's running at 350 MHz. The computer has two frame synchronized video adaptors operating at a resolution of 640 by 480 pixels at a refresh rate of 120Hz. The stereoscopic display uses the frame sequential technique and liquid-crystal shutter glasses, which an observer wears. All of the components of the Wedge system are available commercially which makes it far cheaper than more common immersive virtual

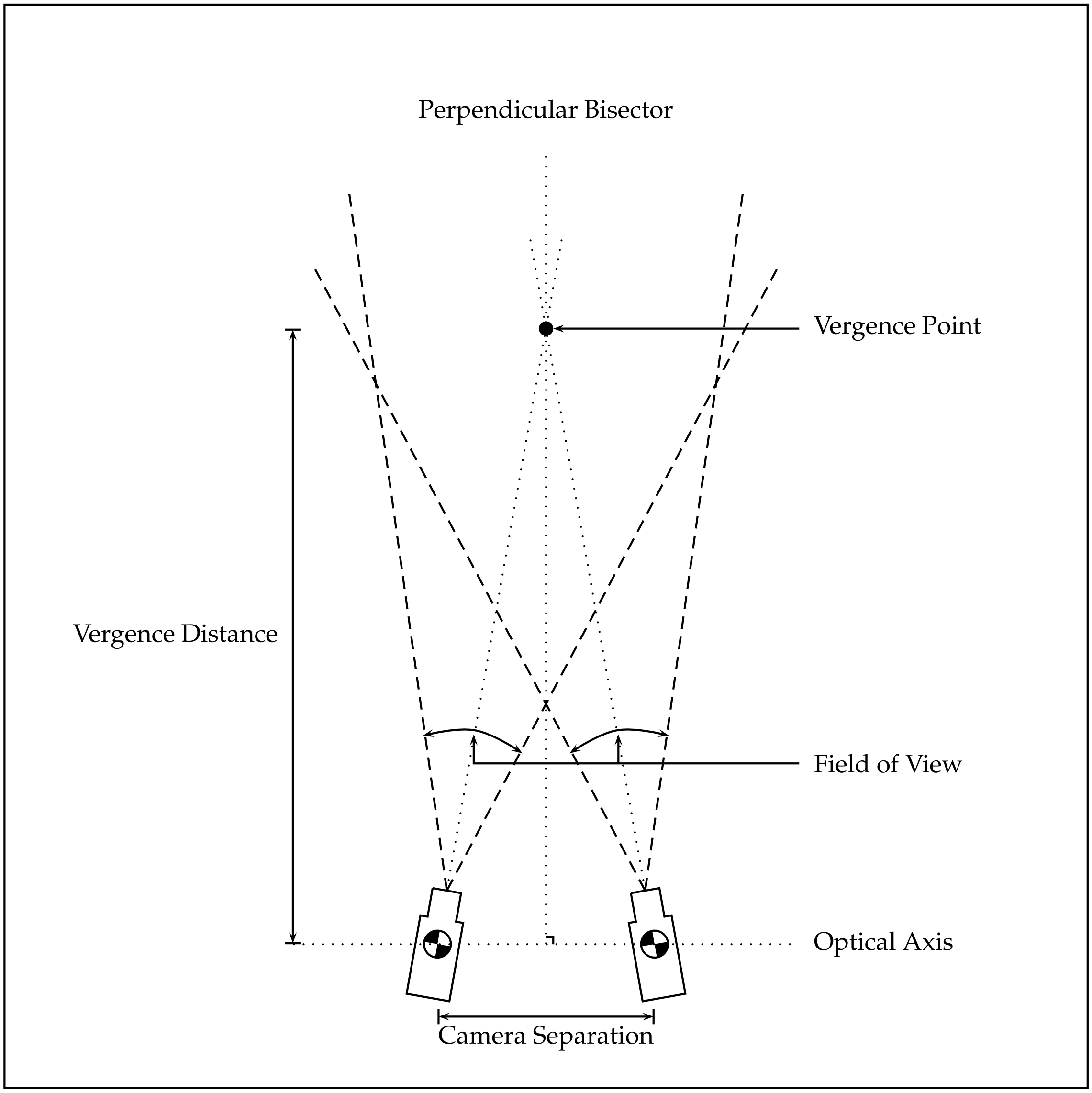


Figure 1.2: A general stereo camera configuration

reality systems.

Wedge systems have been installed at the Powerhouse Museum in Sydney and the CSIRO Discovery Centre in Canberra. These systems are used to provide informative demonstrations to children and the general public to give them a more entertaining introduction to various scientific disciplines. Another system has also been installed at the Australian Defence Force Academy in Canberra for the purpose of further research and development.

The development of the Robot Command Station is seen as the next step in Wedge-style virtual reality systems. Although the Robot Command Station will use different technology to the present Wedge Theatres, the principles will be much the same. The primary difference being the size of the Robot Command Station, which is intended to be a desktop sized system rather than one capable of full immersion.

1.3 Motivation

This project has several very distinct areas of emphasis, each of which has its own set of goals that determine the direction of this research. There have been several outside influences which have shaped the objectives of this project, but they also share a great deal of commonality. It is hoped that this project will give impetus to following projects, and provide a solid grounding in the fundamentals of digital stereo video.

1.3.1 Robot Command Station Project

The prime motivation for this project is the Robot Command Station Project. This is part of the Mobile Robot Project being undertaken by the Research School of Information Sciences and Engineering (RSISE) at the Australian National University. The aim of this project is to develop the infrastructure for the deployment of remote mobile robots, with the ultimate goal of being able to control a highway-speed mobile robot remotely.

The type of mobile robot was inspired to some degree from projects such as Nomad[77], and to a lesser extent, the Pathfinder mission to Mars[1]. The commonality between these projects is the control of a mobile robot from a remote station. This can only be achieved through the use of near real-time transmission of information about the environment in which the robot is currently located. The form of information that most succinctly conveys the current environment is video, with stereo video providing substantially more information.

The primary reason for the construction of a mobile robot is that these robots are to some degree expendable, and therefore can be put into hazardous environments. Perhaps the best example of this is the recent Pathfinder mission to Mars, but some Earth-bound examples of these types of systems are extreme nuclear environment monitoring [50], under-water investigation [81, 78], bomb disposal [23], military surveillance [29] and other dangerous environments.

The original concept of the ANU mobile robot system is shown in Figure 1.3. The mobile robot is placed within two kilometres of a relay station, and communicates with the relay

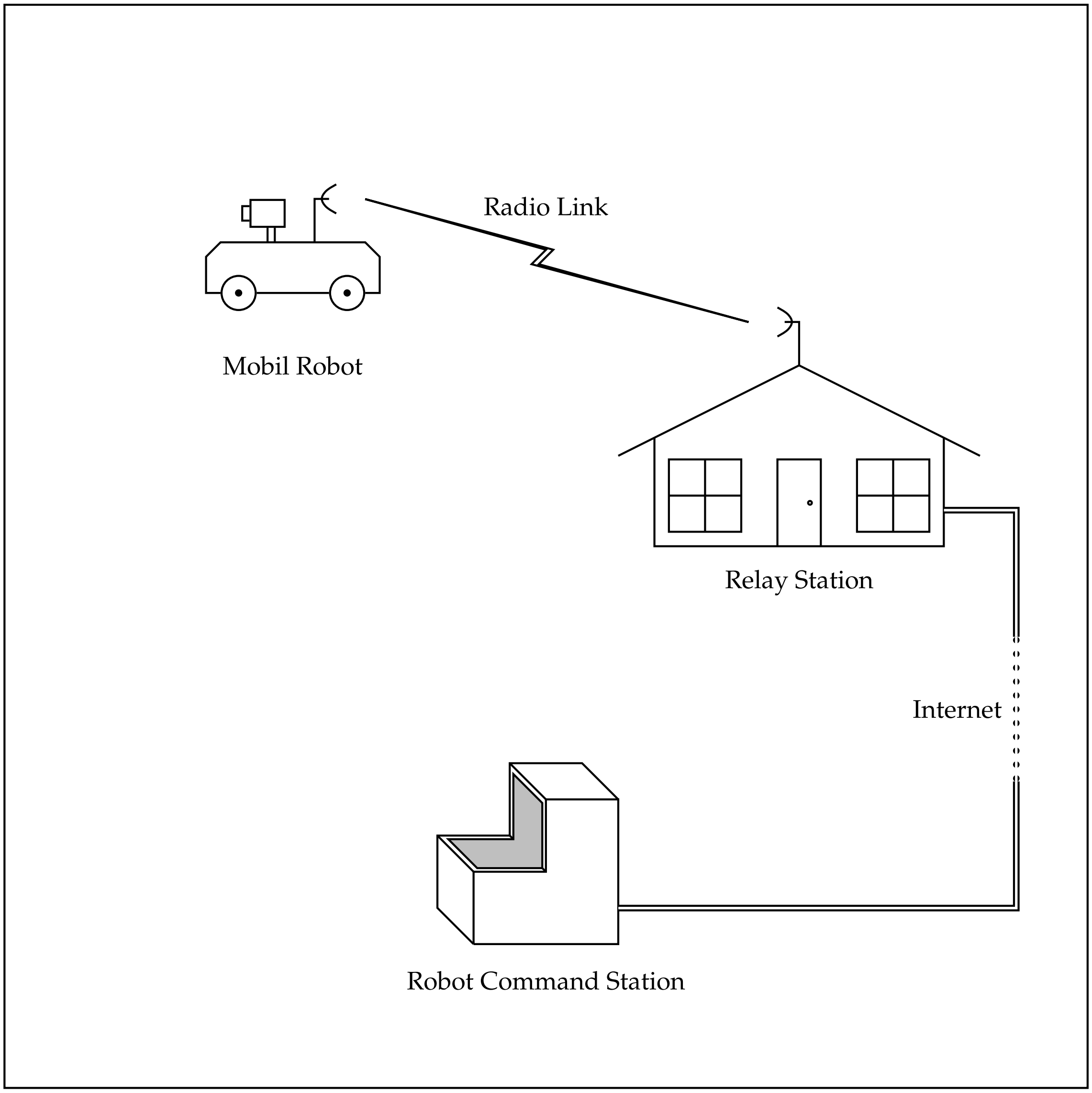


Figure 1.3: Concept of the Mobile Robot System

station via a Radio-Ethernet bridge. The relay station is connected via a high speed link back to the Australian National University, whence the mobile robot is controlled. All of the communications are to be done using the Internet Protocol to handle the sending and receiving of information. The total bandwidth between the mobile robot and the command station is limited by the radio link, which has previously been measured at approximately 6 Mb/s.

The Robot Command Station is to be the command centre for controlling the mobile robot. The initial concept involved an L-shaped table as shown in Figure 1.4. The movement of the mobile robot was to be controlled by the use of a force-feedback haptic device, however this was deemed inappropriate at the time and was replaced by a joystick.

The Robot Command Station was to include head tracking and a 3-D wand pointing device, and these requirements will be the topic of future research. The Robot Command Station was also going to use the same technology as the Wedge, that is, liquid crystal shutter glasses with cathode ray tube projectors. Also, there was a firm desire to investigate the use of polarising glasses due to the reduced cost of the glasses and a belief that alignment problems with the cathode ray tube projectors could be alleviated with different projector technology. A further benefit of using projectors other than CRT projectors is that DLP and LCD projectors are smaller and more portable.

1.3.2 Virtual Reality Collaboration

A secondary project, for which a grant application was made and subsequently rejected, was the Collaborative Virtual Reality system. With Wedge systems installed at the Australian National University in Canberra and at the Powerhouse Museum in Sydney, the project was to essentially link these two Wedge's via the Internet. As part of the grant application, a Wedge system was to also be constructed at the Wollongong University, to which the other Wedge's would also be connected.

The intent in connecting the Wedge systems via the Internet was to perform research and development into stereoscopic teleconferencing and the transmission of stereoscopic datasets and imagery via the Internet. It was hoped that such a system would enable researchers to collaborate more freely, with the ability to simultaneous analyse complex three dimensional data using an advanced stereo display system. Coupled with the ability to teleconference with remote Universities in stereo, it was believed that such equipment would provide much needed infrastructure.

However, with the rejection of the grant for this particular project, it was unable to proceed. There is some commonality between the Robot Command Station project, in that there is a need to develop a high speed, near real-time stereo video compression system and the necessary software for the transmission of the compressed stereo video. Hence, the collaborative virtual reality system is something for which the Robot Command Station project will provide some preliminary data.

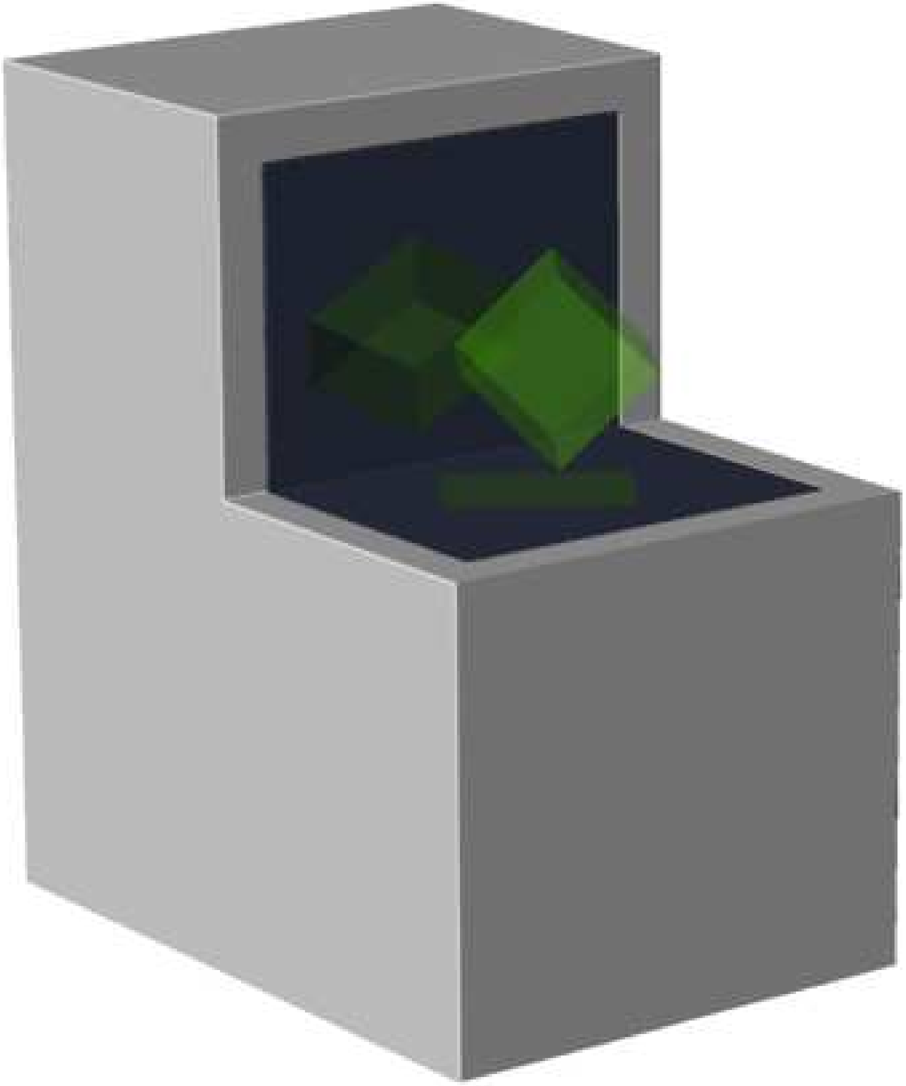


Figure 1.4: The Robot Command Station Concept

1.3.3 Investigation into alternate Virtual Reality systems

The development of the Wedge system has provided many lessons in the construction of virtual reality theatres. There were two problems that were found with the Wedge systems that had been constructed to date. Firstly, liquid crystal shutter glasses are quite expensive and for a system that has the potential for a large audience, the cost of the glasses alone amounts to a fair portion of the total cost. Secondly, the projectors used in the Wedge systems were CRT projectors, and alignment of these projectors is a time consuming process.

The main problem with the projectors is that each time a parameter of the Wedge changes, such as screen resolution or refresh rate, the alignment of the projectors changes and has to be readjusted. This becomes a real problem for when the Wedge must be relocated, as the alignment of the projectors constitutes a large portion of the total setup time. So due to the desire to make the Wedge system more portable, an investigation into the use of different projector technology was required.

Hence, a part of the Robot Command Station project was the investigation into a stereoscopic display using polarising glasses and DLP projectors. Polarising glasses were to be used as opposed to the more traditional liquid crystal shutter glasses to reduce cost, and DLP[3] projectors were to be used in preference to CRT projectors to remove the alignment problems faced with CRT projectors.

1.4 Objectives

1.4.1 Production of Stereo Video

The focus of this project is on digital stereo video, and, as such, there is an implicit requirement to determine how to best obtain realistic stereo video. This is primarily a function of the cameras used in the production of the stereo video, and the orientation of them with respect to each other. There have been several studies into the problems associated with the production of stereo images and video [32, 33, 80] and, in general, these are theoretical and relate to perfectly behaving and adjustable cameras. In our case, we have been provided with specific video cameras, and our objective is to determine the set of parameters which give the best stereo video effect for them.

The measure of the best stereo effect is not something that can be easily quantified, and has the potential to be very subjective and vary from person to person. A set of guidelines for this measure is that the stereo video should give a good and realistic sense of depth without too great an exaggeration, and the viewing of the stereo video should not cause discomfort to the viewer.

[3] LCD projectors were not considered as they interfere with polarisation and DLP projectors are typically brighter

1.4.2 Design and construction of the RCS

Given the desire to trial polarising technology for the stereoscopic display, the Robot Command Station will use polarising glasses, rather than the more familiar liquid crystal shutter glasses. In keeping with the Wedge systems, the Robot Command Station will use commercially available equipment to keep the overall cost of the system relatively low.

It will be a table top system that can be comfortably viewed by a small group, although it is intended to be used by a single operator. There was a firm desire for the Robot Command Station to become a general purpose stereoscopic virtual reality system, rather than a purpose built stereoscopic system. So the Robot Command Station must be able to render polygonal models in three dimensions across both screens in stereo, in the same way as the Wedge. As well as this, the Robot Command Station must be able to show stereo video on at least one screen, in near real-time.

1.4.3 Stereo video compression

The next objective was the time effective compression of stereo video, and, in keeping with the philosophy of using only commonly available products, the compression and decompression were not to be aided by specialised hardware. The first stage was to investigate the various commercial software packages available for the compression of video. Such packages are typically used for broadcasting video over the Internet. An investigation into whether these products could be used for stereoscopic video was to be made as an initial goal.

The main purpose of the compression of the stereo video is to reduce the bandwidth of the transmitted data which allows a greater number of video frames to be transmitted per second, resulting in a better conveyance of information about the environment. The main restrictions on the compression method used were on the latency, that is, the time from when a particular frame is captured to when it is finally displayed on the Robot Command Station. As the controller of a mobile robot needs the video to appear essentially in real-time, the compression and decompression of the stereo video must not be so computationally expensive that it greatly affects the latency of the overall transmission of the stereo video.

As stated previously, the maximum bandwidth of the link between the mobile robot and the RCS is approximately 6 Mb/s. The bottleneck in this system is the Radio-Ethernet bridge, which is susceptible to fluctuations in this limit due to weather conditions, line of sight and distance between the relay station and the mobile robot. Hence the compression of the stereo video should aim to produce a data stream that requires significantly less bandwidth. The ability to produce stereo video of reasonable quality at a bit rate of 1 Mb/s is a suitable objective for the compression of video.

As the stereo video is intended to be used by an operator to control a mobile robot, the number of frames per second must be sufficiently high in order to effectively convey the motion of the robot. The maximum available frame rate from a commercial video camera is generally between 25 and 30 frames per second, however, the minimum frame rate necessary to give a reasonably sense of motion is between 10 and 15 frames per second. The frame rate achieved is dependent on the computational cost of the compression method. This can be seen from the fact that the time taken to compress a single frame of video is the inverse of the maximum theoretical frame rate. Similarly, for bandwidth limited systems such as this one, the frame rate is constrained by the average number of bits per frame. The maximum bandwidth divided by the average number of bits per frame is also the maximum theoretical frame rate. Hence, the frame rate achieved is dependent on several factors, and the choice of compression method plays a very important role in the determination of the ultimate frame rate.

A reasonable quality of stereo video must be available to an observer at the Robot Command Station. The definition of reasonable quality is not one that is easily quantifiable as it is a very subjective measurement. Although this project will not be able to control a mobile robot at its completion, this would have been a good measure of the quality of the video. If an operator could control a mobile robot in an obstacle course, such as in Gärtner & Schneider [29], then this would provide evidence that the quality was sufficient for the purpose. While a good measure, this would not help for the future goal of providing the ability to teleconference in stereo as this has a different set of requirements. Given the desire to be adaptable and the lack of a quantifiable target for stereo video quality, a suitable objective is that the compression scheme should be able to adjust the quality of the final stereo video as seen by an observer to suit the particular application. The compression scheme should also be capable of realistically reproducing the colour of the scene to some degree.

1.4.4 Transmission over Internet networks

Lastly, the compressed stereo video stream must be transmitted from the mobile robot to the Robot Command Station. This is to be accomplished using existing Internet Protocols as the low level protocol for the transmission of information. Many of the existing compression methods are, in part, designed to be transmitted over the Internet, and there are already several commercial packages that provide compression and Internet transmission of video.

The choice of a compression scheme plays a significant part in the transmission as it is directly responsible for factors such as the bit rate, the format of the compressed stream and the tolerance to corruption and partial loss of data. Hence the method of transmission over Internet networks is closely related to the method of stereo video compression used, and they must be considered together.

There are several general-purpose protocols that have been designed on top of the basic Internet Protocol, for example, TCP and UDP, both of which provide different features and services. An initial objective is to study these various protocols and commercial products to determine how these work and if they can be applied to this project, with the ultimate goal of the transmission of the compressed stereo video stream over the Internet.

Chapter 2

Vision

A introduction to the operation of the human eye is given, as well as the principles behind binocular vision.

2.1 Introduction

The following material is a compilation of information gained from a variety of references [55, 62, 16, 53, 84, 12, 27] concerning the human visual system. The display of stereo imagery relies on the ability of the human brain to successfully fuse the two images in order to obtain depth cues from the stereo disparity.

2.2 The Human Eye

A horizontal cross-section of the human eye is shown in Figure 2.1. The cornea is a thin transparent membrane through which light first enters the front of the eye. It is approximately a spherical section and takes up about one sixth of the eye's outer surface area. The remainder of the outer surface of the eye is the *schlera*, or the white of the eye. The eye overall is approximately 25mm in diameter with the volume primarily being taken up with the *aqueous humour* in the front of the eye and the *vitreous humour* at the rear, both of which have water as their main constituents.

The *iris* of the eye controls the quantity of light that passes to the sensory elements. The central aperture of the *iris*, through which light passes unhindered, is called the *pupil*. The diameter of the *pupil* is varied between 2mm and 8mm by the involuntary muscles.

The ability of the eye to focus on objects is due curvature of the *cornea*, *aqueous humour*, *lens*, and *vitreous humour* and their respective refractive indices. The *cornea* of the eye provides most of the power of the eye due to its curvature and refractive index. The flexible lens provides the ability to vary focus by the action of ciliary muscles which alter the curvature of the fore and aft surfaces of the lens. With good eyesight, objects at 15cm to infinity can be focused on by changing the shape of the lens.

The eye can sense light in the wavelengths from 380nm to 760nm. Light in the ultra-violet end of the spectrum tends to be absorbed by the *cornea* and the lens. Similarly, light at the

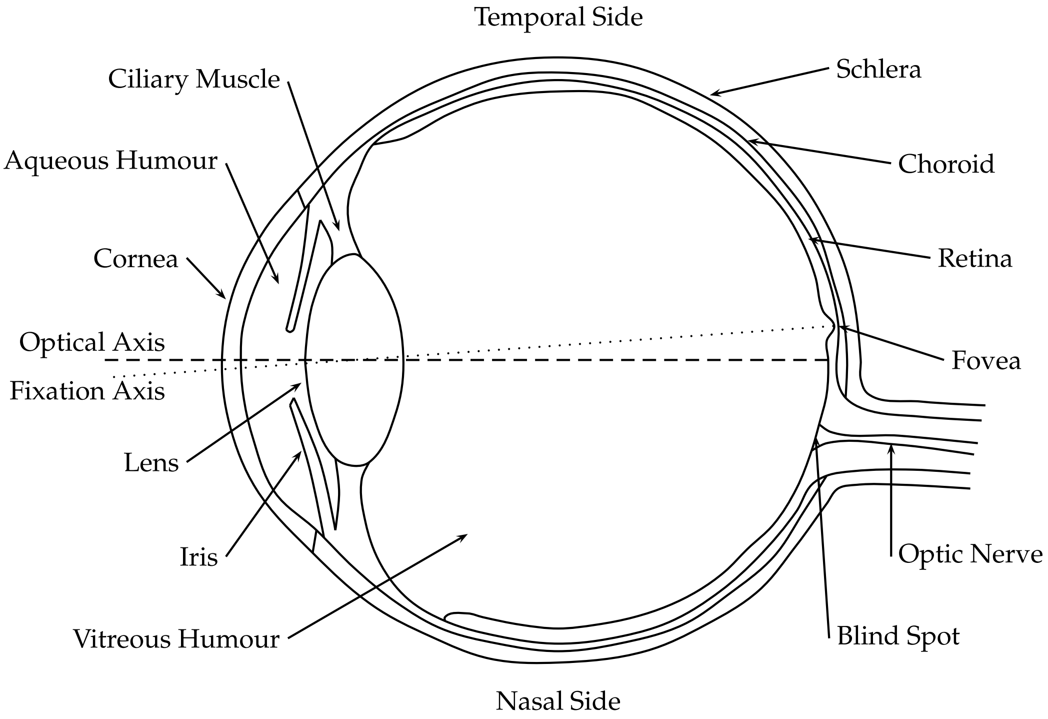


Figure 2.1: A horizontal section view of the human eye through the meridian [16]

infra-red end of the spectrum tends to be absorbed by the *aqueous humour* and the *vitreous humour*.

At the rear of the eye is the *retina*, which covers the rear hemisphere of the eye. This is the component of the eye which converts the incident light into electrical impulses that can be processed by the brain. Of interest in the rear of the eye is a small indentation in the retina called the Fovea, which is the centre of vision. When an object is fixed upon by a person, light strikes the retina here. Approximately 16 degrees off the fixation axis towards the nasal side of the eye is the optic nerve which communicates impulses from the *retina* to the brain. Due to the high density of nerve fibres here, there are no sensory cells and it is therefore blind (this area of the *retina* is commonly called the *blind spot*).

2.3 The Retina

A cross-section of the *retina* reveals a series of layers which, for simplicity can be separated into two distinct areas, the receptors and the transducers. The receptor elements are rod and cone cells, which convert incident light into nerve impulses, whereas the transducer cells, such as the bipolar and ganglion cells, transcode the nerve impulses before they are conveyed to the brain via the optic nerve. The boundary of the retina is the pigment epithelium which lies beyond the rod and cone cells. A high percentage of the light that reaches the retina is sensed by the receptors even though it must pass through several layers of transducer cells. In each eye, there are approximately 120 million rods, 7 million cones and only about 1 million nerve fibres linking the eye to the brain. At the periphery of vision, several rods and cones share nerve fibres whereas in the *fovea*, there is generally a one to one connection between nerve fibres and receptors. This, presumably, increases the visual acuity in the *fovea*.

In the centre of the *fovea* is a small indentation, where the transducing cells that normally interfere with the light striking the sensory cells, are less inhibiting. This area, the *fovea centralis*, occupies a circular region of approximately 1mm in diameter which is responsible for sensing the central 2 degrees of the field of vision. Hence, the ratio of rods to cones increases, as the angular displacement from the centre of the fovea increases and the properties of vision in the central 2 degrees of vision is governed by the properties of the cones.

Of the rods and cones, a great deal more is known about the rods because at the edges of vision, only rods are involved, and this means that the rod cells can be isolated in experiments on living or deceased eyes. The rods contain a pigment call rhodopsin, which seems to be a common chemical found in the eyes of several animals such as fish, cats, monkeys and indeed humans. As such, much is known about this pigment and the properties it has when excited by light. When deprived of light, rhodopsin is a reddish-purple colour. The action of light causes it to bleach to yellow, and, if the light is sufficiently intense or the exposure is long, the rhodopsin becomes colourless. Generally, only 60 percent of the light absorbed actually bleaches the rhodopsin and thus the remainder does not effect vision. Rhodopsin is most sensitive to wavelengths of light around 500nm (blue-green), and tails off at wavelengths below 400nm (violet) and above 600nm (orange). The regeneration of rhodopsin from fully bleached to the original reddish-purple state is an exponential process with 50 percent regeneration taking about 7 minutes. The reddish-purple state is commonly called the dark adapted state.

Given higher sensitivity of the rods in the dark adapted state, and their higher concentrations at the periphery of vision, it is evident that they are primarily for night vision and provide a low fidelity peripheral vision to detect motion. The human visual system has three separate stages depending upon the intensity of light in the environment, these stages are known as scotopic, mesopic and photopic. In scotopic vision, the light levels are too low for the cones to operate, hence vision is exclusively governed by the action of the rods. Mesopic vision is the transitory state where both cones and rods can operate effectively. It is, however, a narrow region and not of much consequence. Photopic vision occurs at higher light intensities and is used for normal daytime vision. At these light levels, the effectiveness of the rods is greatly reduced since they are bleached by the light, hence vision is primarily governed by the cones, although the rods still play a part, especially in the periphery of vision. A general approximation is that cones will only respond to light levels above 10^{-3}candela m^{-2}, similarly, rods can respond to light levels down to 10^{-6}candela m^{-2}.

Towards the centre of vision rods and cones are interspersed, and cannot be isolated to discover their individual properties. In the past 50 years there have been several investigations into the cone receptors which have shown that there are three types of cones, distinguished from each other by the wavelength of light they absorb strongest. The so called blue, green and red cones have their peak responses to wavelengths of approximately 440nm, 540nm and 590nm respectively.

2.4 Colour vision

Many theories have been proposed over the centuries as to how human colour vision operates, and it has only been through recent experiments that confidence in these theories has been gained. The oldest and still the most accepted theory of colour vision was first proposed by Thomas Young in 1802. His reasoning was that since the number of nerve pathways from the eye to the brain was limited, and that the triplicity of colour had no foundation in the theory of electromagnetic radiation, the eye therefore has three separate signalling mechanisms which respond to red, green and blue. Herman von Helmholtz showed that more than three primary colours were required to match some colours exactly which was contrary to Young's theory and led to its rejection for many years.

Subsequently, Helmholtz showed that if the three primaries were not pure, then this was sufficient to match any colour. Hence, while a red receptor will respond more vigourously to light in the red wavelengths, it will still respond in less vigourously to green and blue wavelengths. This notion of the eye having three receptors with overlapping response curves became known as the Young-Helmholtz theory and has been widely accepted since its inception. It has only been this century that direct evidence for the existence of three cone receptors has been found.

Ewald Hering, at the begining of the 20th century, discovered that some colours tended to cancel each other out, for example, red cancels green and blue cancels yellow. Hering subsequently postulated that there were indeed three different receptors in the eye, however, rather than responding to red, green and blue, they respond to red versus green, yellow versus blue and white versus black. This theory became known as the Hering Opponent Colour theory and was the most accepted theory in opposition to that of the Young-Helmholtz theory.

The ability to perform electrical analysis of the process of seeing has provided great insight into the operation of the eye, and in particular, communication between the eye and brain. These analyses have provided much understanding of how the human colour vision system works and its limitations. A reconstruction of the results obtained by Arakawa [6] is shown in Figure 2.2. These graphs show that colour vision extends from the centre of the field of view to about 50 to 60 degrees either side, which corresponds to approximately 100 degrees of the entire 180 degree field of view. Naturally, as the density of cones decreases proportionally with angular distance from the fovea, the colour acuity also decreases, and this can be seen from the abrupt decrease in response to different wavelengths as soon as the incident light moves away from the fovea.

Whilst there is evidence that there are indeed three types of cones as predicted by the Young-Helmholtz theory, arguably the most interesting result from electrical analysis is that early analysis of the retina of fish showed that certain cells gave potentials that became more negative for increasingly blue light, and more positive for increasingly yellow light. Other cells were also found to respond in a similar manner for red versus green light. This experiment was repeated on the macaque monkey which has a colour vision system that more closely resembles that of humans. Similarly, cells were found which produced spike potentials that increase the rate of firing depending on whether red versus green, or yellow versus blue light is incident upon them. Such cells are called opponent cells, and conversely there exist non-oponent cells that respond proportionally with luminosity.

Current understanding of the human colour system can be summarised as follows: rods are receptors which detect luminosity levels, primarily at low light levels and at the periphery of vision. There are three different types of cone receptors, responding to blue, green and red light (the red cone actually responds more strongly to yellow), as predicted by the Young-Helmholtz theory. The outputs of the cone and rod cells are coded into spike potentials with the frequency being the measure of intensity. There are four distinct type of signals sent to the brain, the luminosity from the rods, the luminosity from the sum of the cones, a red versus green signal derived from the red and green cones, and a yellow versus blue signal derived from a compounding of red and green cones and the blue cone. This signalling method is close in concept to the Hering Opponent Colour theory.

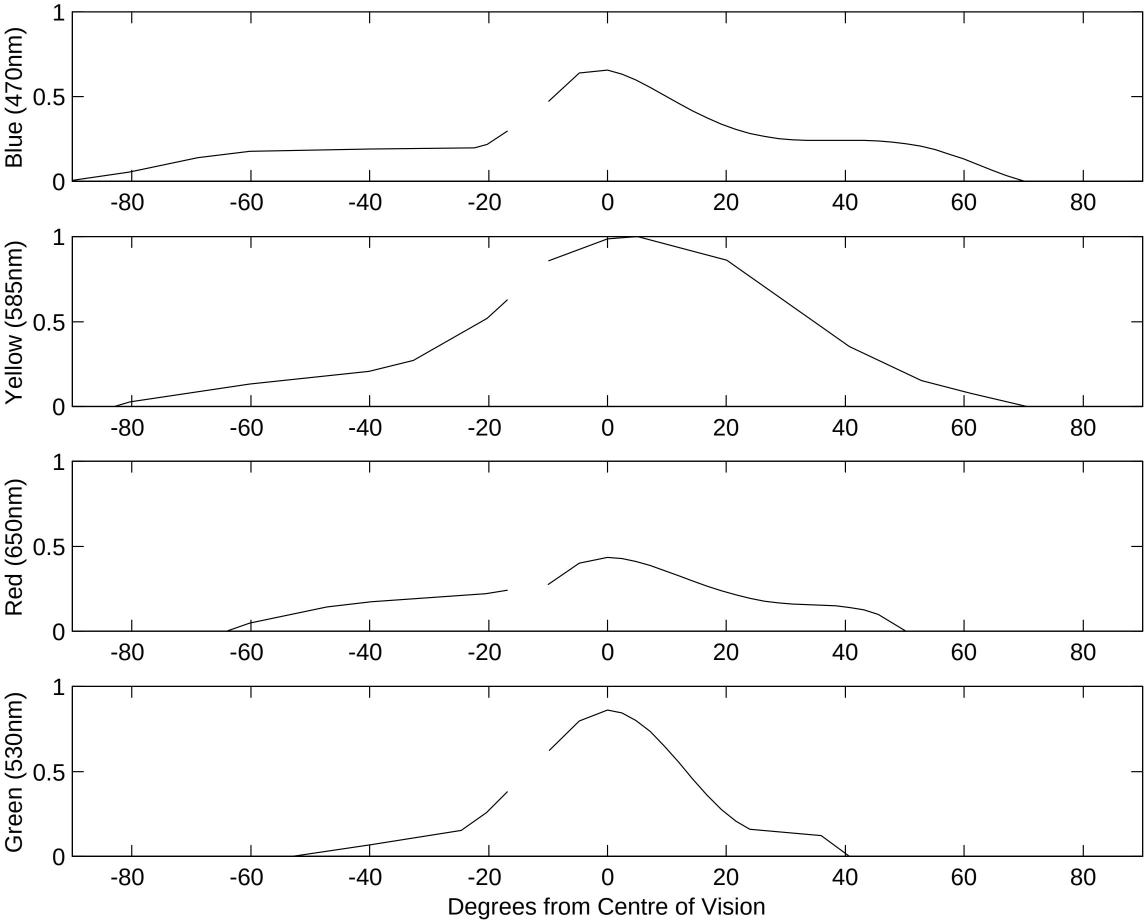


Figure 2.2: Limits of the field of colour vision at the horizontal meridian

2.5 Temporal Aspects

The coding of luminosity as potential spikes, where the frequency of the spikes is a measure of the intensity of the light (higher frequencies correspond to higher intensities), has several interesting side effects. Firstly, it takes approximately 10ms for a light stimulus to be perceived, and similarly the stimulus is still perceived a short period after it has gone, this is known as the persistence of vision. An important consequence of this persistence is that, at high enough frequencies, a flickering light stimulus will appear to have constant brightness. The point at which this occurs is called the flicker fusion frequency, which has been found to vary with the logarithm of the luminance to which the eye has adapted. An emperical formula which gives an approximation for the frequency is the Ferry-Porter law:

$$f_c = a \log_{10} L + b \tag{2.1}$$

where L is the luminance to which the eye has adapted, and a and b are empirically derived constants. This frequency varies from about 10 Hz at low luminance levels to a maximum of approximately 55 Hz. The application of this fact can be seen in the engineering of several devices. For example, fluorescent lights flicker at 100Hz when operated from a 50Hz supply (incandescent bulbs emit light via black body radiation and do not flicker). Movies are filmed at 24 frames per second, and at a cinema, each frame is shown twice to increase the flicker frequency to 48Hz. Television has 25 or 30 frames per second, depending on the particular broadcast standard, where each frame is split into two fields which can be interlaced back into a frame. Fields are shown sequentially (frame interlaced) so that, like cinema, the flicker frequency is doubled. Computer monitors generally refresh the screen at much higher frequencies, 60 to 100 Hz, and this is primarily because they are typically used in bright office environments, when the ability of humans to perceive flickering is highest.

2.6 Eye Movements

The orientation of the eye is controlled by six muscles that allow the eye to tilt up and down, pan left and right, and rotate about the fixation axis in either direction. There are several involuntary movements which prevent the eye from being able to fixate on a point for a long time. These movements are called *tremors*, *saccades*, and *drift*. The tremors have a frequency of approximately 50 Hz and an amplitude of less than 1 minute of arc. The saccades occur approximately once a second and the fixation point moves up to 5 minutes of arc at high speed. Lastly, the fixation point tends to drift in between saccades, and this is probably due to the tremors. It has been shown through experiment that if an image is fixed on the retina, after a few seconds pattern and colour grey out and only a dim uniform field is seen. From these experiments, it has been theorized that the involuntary movements are responsible for limiting fatigue or enhancing contrast in the eye.

The movement of the eye plays an important role in binocular vision and many believe that the brain may use the degree to which the eyes are verged as another depth cue. The rotation of the eyes has been shown to counter rotations of the head, where a rotation of the head of 45 degrees results in a rotation of the eye of about 37 degrees. However, studies have also shown with rotated stereograms, that the eyes are also rotated to better accomodate the disparate images. Indeed, when viewing a stereo scene, the fixation point of each eye is changing constantly, including the vergence of the eyes, thus the scene is continuously scanned.

2.7 Binocular vision

Not a great deal is known about how depth is perceived from the two uniocular images that are sensed by the eyes. The optic nerves from each of the eyes remain separated until they fuse in the *striate* area of the brain. It is assumed that this is the centre for binocular disparity processing. The discussion of human binocular vision will be limited to definitions, basic functions, and limitations.

Before discussing the fusion of two uniocular images, it is worthwhile to consider other depth cues which feature in human vision that can be gained from monocular vision and learned experience, as given by Helmholtz:

1. Judgement of distance from the size of the retinal image of objects whose approximate size is known, such as men and cattle.
2. The apparent converging of parallel lines in perspective as they recede.
3. The overlapping of one contour by another when a more distant object is partially concealed by another.
4. Clues given by shadows.
5. Aerial perspective or the softening of outline and colour changes of distant objects due to the atmosphere.

A common thread of thought is that binocular vision, rather than giving a quantitative sense of depth, enables an observer to judge the relative depths of objects in a scene, and thus construct a 3-Dimensional mental model of the current environment.

The two disparate images sensed by the left and right eyes are stereoscopically fused so that in general, we only perceive one image of the world, albeit, that image contains additional *stereopsis* [1] information. There are, of course, limits to this fusion process and these are governed primarily by past experience providing conflicting depth cues and a limit on the amount of angular disparity that can be fused. When fusion fails, one of two things occur. Two ghost images of the object that cannot be fused are seen, which is known as *diplopia*. Alternatively, the image received from one of the eyes is suppressed or blended

[1] Stereopsis is a descriptor for the depth cue in a scene caused by stereoscopic disparity.

in preference for the other eye. This is called *binocular rivalry*, or sometimes *eye dominance*. Sometimes this phenomenom oscillates between left and right eyes with a period of 1—5 seconds.

The angular areas within which stereoscopic fusion is possible are called Panum's fusional areas. The limiting angular disparities are commonly known as the diplopia threshold. The accuracy of the measurement of Panum's area is not high due to wide discrepancies between experimental results. Figures quoted by Davson [21] give the value of 3 to 4 minutes of arc for vertical disparities and 5 to 26 minutes of arc for horizontal disparities. A more recent study, [25], showed that the diplopia threshold varies significantly between people and that training and various other factors greatly influence the threshold. A study, [83], with perhaps more relevance to this project showed that in stereoscopic colour displays, the limits of fusion were:

- 27.1 minutes of arc for crossed disparity, and
- 24.3 minutes of arc for un-crossed disparity.

These figures provide a useful reference later for measurements taken using stereoscopic display equipment. It should be noted that these measurements were taken for a stimulus whose duration was 200ms, which is believed to circumvent the verging of the eyes to accomodate greater disparities. The results obtained from the same study using a stimulus with a duration of 2 s were:

- 4.93 degrees of arc for crossed disparity, and
- 1.57 degrees of arc for un-crossed disparity.

The figures for the 200 ms stimulus can be thought of as the maximum disparity from the current vergence point that can be stereoscopically fused. Disparities of angles greater than these must be scanned by an observer by adjusting their vergence. On the opposite side of the scale, Regan et al [62] quotes the minimum observable disparity as 2 seconds of arc for the best observers.

The human binocular vision system can also tolerate some degree of vertical disparity. Presumably this in cooperation with the rotation of the eyes enables depth perception when the head is tilted. In Duwaer et al [25], vertical disparities of between 1.2 and 19 minutes of arc could be stereoscopically fused, depending on the observer. This has some important ramifications for both the display system and the camera setup as will be explained later.

The rotation of the eyes about the optical axis can also play a part in the stereoscopic fusion of two uniocular images. It is possible to stereoscopically fuse a stereo image pair rotated slightly with respect to each other, albeit with some effort. While this is not conclusive proof that the eyes are rotated, it implies that some rotation operation is applied to the images, either of the eye or in the brain.

The human binocular vision system is also able to stereoscopically fuse two uniocular images with differing intensities. This ability is most likely to be necessary for observing objects with high reflectance whose intensity can vary greatly with a small change in the viewing angle.

The last factor of importance, especially for stereoscopic display systems, is interocular cross-talk. In fact, this phenomenon is strictly restricted to stereoscopic display systems, which are sub-optimal. The projection of two uniocular images into separate eyes generally involves some degree of cross-talk, where some fraction of the image for the left eye is perceived in the right eye and visa-versa. The common result of this is the perception of "ghosting", particularly about high contrast edges. A study [83] using liquid crystal shutter glass (described in a later section) found that extinction ratios [2] between 11:1 and 17:1 did not greatly effect fusional thresholds, nor estimation of depth. However, it was also found that increased levels of interocular cross-talk decreased the subjective ratings on quality and comfort in the fusion of the stereo images. This seems to indicate that the observer is required to exert greater effort in accomodating stereo imagery as levels of interoccular crosstalk increase.

As can be seen from these studies, the human binocular vision system is flexible and able to adapt to differences in two uniocular images, where some of these differences are not related to stereoscopic disparity. It is important though, to realise that binocular vision cannot give a quantitative measure of depth, but allows the perception of the relative depths of objects in a scene. Past experience and acquired knowledge give depth information that is used in conjunction with the stereopsis of the scene to provide the brain with a spatial representation of the environment in front of an observer. Hence, binocular vision requires other factors to give the information it contains a reference point. It should also be noted that monocular depth cues are given greater credence than binocular disparity, as is evidenced by the occurance of diplopia when monocular depth cues contradict with the binocular disparity of a scene.

While this discussion has concentrated on the ability of binocular vision to provide depth information, other properties of human binocular vision have been neglected. Binocular vision has somewhat higher acuity than monocular vision, and it is believed that this is caused by some process in the brain that performs a form of averaging between the eyes to improve acuity. It is believed that the constant tremor and saccades of the fixation point of the eyes, performs a similar process, since the blind spot of the eye is not perceived in either monocular or binocular vision except in rare circumstances. This averaging process prompted some theories that binocular vision gives humans a greater chance for survival, in that damage to portions, or the entirety of one eye can to some degree be replaced by the use of the fellow eye.

[2] Extinction ratio here is defined as the ratio of luminance for the correct eye to luminance for the fellow eye: Extinction ratio = $\frac{L_{correct}}{L_{fellow}}$: 1.

2.8 Summary

The following features of the human visual system are of interest in the design of stereoscopic display systems:

- up to 2 minutes of arc of disparity can be accomodated without the need for the eyes of a subject to be verged,
- a small degree of vertical disparity and rotation of two uniocular images can be tolerated,
- scale of known objects is an important depth cue, and it is possible for this depth cues (and others) to override the information provided by stereopsis.

Chapter 3

Display

The options available for the construction of the Robot Command Station are discussed. The final concept of the Robot Command Station is described, and the results of the construction of a prototype are given.

3.1 Overview

The display of stereo images is not a new technology, it has been available in many forms for several decades, but recent advances in stereo display technology has boosted its popularity. There are, however, some problems with stereo displays, in that they tend to cause fatigue in viewers who use them for long periods of time.

3.2 Stereoscopic display technologies

Present methods for the display of stereo imagery have many principles in common. The requirement is to project, into each eye of a viewer, two uniocular images whilst limiting the amount of interocular crosstalk. This is generally accomplished by projecting two images onto a projection screen or displaying two images on a monitor, and using filters or optics to direct the correct image to the desired eye.

As shown in Figure 3.1, an observer standing at some distance in front of the projection screen or monitor perceives different images in each eye. If we consider the image of a sphere placed between the observer and the screen, the projection of the sphere as seen by the observer will have a certain amount of crossed disparity (see Figure 3.1a). Similarly, if the sphere is placed behind the screen, then the projection of the object onto the screen has an uncrossed disparity relative to distance away from the screen (see Figure 3.1b). This means that virtual objects in the plane of the screen will have zero disparity. An interesting thing to note is that if the display is set up to provide a true to scale representation of the world, then the maximum uncrossed disparity will be the distance between the observer's eyes. This can be seen from the fact that as an object moves further behind the screen, the projection lines from the object to the left and right eyes of the observer become more parallel, the limit at a

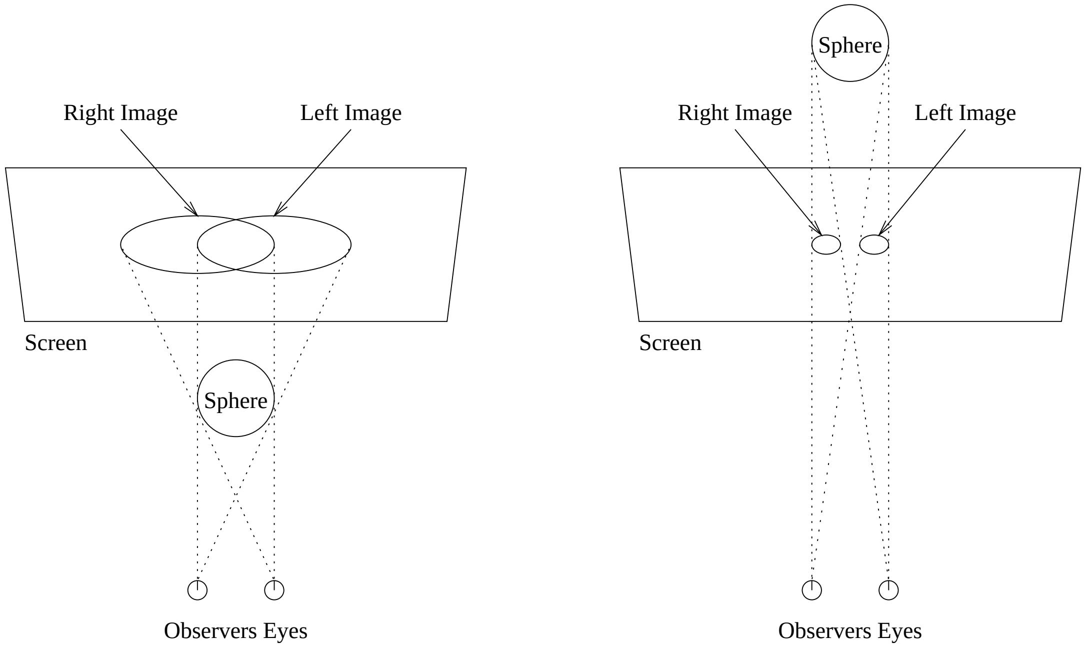


Figure 3.1: The disparity in stereo displays

distance of infinity equating to a disparity on the display screen equivalent to the distance between the eye centres.

It has been shown in several studies [23, 58, 49, 52, 51] that stereoscopic displays improve the performance of operators in teleoperator tasks, particularly those involving manipulation of the environment, obstacle avoidance or traversing of rough terrain. Similarly, studies [54, 56] have also shown that adding stereo to teleconferencing systems improves videoconferencing and allows a greater feeling of connection to the remote participants of a teleconference session. While the benefits of stereo displays have been proven, there remains some problems with them, particularly problems with observers becoming fatigued and developing eye strain.

It is believed that the primary cause of the fatigue in stereo displays is due to the brain having to focus the eyes on the screen whilst verging the eyes on an object in front or behind the screen. This situation is shown in Figure 3.2. It can be inferred that as the disparity for an object increases, when viewed on a stereo display, an observer requires greater effort to fuse the two uniocular images into one. Therefore, objects in front of the screen are more tiring to observe in stereo. The stereo displays described in the following sections all have this particular problem, as it seems very much unavoidable. The only way to adequately deal with it is to reduce the disparity in the scene. Indeed, if disparity is reduced to less than

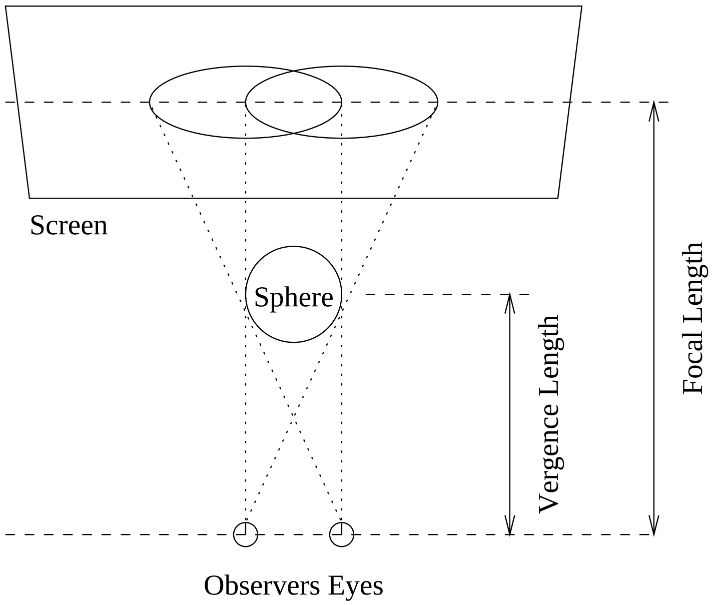


Figure 3.2: Differences between vergence and focal length

2 minutes of arc, then an observer can fuse the stereo images on a stereo display whilst being verged on the plane of the screen. This would ensure that the vergence and focal lengths of the eyes would be approximately the same and thus, reduce discomfort and fatigue caused by their differing. However, in practice, having such small disparities is quite difficult as this requires finer resolution which is especially costly for systems with large screens.

3.2.1 Coloured Anaglyph

The simplest stereo display technique is coloured anaglyph. A normal RGB image is displayed on the screen. This RGB image is a composite of left and right stereo images, where the intensity of left image becomes the red channel, and the right image becomes the blue or the blue and green channels. An observer then needs to wear glasses with a red filter over the left eye and a blue or cyan filter over the right eye [1].

The use of coloured anaglyph is quite simple and cheap, and can be applied to any display technology without the need for specialised hardware. Indeed, much testing was done during this project using coloured anaglyph stereo due to its simplicity. There are some problems with the coloured anaglyph method though, primarily related to the use of differ-

[1] This is a defacto standard, however it is not that critical as most anaglyph glasses are easily reversed.

ent colour filters over each eye.

Having a red filter over the left eye and a blue filter over the right eye for any length of time causes the individual eyes to become adapted to the different colours in each eye. This adaption is so strong that removal of the glasses is actually quite uncomfortable until the eyes readjust to normal colour levels. For small screens, such as computer monitors, areas external to the stereo display can be difficult to look at because of the colour filters. Attempting to read blue text or looking at a normal image that has areas of red and blue in it are very difficult. For this reason, for maximum effect and minimum discomfort, coloured anaglyph should be shown using the entire resolution of the screen and not in a window with colourful surrounds.

Perhaps the greatest problem with coloured anaglyph is the loss or distortion of the colour information. While it is true that some degree of colour can be present and perceived in the image, there is a substantial degradation of the colour in the scene.

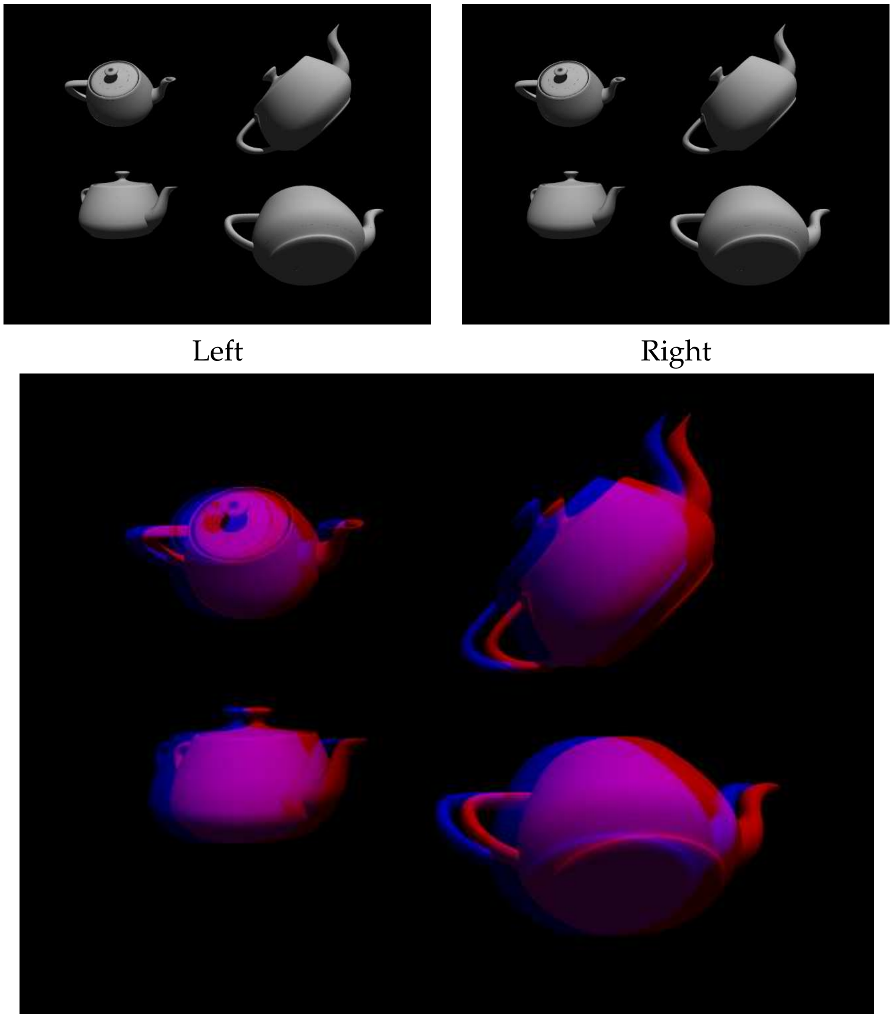
For greatest stereo effect, it is preferable to only use the intensity information from the left and right images to construct the stereo image. This means that the stereo image will contain no green, and also that all colour information from the original images is discarded. However, for all its flaws, coloured anaglyph is still the most popular and accessible stereo display format, and the only one that can be displayed on paper with relative ease, as shown in Figure 3.3.

3.2.2 Liquid Crystal Shutter Glasses

A stereo display technology, which in the past few years has become affordable and available for desktop computers, is the liquid crystal shutter (LCS) glass system. The LCS system uses a liquid crystal film which can be electrically switched to allow or block light from passing through. The physical mechanism for achieving this is linear polarisation but a more detailed description is beyond the scope of this thesis (see [82, 30] for a detailed description).

The transition time between on and off states of the liquid crystal shutter glasses is very fast, and the glasses can be cycled at greater than 100Hz. The way in which these glasses work is that the left and right eyes are alternately switched on and off so that only one eye can see the image at a time. This cycle is then synchronised to the display so that the display shows the image for the individual eye while that eye can observe it. If this is done at a sufficiently high rate, then stereo can be perceived. Rates as low as 30Hz have been tested using this technique, however the main factor in determining the speed of the switching is the desire to reduce flickering. This implies a frequency higher than the flicker fusion frequency of approximately 55 Hz in low light conditions, with higher rates being required as ambient lighting levels increase.

There are two methods used to display the stereo images on the monitor or projection screen, and these are known as field sequential and frame sequential stereo. The field sequential method uses a display mode that is interlaced. Interlaced modes display a whole screen by splitting it into odd and even rows, and instead of rendering the entire frame in one pass of the electron gun, only the odd rows or even rows are displayed in a single scan.



Left

Right

Figure 3.3: An example of a printed colour anaglyph stereo image

This has the benefit of increasing the refresh rate of the monitor, while using less bandwidth. With field sequential stereo, the odd and even fields correspond to the left and right channels of the stereo image. The standard for interlaced modes have signalling in the RGB signal for a computer monitor that indicates which field is currently being communicated. This can be used to synchronise the shutter glasses with the display so that, for example, the left eye sees only the odd field and the right eye sees only the even field. This syncronisation is usually provided by an adaptor that connects to the monitor port of the computer and either connects directly to the shutter glasses or to an infra-red transmitter which broadcasts a signal that the LCS glasses receive.

The main problem with the field sequential method is that half the vertical resolution is lost. Another fault is that there is often a gap between the scan lines due to the interlacing. If specialised hardware is not used, it is more difficult to process interlaced images using computers due to the necessity to interleave two separate images. Most of the problems of the field sequential method are solved by the frame sequential method.

The frame sequential method displays alternate frames that are the left and right images of a stereo image at very high refresh rates. This means that the full vertical resolution is available, and problems with interlaced displays are not an issue. Given that the display must have a refresh rate that is more than two times the flicker fusion frequency to prevent an observer from perceiving flickering (approximately 110 Hz), these types of systems require high quality video adaptors and monitors or projectors that have sufficient bandwidth. Also, the video adaptor must have support for frame sequential stereo, as this feature cannot easily be performed with the required speed in software. So although frame sequential stereo provides several advantages over field sequential, it requires specialised hardware to operate the display, and therefore the costs are significantly higher compared to a field sequential system.

The interocular crosstalk for LCS systems are quite low, in Silverstein et al [83] it was shown that for a white image, the extinction ratio varied between 17:1 to 11:1. The extent of the crosstalk in these systems is governed by three factors:

1. the amount of light transmitted by a shutter in the closed state,
2. the amount of time the phosphor remains bright after being struck with the electron beam, and
3. the vertical position of the image on the screen.

The colour of the image also plays a small part due to the fact that the different phosphors used in CRT display systems can have different brightness decay times. For example, the interocular crosstalk for a red image is significantly less than for a similar green image.

The main problem that faces both the field sequential, and the frame sequential LCS stereo display method is that the glasses are quite expensive. The cost comes from the fact that each set of glasses has electronics that detect the infra-red signal and control the switching of the liquid crystal films.

3.2.3 Polarising Glasses

Another stereo display method is to use polarising glasses. These are simply glasses with a polarising film in each lens which is crossed with respect to the other lens, making the glasses significantly cheaper to manufacture than the LCS glasses. This method can be used for both monitor displays and projection displays, although the way it works for these different methods is quite different.

For monitor displays, it is necessary to place a liquid crystal film over the surface of the monitor. This film is then alternately cycled between two different polarisation orientations which correspond to the orientation of the left and right lens of the glasses an observer wears. This liquid crystal film is synchronised to the display, so that it can be operated in a field sequential or frame sequential manner, using the same principles as LCS glasses system. In essence, it is equivalent to the LCS glasses system where the liquid crystal film has moved from the glasses to the screen. In terms of hardware for the driving of the display the two systems are much the same.

For projection displays, two projectors are used for a single screen. Polarising filters are placed over the lenses of each projector that correspond to the filters in the lenses of the glasses worn by an observer. Since there are now two projectors for a screen, this adds complexity to the system since images to two separate projectors must be generated by a computer. The simplest method is to use two separate video adaptors to connect to the separate projectors. The only problem that can be encountered is a lack of synchronisation between the images displayed by each of the projectors. With a modern high performance computer, this problem can easily be avoided.

The biggest problem with using polarisation glasses is that it is more common to use plane polarising filters. This means that as an observer tilts their head, the degree of blocking of the image for the opposite eye becomes less, which increases the interocular crosstalk. The limit of head tilting is between 10 and 15 degrees, in which the stereo images can be fused. It is a difficult thing to measure, since there appears to be pronounced hysteresis at the limit of fusion. If starting from a vertical head position and gradually tilting to one side until diplopia occurs, the maximum angle is significantly larger than if the head is first positioned at a large angle and gradually made more vertical until fusion occurs.

To overcome the problem of head tilt increasing interocular crosstalk, circular polarisation is used in preference to plane polarisation (for a comparison of linear and circular polarisation, see Appendix A). This requires the use of quarter wave plates in conjunction with the linear polarising film to produce and filter circularly-polarised light. Whilst this allows an observer to tilt their head to any angle and still perceive stereo, there are some undesirable aspects of this technology. The quarter wave plate used is naturally only a quarter wave for a particular wavelength of light. This wavelength is chosen optimally so that it lies in the middle of visible spectrum, approximately 450 nm. Which means that the filtering of wavelengths of light at the periphery of the visible spectrum, i.e. red and violet, is less than ideal. Hence, the amount of interocular crosstalk for these wavelengths is higher than for those closer to the quarter waveplate wavelength. A result of this is that, as an observer tilts his or her head, different colours are perceived. These different colours are the result of the linear polariser-quarter wave plate filter producing elliptically polarised light with orientations dependant upon the wavelength and angle of the linear polariser. At various angles, the tint of the crossed circular polarisers becomes blue, purple and red.

The interocular crosstalk for polarising systems varies depending on the display method used. For displays using liquid crystal films over a monitor, the crosstalk is comparible to that of the LCS glasses method, quantitatively, extinction ratios in the range of 15:1 to 10:1. For projection display systems, the crosstalk is affected by the scattering of polarisation from the screen material, so the crosstalk of these systems is slightly higher. Due to the fact that the creation of circular polarisation with a quarter waveplate is non-ideal, the extinction ratios are much lower then thay are for linear polarisation (see Tables 3.1 and 3.2) .

3.2.4 Lenticular Displays

Lenticular display systems typically operate using either two projectors or a single monitor. On the front of the screen a special lens is placed which has hemi-cylindrical lenses aligned vertically that refract the light in an attempt to direct light from the screen directly into an observer's eye. For monitor screens, alternate horizontal pixels are refracted in different directions in order to direct their light directly into an observer's left and right eyes. With projected systems, the two projectors are positioned at the rear of the screen, and a lens behind the screen is used to direct light from the two projectors to the lens on the front of the screen. The operation of a lenticular display is shown in Figure 3.4.

The primary disadvantage of lenticular displays is that, at the minimum, the horizontal resolution is reduced by half because half the pixels are used for the left eye image and half are used for the right eye image. The horizontal resolution is the most important for stereo display systems since horizontal resolution directly influences the quality of the disparity information, so this reduction is quite critical. The benefit of this system is that an observer does not need to wear any sort of glasses to perceive the stereo. All that is required is for them to be within a certain angular range of the display. This also means that multiple viewers can still see the display, all without glasses.

While much research has been performed on lenticular systems, they are still very much in their infancy and there is very little available in terms of hardware. It is possible that such systems, coupled with automatic eye-tracking to enable stereo viewing from any angle, could be the way of the future due to the fact that a user will require no glasses, particularly for applications such as stereoscopic video phones and teleconferencing.

3.3 The Robot Command Station Prototype

The Robot Command Station will use the polarisation technique with rear projection screens to give an observer a stereo view of the mobile robot's environment. Mirrors were used in order to reduce the required projection length so that the overall size of the RCS was reduced. For controlling a remote mobile robot, the vertical screen would be used for displaying video

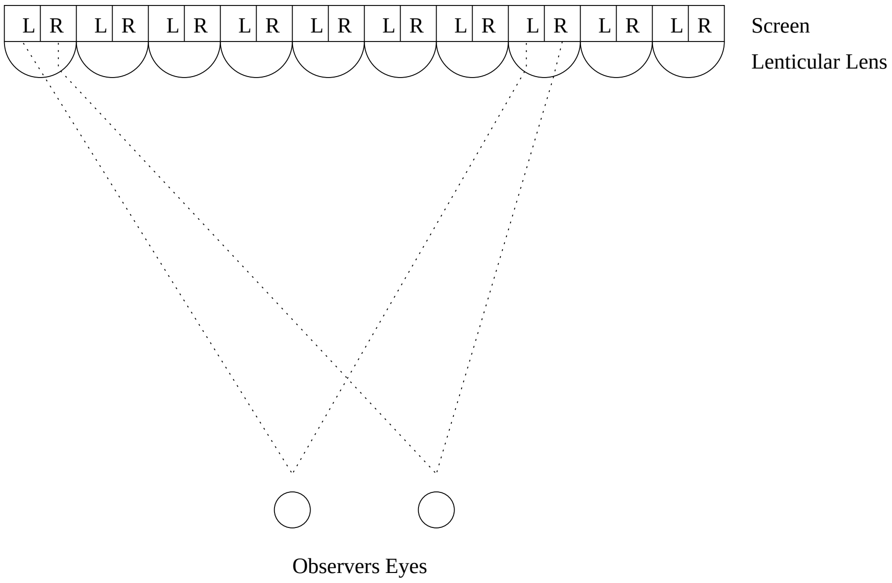


Figure 3.4: The lenticular lens and its operation

and a heads up display, while the bottom screen could be used for a virtual dashboard to provide other control information to an operator.

The screen size was set to approximately match the resolution of the projectors used. The resolution of the projectors was 800 by 600 pixels, and since the viewing distance was expected to be between 0.8 and 1.2 metres, a size was chosen so that the pixels were 1 mm square. This gives an angular resolution of between 3 and 4 minutes of arc, depending on viewing distance.

Since it was desirous to allow both an operator to sit at the RCS and control a mobile robot, and a small group to stand and observe stereo displayed on the RCS, circular polarisation is the obvious choice (This is because a group observing stereo on the RCS would stand around the bottom screen, which is equivalent to an observer tilting their head). This fact all but eliminates the use of linear polarisation for the RCS if stereo is desired on the bottom screen.

Four NEC LT84 projectors were purchased for the Robot Command Station, and a prototype was constructed to determine the feasibility of the RCS and determine the best method of constructing the final RCS. There were several factors that have contributed to the design of the RCS in its prototype phase.

Firstly, the projectors used for the RCS use Digital Light Processing (DLP) technology that was created by Texas Instruments. They operate by using a lamp reflecting off an array of micro-mirrors, the light of which then passes through a rotating colour wheel. The refresh rate of the projectors is fixed at approximately 63Hz, and, as a consequence, the colour wheel must rotate at 3–4 times this speed. A side-effect of having such a fast moving part in the projectors is that if the projector isn't mounted near horizontal (either upside down or the right way up), then the lifetime of the projector is significantly reduced [2]. This reduces the options available for the geometry of the projectors and mirrors within the RCS system. It was deemed that for the RCS, the top screen would be projected directly rather than using a mirror, which increased the size of the system quite dramatically.

The material to be used for the screens had to be determined. It was found that a rigid material needed to be used because the bottom screen was horizontal, and needed to support its own weight. This effectively eliminated the use of normal rear projection screen material which is a thin polymer film. Hence, it was decided to use either glass or perspex with one surface sand-blasted, which also preserved, to some degree, the polarisation of the rear projected light.

The sand-blasting of the screen material is necessary to cause scattering so that the image can be seen on the screen. If the scattering is too little, then the lamp of the projector becomes evident when viewing the screen (so called "hot-spots"). However, the scattering also affects polarisation so that increased scattering results in higher interocular crosstalk, which degrades the quality of the stereo display. Hence there is a trade-off between the appearance of hot-spots and the amount of interocular crosstalk in determining the amount of sand-blasting for the material.

Two different materials were tested to determine the amount of scattering of polarisation on transmission through the material. These results were obtained by measuring the extinction and transmission of linear and circular polarising filters with a particular intermediate material. Briefly, the output of the projector was polarised using a filter (either linearly or circularly) and passed through the screen material being tested. The light intensity was measured after it had passed through another polariser which was either crossed or uncrossed with respect to the original filter.

For linear polarisation, crossed refers to the planes of polarisation of two polarisation filters being perpendicular to each other. Hence, two crossed linear polarisation filters extinguish most of the light passing through and two uncrossed linear polarisation filters allow most of the light to pass through. With circular polarisation, the same terminology is used but crossed refers to the rotation direction of the polarisation (either left-circular or right-circular). Hence if two circular polarisation filter have different rotation directions then they are said to be crossed and conversely, if two circular polarisation filters have the same rotation directions then they are said to be uncrossed. It should be noted that the order of circular polarisers does not affect the extinction of light, ie a left-circular polariser followed by a right-circular polariser gives the same extinction as a right-circular polariser followed by a left-circular polariser.

A summary of the results of these experiments is shown in Table 3.1 for linear polarisation and Table 3.2 for circular polarisation. The abbreviations "S.F." and "R.F." stand for

[2] According to the user manual for the projectors

Material	Incident Light	Uncrossed	Crossed	Extinction Ratio
	(Lux)	(Lux)	(Lux)	
Air	280	160	1	160 : 1
Glass (S.F.)	247	123	16	7.7 : 1
Glass (R.F.)	252	127	16	7.9 : 1
Perspex (S.F.)	245	98	33	3.0 : 1
Perspex (R.F.)	247	103	32	3.2 : 1

Table 3.1: Extinction ratios for different materials using linear polarisation

Material	Incident Light	Uncrossed	Crossed	Extinction Ratio
	(Lux)	(Lux)	(Lux)	
Air	322	172	36	4.8 : 1
Glass (S.F.)	278	139	38	3.7 : 1
Glass (R.F.)	280	140	37	3.8 : 1
Perspex (S.F.)	267	117	47	2.5 : 1
Perspex (R.F.)	273	81	44	1.8 : 1

Table 3.2: Extinction ratios for different materials using circular polarisation

smooth facing and rough facing respectively, and indicate whether the measurements were taken with the smooth side or sand-blasted side facing the light source. It should be noted that the circular polarising filters used in this experiment where the quarter-wave plate type, which are optimal for a particular wavelength of light (chosen in approximately the middle of the visible spectrum). It is therefore not surprising to find that circular polarisation does not perform as well as linear polarisation. Several trends can be seen from these results, with respect to the extinction ratios, arguably the most important property for the screen. These can be summarised as follows:

- sand-blasted glass is less detrimental to polarisation than sand-blasted perspex,
- the rough side facing the light source is less detrimental to polarisation for sand-blasted perspex, and
- linear polarisation provides a higher extinction ratio (peak) than circular polarisation.

While having the rough side facing the projector has been shown to provide higher extinction ratios, for a high quality display the smooth side of the screen must face the projector, because with the rough side facing the projector, the image seen on the smooth side is significantly blurred. There are other practical reasons for using sand-blasted glass screens. Firstly, the screen material has to span 800 mm horizontally, which means that the screen material must be rigid enough to support its own weight over this distance with minimal bowing. Tests showed that perspex spanning this distance had appreciable bowing. It was also found that the perspex screens were quite easily damaged due to the fact that the rough surface is quite fragile. For these reasons, sand-blasted glass screens with the smooth side facing the projectors were used.

Another problem is the alignment of the projectors so that they project the same image onto the projection screen. The projectors used in this project have electronic keystone adjustment, however, this is accomplished using pixel interpolation which has a tendency to blur the image. This is not so much a problem with video or real life images, however it makes text very difficult to read and so the use of this feature of the projectors should be minimised. In order to minimise the keystoning of the display, the projectors need to be positioned optimally.

The positioning of the projectors is largely determined by the properties of the lenses on the projectors. The lens has a manual zoom facility that gives some degree of control over the size of the image. However, the projection length required for an image size of approximately 800 by 600 mm is 1.7 m, with the zoom lens able to alter this length slightly. The zoom lens, when making the image larger, expands equally on the left and right sides of the projected image, but the bottom edge of the image remains largely static whilst the top edge rises vertically to preserve the aspect ratio. The action of the zoom lens is shown in Figure 3.5.

This property of the zoom lens of the projectors means that some degree of keystoning is unavoidable. Hence, the next decision that had to be made was which orientation of keystoning was preferable. If the projectors are mounted above each other then horizontal keystoning is introduced, similarly if the projectors are mounted beside each other then vertical keystoning is introduced. A picture of horizontal and vertical keystoning effects is shown in Figure 3.6.

It was determined that horizontal keystoning was preferable for several reasons. Firstly, the projectors that are being used are significantly wider than they are high, which means that the lenses can be closer together when mounted above each other then when they are mounted side by side. This means that the keystoning produced from the projectors when side by side is more pronounced than when the projectors are mounted above each other. Also, the keystoning correction that is provided by the projectors only improves horizontal keystoning, hence vertical keystoning cannot be adjusted for. Lastly, horizontal and vertical keystoning effectively introduce noise into the horizontal and vertical stereo disparities respectively. Since the human visual system can accomodate horizontal disparities with greater ease than vertical disparities (see Chapter 2), horizontal keystoning is preferable to vertical keystoning.

Another problem with the alignment of the projectors is that the light for the top and bottom screens are projected from significantly differing angles. This has the affect of causing the top and bottom screens to appear to be of different intensities depending upon the observer's viewpoint. Without introducing further keystoning, this problem is unavoidable since the top screen must be projected horizontally and the bottom screen must be projected

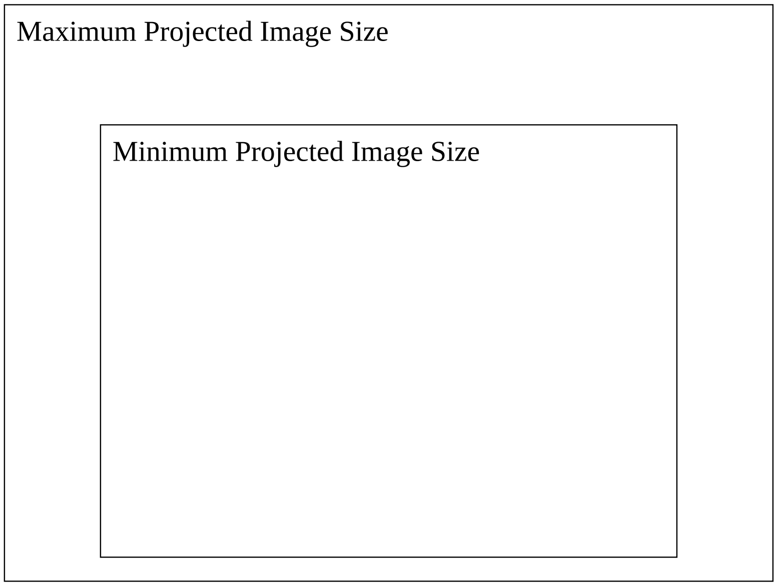


Figure 3.5: The effect on the projected image position of the zoom lens

vertically. Since the magnitude of the difference in intensity of the top and bottom screen varies as an observer's viewpoint changes, head tracking with software adjustment of intensity is required to remedy this problem. Since the initial use of the Robot Command Station is for the display of stereo video on the top screen and a virtual dashboard on the bottom screen, this is not a great problem and can be tolerated. However, for the display of virtual reality and other graphics which are displayed across both screens, this is quite a serious problem and causes a distinctive border to be perceived. Although, given the thickness of the screen (6mm) such a border is unavoidable.

For the bottom screen, a mirror was used at an angle of approximately 45 ^{o} to fold the optical path of the projectors. Early operation of the prototype Robot Command Station gave the impression that the bottom screen had noticeably higher interocular crosstalk than the top screen. Front silvered mylar mirrors, which are commonly used in projection systems due to their low weight and high clarity, were used in the RCS. Whilst the effects of this type of mirror on both linear and circular polarisation are well known (see Appendix A), it was not until close inspection of the mirror revealed a fine layer of dust had accumulated on the mirror. Subsequent measurement revealed that the dust on the mirror had a significant effect on the extinction ratio, quantitatively, the extinction ratio on the bottom screen improved from $3.3:1$ to $4.1:1$ when the mirror was blown clean.

Initial setting up of the Robot Command Station with the polarising filters showed that,

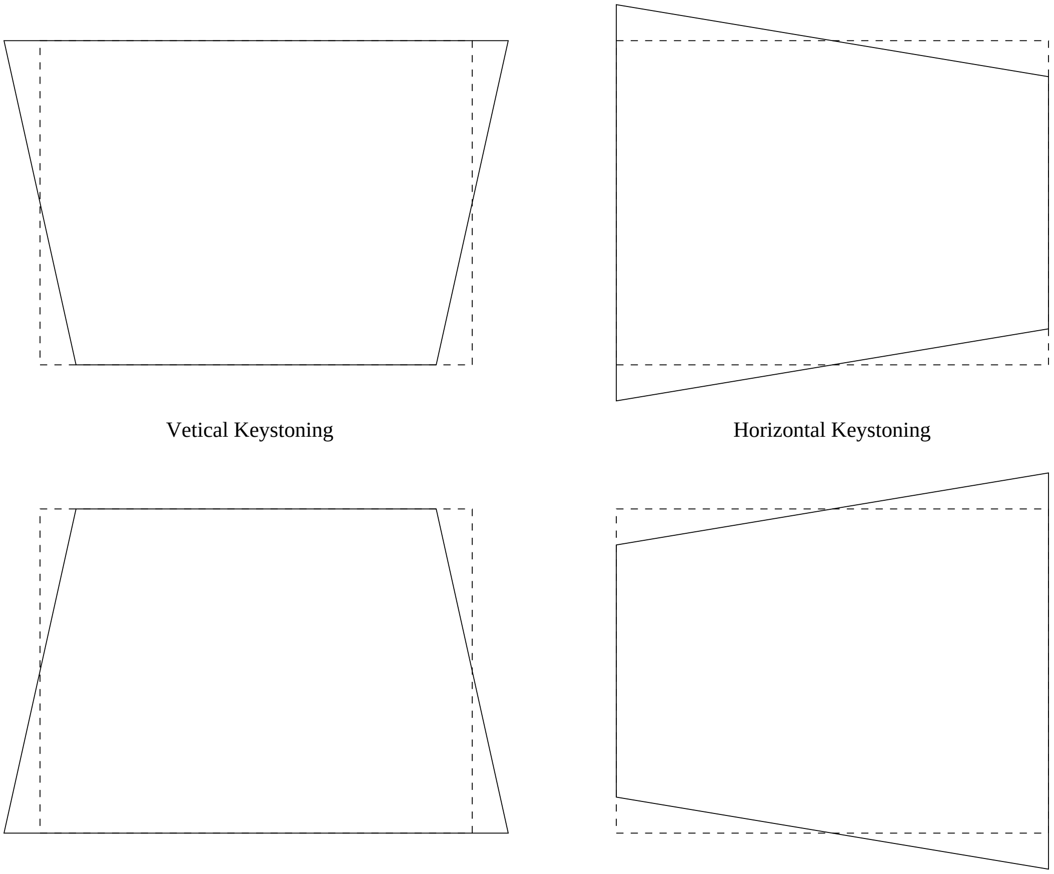


Figure 3.6: The different types of keystoning

due to the heat produced by the lamp of the projectors, the polymer polarising filters would warp if placed too close to the projector lenses. This eliminated the possibility of attaching the polarising films directly to the projector lenses. The polarising filters must be relatively close to the projectors, since at a short distance the two projection frustums intersect, at which point individual polarisation becomes impossible. In order to prevent the polymer filters from warping, they were glued to glass squares. The glass squares keep the polymer filter rigid and flat, and provides a measure of insulation from the heat of the lamp. Stands were constructed to hold the glass squares with filters attached, at a distance of approximately 5 cm from the projector lenses. While this technique of mounting the polarising film has proved successful, one problem that remains is that air bubbles in the glue cause severe distortions in the projected image. However, these air bubbles can be removed using greater care when glueing the polarising filters to the glass squares.

The final component of the Robot Command Station was the computer used to control the system. Several different hardware solutions from various vendors were discussed, however the simplest and cheapest method for the prototype was found to be a video adaptor available for personal computers. The video adaptor, a Matrox Productiva G100 MMS, had the capacity to simultaneously drive four displays at a resolution of 800 by 600 pixels. This video adaptor was fitted to a Pentium II 300MHz personal computer using the Microsoft Windows NT 4.0 operating system.

3.4 Generation of demonstrations

To test the effectiveness of the Robot Command Station with respect to the ability to display stereo imagery, some demonstrations were prepared. These demonstrations were written in the C programming language [48] using the OpenGL graphics library [79, 9]. A thorough description of the methods used for the calculation of the various parameters for the demonstrations is provided in Appendix B.

Since there is no head tracker on the RCS, a suitable static viewing position had to be specified for the perspective calculations. The point decided upon was in the plane of the horizontal centre of the screens, 600 mm above the bottom screen and 600 mm in front of the vertical screen. This point requires an observer to stand and look into the corner of the Robot Command Station to see the polygonal objects in stereo undistorted.

With the Matrox Productiva video adaptor, for four screens of 800 by 600 pixels, a virtual resolution of 1600 by 1200 pixels is used to draw on separate screens. This was logically set up so that the top half of the virtual screen was the top screen of the RCS and similarly for the bottom screen. The left and right halves of the screen then corresponded to the left and right projectors of the stereo images.

Given that OpenGL is used for the rendering of three dimensional objects, the application of this to the display of said objects in stereo on the RCS across both screens is relatively straight forward. In essence, this is accomplished by constructing the virtual world, and then rendering it from different viewpoints in order to generate the correct perspective and

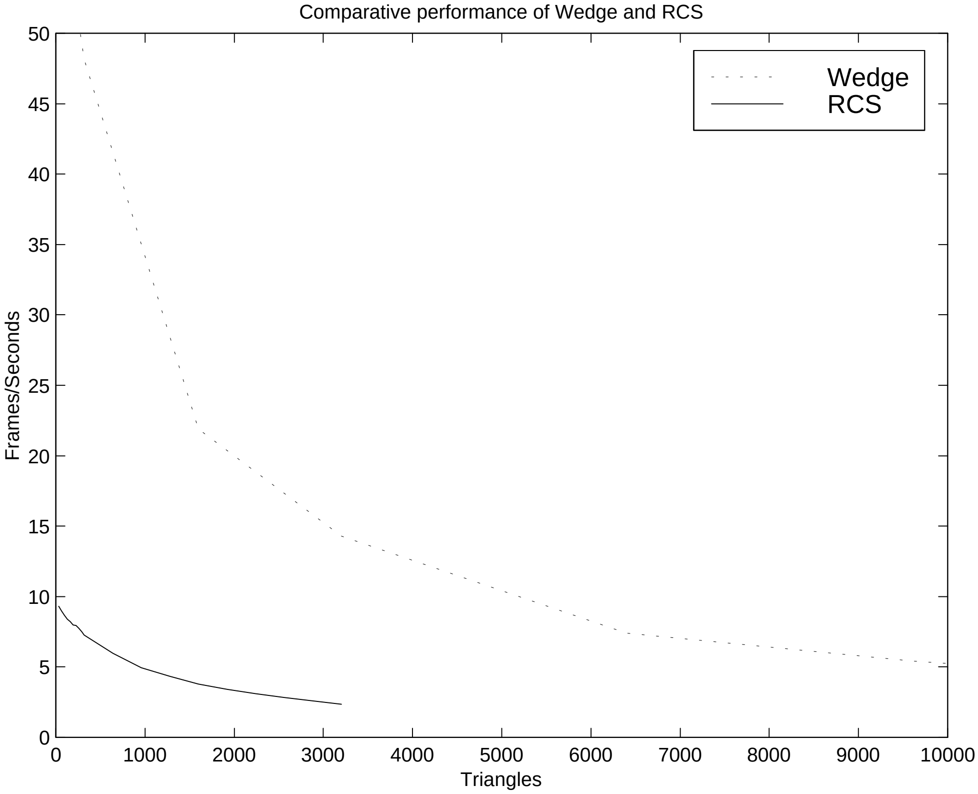


Figure 3.7: Robot Command Station Vs Wedge Performance

stereo disparity for each eye and screen. Hence, the virtual world must be rendered a total of four times, since we require a left and right view for the top and bottom screens. When displaying the four separate views of the virtual world, each of the views is rendered into the quadrant of the screen it is made for.

In terms of hardware and configuration the RCS and the Wedge are very different. The RCS uses the Matrox video adaptor which does not have any hardware acceleration for 3D graphics, whereas the Wedge has two 3D accelerated video adaptors which have both frame sequential stereo capability. The results of the performance tests are shown in Figure 3.7, and as can be expected, the Wedge shows significantly greater performance than the RCS. Also in the favour of the Wedge is the fact that the RCS is running at a slightly higher resolution than the Wedge, 800 by 600 compared to 640 by 480 pixels.

3.5 Summary

Computer generated graphics were displayed across both screens of the RCS, and stereoscopic video was displayed on the top screen with moderate success. The current problems with the RCS are:

- since a normal PC is being used graphics performance is poor, which results in unacceptable frame rates,
- circular polarisation does not provide high enough extinction ratios, which reduces the stereo quality, and
- alignment of the four DLP projectors is difficult and imprecise.

Chapter 4

Capture

The general capabilities of modern video cameras is discussed. The difficulties in the configuration of the stereo cameras is discussed and some simple methods of achieving satisfactory results are given. Different methods of interfacing stereo video to a computer are proposed and the method used in this project for the capture of stereo video into a computer is selected.

4.1 Introduction

The first step in the transmission of stereo video from a remote location to the Robot Command Station is the imaging of the environment and conversion into an electronic signal that can be interpreted by a computer. The first stage of this process is accomplished using a video camera, and, to capture stereo video, two such cameras are required. The information captured by the two cameras can then either be multiplexed into a single signal or treated as separate signals for later processing by a computer.

The cameras are used to emulate an observer's eyes, however video cameras are in many respects deficient in the capabilities that our own eyes have (see Chapter 2). The conversion of the information captured by the two stereo cameras into a signal that can be interpreted by a computer can occur in many different ways. There are differences in quality and functionality of these different methods, and these will be discussed later in this chapter.

For this project, two Sony DCR-TVR900E PAL consumer video cameras were supplied for the capture of stereo video. An attempt has been made to give a general treatment of video cameras in this chapter, however in many cases discussion is purely focused on the characteristics of these particular cameras.

4.2 Video Camera Operation

The typical consumer video camera has a multitude of functions, most of which are of limited utility. In this section, information relating to the operation of video cameras will be described. This has been separated into four different functional subsections which will, hopefully, be of a general enough nature to cover the operation of most video cameras.

4.2.1 The Lens

The lens of a video camera allows the manipulation of the field of view, accommodation distance, and the volume of light which can pass onto the sensor. These three functions are known more commonly as *zoom*, *focus* and *aperture control*, and these control the image which is formed on the sensor array and subsequently converted into electronic signals.

The focus of the lens specifies the distance at which an object appears perfectly focused. The individual lenses in the video camera lens are glass and thus have fixed focal length. By altering the distance between them, it is possible to change the overall focal length of the entire lens assembly.

The aperture performs much the same task as the iris in the human eye. The aperture is fitted in between the lenses of the camera lens assembly and alters the amount of light that can pass through the lens. This is accomplished using a mechanism that, by rotation, expands and contracts the opening in the aperture. The operation of the aperture also has a noticeable affect on the focus, as altering the aperture affects the depth of field (the depth of field is the depth in front and behind the focal distance that is still in focus). This is related to the aperture in that the smaller the aperture, the larger the depth of field (it is for this reason that pinhole cameras do not require lenses for focusing).

Lastly, most video cameras provide a zoom facility in the lenses. This simply enlarges or reduces the field of view through the use of optics. In general, the optical zoom capability is a factor of about 8 to 10 times. Many manufacturers claim figures like 180 times zoom, but this is done digitally using interpolation to scale the image formed by the lens by up to 20 times. The quality of digitally scaled images is therefore significantly less than those obtained through purely optical means. The factor given is the magnification of the lens.

4.2.2 The Sensor

The majority of video cameras use Charge-Couple Devices (CCD)[7] for the conversion of light into electronic signals. These devices are based on metal-oxide-semiconductor (MOS) technology, and the basis of CCDs used in imaging is the MOS capacitor. This component essentially converts the incident light into electric charge, which occurs when an incident photon of sufficient energy causes an electron bound to a nucleus of an atom to be freed. The free electron is stored in the MOS capacitor as charge, which can later be measured electronically.

To sense an entire image, a large number of these MOS devices are required in a grid arrangement. These are fabricated into a single device with the number of sensing elements in the device of the order of 500,000 for a video camera CCD. However, higher quality video cameras often have more than this to allow for features such as image stabilisation. For each field of video (or frame if using the camera in progressive scan mode), the array must be exposed to the image formed by the lens for a period of time, the charge stored in the individual capacitors is recorded electronically, where the amount of charge is proportional to the intensity of the image at that point, and finally the charge is removed by applying an appropriate voltage. Hence, the brightness is measured over a two dimensional grid of the image, with resolution of the video being directly related to the density of the grid and the number of sensory units in the grid.

An individual CCD cannot distinguish between colours by itself. Although a typical CCD used in photo and video applications is most sensitive to light of wavelengths around the red-yellow area of the visible spectrum (600 – 700nm), a sensing unit of the CCD merely gives a measure of the number of incident photons of light. Indeed, some CCDs are sensitive to wavelengths outside the visible spectrum.

4.2.3 Filming Colour

Since Charge-Coupled Devices cannot distinguish colour, several methods have been devised for the filming of colour, however we will concern ourselves only with the most straightforward of these. This involves using three CCDs, which are used to measure the intensities of different wavelengths of light, much like the three different types of cones in the human retina.

The structure of a video camera is shown in Figure 4.1. Light enters through the lens and is split into three separate beams through the action of a prismatic beam splitter. Each of these three beams then passes through red, green and blue filters. Finally, the red, green and blue beams are sensed by an individual CCD array, giving three different images of the scene being filmed. These separate images can then be blended to form a full colour image.

4.3 Video Camera Capabilities

As stated before, two video cameras were supplied for this project to affect the capture of stereo video. These cameras used the Digital Video (DV) format [35, 40, 65] for the storage of video onto cassette tape. The cameras used in this project use the 625-50 variant of this format, and produce PAL[1] output signals via S-Video or Composite connections.

The storage of video on the cassette tape is done using a fixed bit rate compression scheme where audio, video and data are packed into 144,000 bytes per frame for the 625-50 system. At 25 frames per second this gives a data rate of 3.6MB/s or nearly 30Mb/s. It should be noted that the 525-60 system which is compatible with NTSC has precisely the same data rate since it has 120,000 bytes per frame at a frame rate of 30 frames per second.

In the 625-50 system, there are 720 active samples per line with a total of 864 samples, and 576 active lines per frame out of a total of 625 lines. Hence, the total number of pixels is

[1] PAL more correctly refers to the colour system but in general PAL is used to signify the 50Hz television signals used throughout Australia.

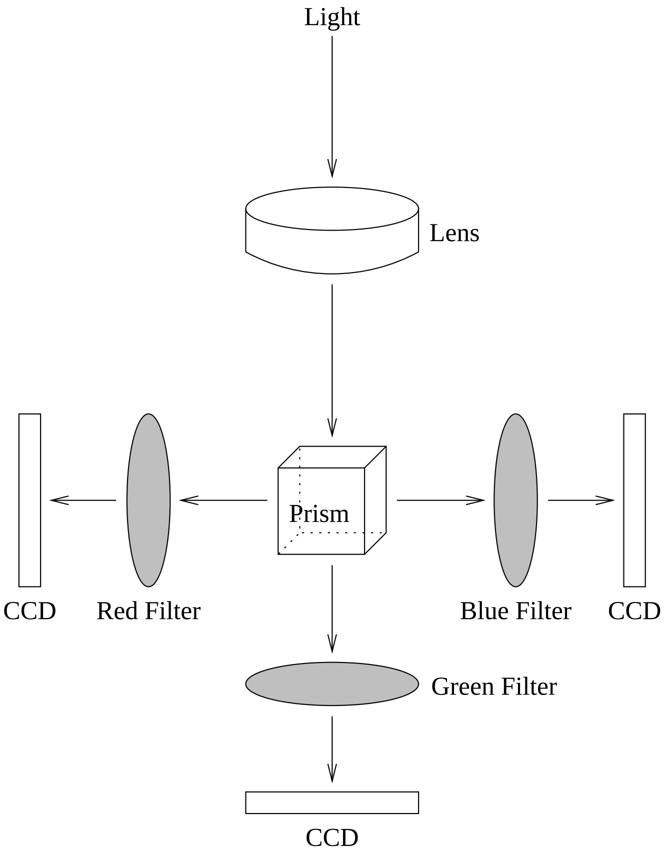


Figure 4.1: Operation of a 3 CCD video camera

720 by 540 which in computer memory in true colour would consume nearly 1.2MB, thus the compression ratio for the DV format is of the order of 10:1 while retaining very high quality.

4.3.1 Video Colour Processing

When television was first developed, the only sets available were monochrome and all transmission was also monochrome. With the creation of colour television sets, a method was required for the transmission of the television signal that was compatible with both the older monochrome television sets and the newer colour sets. The scheme created used a colour differencing approach where a luminance component and two colour difference signals were used. These colour difference signals were sub-sampled with respect to the luminance signal and quadrature modulated with the luminance signal. The result was that the modulated signal containing both the luminance and colour difference signals could be received by monochrome sets with limited interference and received by the colour television sets perfectly. As television is the predominant force behind the development of video technology, virtually all video technology uses this style of colour differencing. For a more thorough description of television and video signaling, see Poynton [61].

Most digital video standards use the ITU-R Recommendation 601 [44] described below, which specifies how colour should be encoded for digital video systems. The display of colour generally uses a different method, and therefore, encoded video must be transcoded to a different representation before being displayed.

Most colour display systems, such as cathode ray tube monitors and televisions, use the illumination of the tristimulus colours to give a full colour image. Hence, close inspection of a colour television set reveals a triangular pattern of red, green and blue dots, which at a certain distance blend into a full colour image where the individual dots aren't discernible (most computer images are represented in memory this way as well, by having for each pixel a measure of the intensity of red, green and blue, with a true colour image having 8 bits for each colour component).

In physics, luminance is proportional to the intensity of light and is a linear quantity usually represented with the symbol Y. The term luma has been introduced into the television and digital video industry to avoid confusion with this term, as luma is a weighted sum of non-linear tristimulus components. The non-linearity comes from the weighting of the spectral radiance of a scene against the efficiency curve of the CIE Standard Observer [18], which is a spectral response curve with its peak at 555nm (yellow-green). These tristimulus colours are weighted or *gamma-corrected* to account for variations in displays. The various standards for the calculation of luma from the relative intensity of the gamma-corrected tristimulus colours (labelled R, G and B for red, green and blue respectively) is shown in Table 4.1.

The calculation of luma is, in principle, the same for television signaling, however the coefficients are different. The most common colour standards for television are NTSC, PAL and SECAM, but these standards define more than just the weighting coefficients of the tristimulus colours. To reconstruct the tristimulus colours, two colour difference signals are required as well as the luma signal. Because humans have a higher tolerance for noise in the

ITU-R Recommendation 601	$Y = 0.2990R + 0.5870G + 0.1140B$
SMPTE 240M	$Y = 0.2120R + 0.7010G + 0.0870B$
ITU-R Recommendation 709	$Y = 0.2125R + 0.7145G + 0.0721B$

Table 4.1: Standards for the calculation of luma from RGB

$\mathrm{YC_BC_R}$	$C_\mathrm{B} = \frac{0.5}{1-0.114}\,(B-Y)$ $C_\mathrm{R} = \frac{0.5}{1-0.299}\,(R-Y)$
YUV	$U = 0.492111\,(B-Y)$ $V = 0.877283\,(R-Y)$

Table 4.2: Colour Difference Calculation

green component, these colour difference signals are based on the red and the blue components. Two common colour difference signals are $\mathrm{YC_BC_R}$ and YUV and the calculation of these signals is shown in Table 4.2.

To reduce bandwidth, the colour difference signals are sub-sampled with respect to the luma signal. This can be done with little reduction in perceived image quality due to the relatively poor colour acuity of human vision. The more common methods of sub-sampling of digital component video are shown in Figure 4.2. The use of 4 as the numerical basis for sub-sampling notation is for historical reasons. Of these sub-sampling methods, $\mathrm{YC_BC_R}$ 4:2:0 is used most often in computer systems, as it is this method commonly used by the JPEG and MPEG compression methods. The DV format specifies the use of 4:2:0 or 4:2:2 sub-sampling, however the the cameras for this project use 4:2:0 sub-sampling.

The highest quality image available from the cameras is therefore dictated by the DV format, and we know that in 625-50 system this is 720 by 576 pixels per frame in $\mathrm{YC_BC_R}$ 4:2:0 format. It should be noted that the fixed bit rate compression scheme used is not lossless, so there is a measure of quality reduction caused by this.

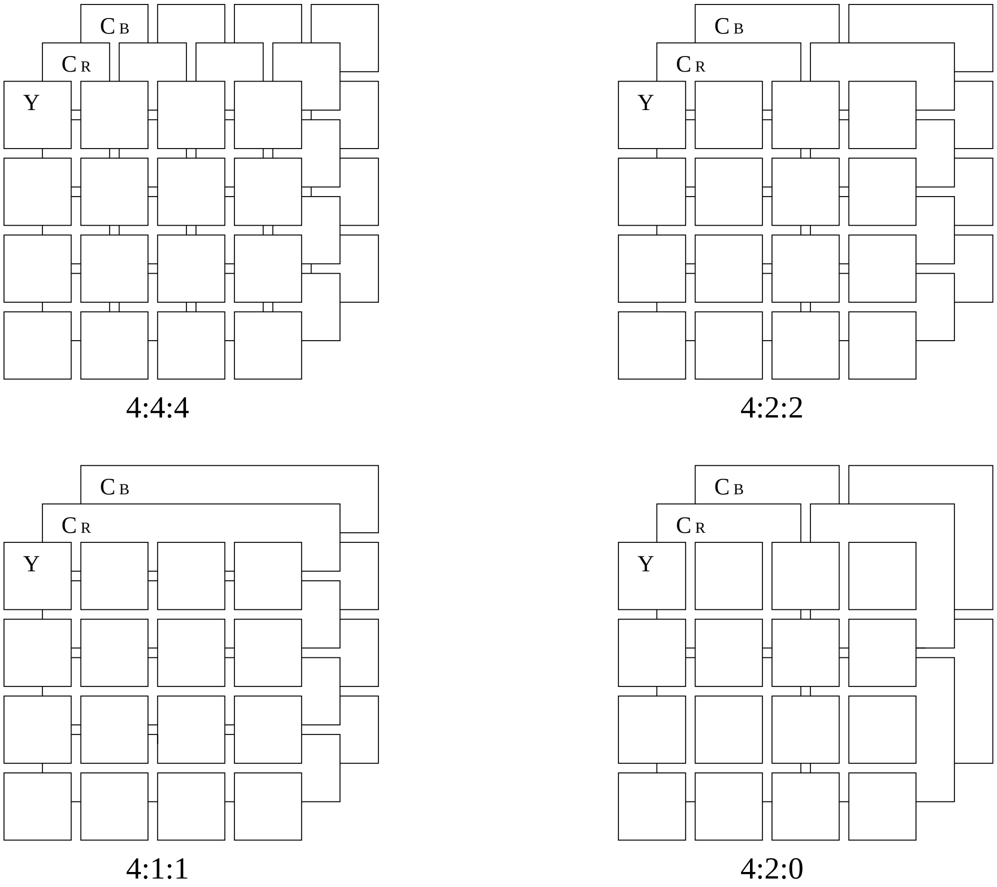


Figure 4.2: Common Video Colour Sub-Sampling Methods

4.3.2 Overview of different camera outputs

The cameras used in this project had three outputs, from which video could be received from the camera, these are:

- IEEE-1394 [36], also known as Firewire[2] or iLink[3],
- Composite Video, and
- S-Video, also known as SVHS

All of these different methods have advantages over each other, and these will be discussed in the following sections.

4.3.2.1 IEEE-1394

IEEE-1394 is a standard for a high speed serial bus, which uses bus-mastering, and a network topology. It can support a variety of speeds, such as 100 Mb/s, 200 Mb/s, and 400 Mb/s. Current revisions of the standard have maximum transmission speeds up to 1600 Mb/s. There are two different communication methods in IEEE-1394 which are known as asynchronous and isochronous. Asynchronous transmission is used for sporadic communication between nodes, whereas isochronous transmission is used for period transmission that requires a certain amount of guaranteed bandwidth. For this reason, isochronous transmission is particularly suited to the transmission of audio and video. Another benefit of the IEEE-1394 standard is that, unlike many other serial standards, there is no need for a computer or other such master device, which means that any set of IEEE-1394 enabled devices can be connected to each other.

For these reasons, many video cameras that store video and audio data digitally include an IEEE-1394 connection to enable transmission of video, to and from the camera. This enables the video to be transmitted from one device to another, in digital form (e.g. DV) without any loss due to noise.

Many computers also now have IEEE-1394 connections included as standard, and there are several vendors supplying adaptors that can be fitted to any computer. The only problem with connecting a video camera that uses the DV format to a computer is that the DV data must be decoded before the video can be seen. This is quite a computationally expensive process because the video encoding method uses the Discrete Cosine Transform (DCT).

To give an idea of the time involved, to decode a full frame of video in 625-50 DV format, there are 720 by 576 pixels of luma samples and 360 by 288 pixels of chroma samples for each colour difference component. The DCT is performed on an 8 by 8 block of samples, therefore there are 6480 blocks of luma samples and 1620 blocks of chroma samples for each colour component. Therefore, for each frame, 9720 Discrete Cosine Transforms are required. An

[2] Trademark of Apple Computers Inc.

[3] Trademark of Sony Inc.

unoptimised DCT running on a Pentium III clocked at 600MHz takes approximately 1.4ms to run which means that this would take approximately 14 seconds to decode a single frame. Through optimisation it is possible to reduce the time required to decode a single frame of DV so that a video can be decoded in near real-time.

4.3.2.2 Composite

In Composite video, the luma and colour information is summed into a single analog signal which means that the video can be transmitted using a single wire. The luma and two colour difference signals of the video are combined using the technique of quadrature modulation. The main disadvantage of Composite video is that the combining of the three signals into one causes a degree of mutual interference between the luma and colour difference signals.

As the Composite video is an analog signal, to transmit from the video camera to a computer it must go through a digital to analog conversion, transmission over a wire, and be converted back to a digital signal [4]. This process introduces noise in the video at different stages, resulting in added video degradation compared to the original DV format.

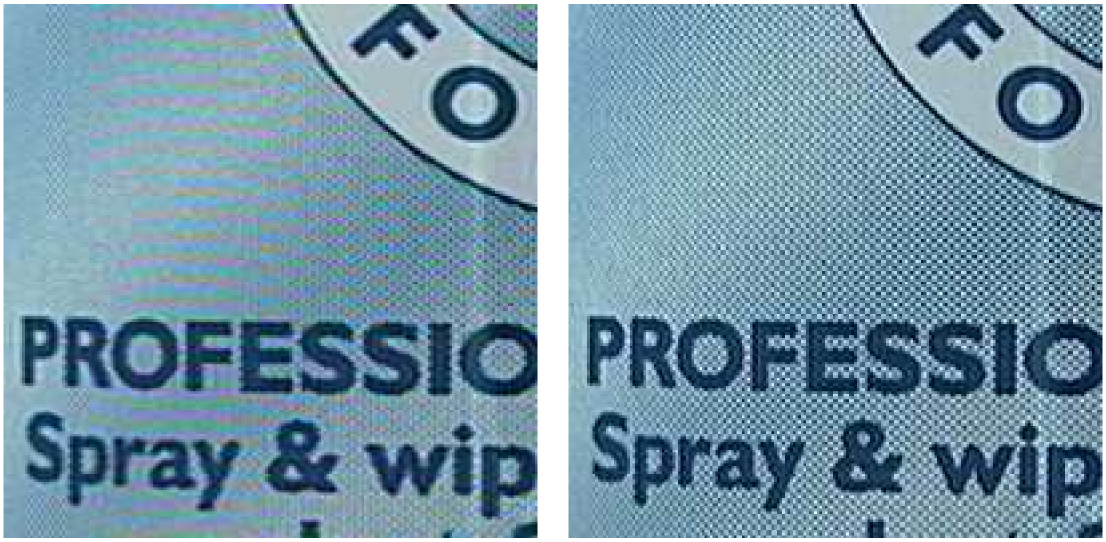
4.3.2.3 S-Video

The S-Video connection is also an analog connection, however it has a separate connection for the luma and two colour difference signals so that there is no mutual interference between the two signals. This removes the problem found in composite video, where mutual inference and quadrature modulation artefacts cause video degradation. A particular example of the difference in quality between Composite and S-Video is shown in Figure 4.3.

Like composite video, to transmit a S-Video signal from a video camera to a computer requires a conversion to analog signal and back again which introduces noise. To convert the analog signal into a digital one that can be interpreted by a computer, an Analog Frame Grabber, more commonly known as a TV Tuner, is required. These adaptors are sold by many different vendors and are able to capture from an Antenna, Composite or S-Video sources.

One of the benefits of these types of adaptors is that they can do hardware rescaling of the video image without any computational cost. They can also provide the video in many different colour spaces, and this is all accomplished with off-line processing and therefore does not interfere with the operation of the computer. If IEEE-1394 transmission is used then processing by the computer is required to convert this into an image where the image size is fixed. So while some image quality is lost by using an analog transmission method like Composite or S-Video, some functionality is gained.

[4] This is in addition to the conversion from analog to digital form the CCD to the DV format



Composite

S-Video

Figure 4.3: Comparison of S-Video and Composite Video

4.4 Stereo Video Camera Configuration

4.4.1 The Ideal Configuration

Several studies [32, 33, 80] have given a theoretical framework for the capture of stereo imagery using cameras. These are theoretical and are therefore based on an ideal model for a camera, which simply does not apply to the cameras that were used for this thesis. However, the findings of Grinberg et al [32] give a simple set of conditions for the capturing of stereo imagery for a particular display system.

Figure 4.4, which is a direct translation from Grinberg et al [32], shows the geometry of the camera configuration and the display system. The set of conditions that provide an ideal representation on the display system from the camera configuration according to Grinberg et al [32] are;

$$a = p \tag{4.1}$$

$$\alpha = \beta \tag{4.2}$$

$$\frac{f}{l} = \frac{w}{s} = \frac{2d}{p} \tag{4.3}$$

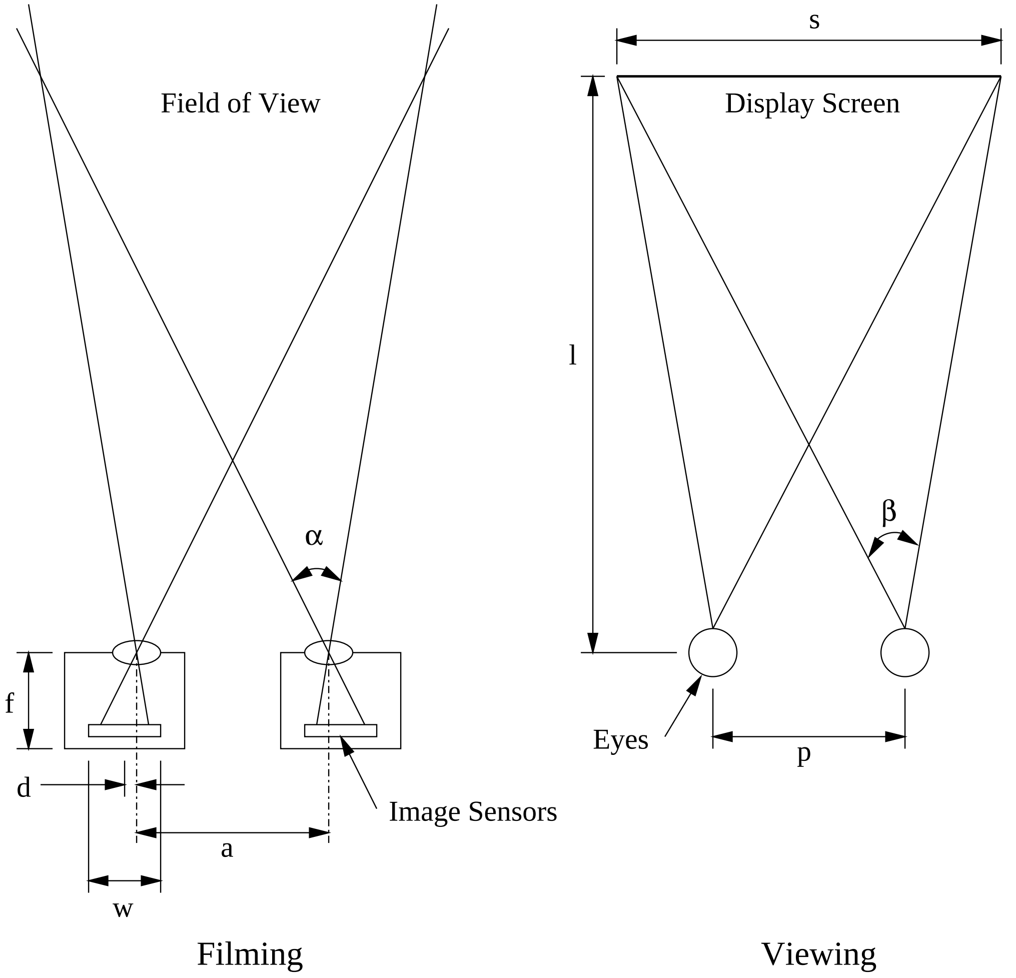


Figure 4.4: Matching the camera configuration to the display

Equation 4.1 simply states that the interpupiliary distance must equal the separation of the cameras. Equation 4.2 states that the horizontal field of view of the camera must equal the angle projected by the horizontal extents of the screen from the eye. Lastly, Equation 4.3 states that the image sensor, the CCD, must be offset by a small amount relative to the centre of the lens, and this amount varies with the viewing distance from the screen.

In humans, the distance between the eyes ranges from between 65 and 70mm. The cameras used in this project are approximately 95mm wide, which means that at best the camera separation is going to be 95mm. However, the IEEE-1394, S-Video and Composite connections are on the side of the camera, and if the cameras are placed to close together then one of the camera connections to the computer will be fouled. The smallest camera separation that can be obtained with these particular cameras while still being able to transmit video to a computer is approximately 110mm, which tends to exaggerate the depth information.

The field of view of the cameras can be adjusted using the zoom function. However, the major problem with this when using the zoom for stereo video capture is that there is no way to ensure that both cameras have precisely the same field of view. The zoom function can be controlled by a button on the camera, remote control and through the IEEE-1394 interface using AV/C commands [4]. All of these methods only enable the field of view to be altered using relative movements and there is no way to command the camera to set the field of view to x°. The only sure way to ensure that both the cameras have the same field of view is to set both of the cameras at the limits of the zoom. This assumes that the tolerances in the manufacturing of the camera are low, but it is the best that can be done.

The focal length for the stereo cameras should also match each other as it is important that the two stereo images have the same objects in the scene in or out of focus for stereoscopic fusion by an observer. There are different ways that this can be done. Firstly, both the cameras can be placed into auto-focus mode. The main problem with this is that auto-focus systems generally give greater weight to the centre of the image and, when filming stereo video, the centre of the image for each camera is likely to be different. Secondly, both of the cameras could be set to a manual focal length. This method would work quite well if the iris could be fixed to a small aperture while the exposure is still controlled automatically. Unfortunately, the cameras for this project do not have an iris priority auto-exposure mode. The last option is to use the IEEE-1394 connection and AV/C commands to effect a master-slave system, where one of the cameras uses auto-focus and the fellow camera's focal length is set accordingly. This will still be deficient in many ways since an object appearing in the master camera's field of view will dramatically affect focal length and may cause blurring of the entire scene. In the end, it was thought that having both the cameras use their own auto-focus systems was best, as this will be more likely to give a clear picture even if the focal lengths of the cameras are slightly different.

Where the focus may differ slightly between cameras, the exposure level should not as this can cause artifacts. A master-slave system would be the best method of achieving this, however the AV/C commands for controlling the exposure manually are not very flexible. The iris can be controlled by specifying the aperture size which is good, but like the zoom control the AGC (automatic gain control) can only be adjusted with relative commands.

Therefore, the best that can be achieved is to place both cameras into auto-exposure mode and accept the small differences in scene illumination that may occur.

Perhaps one of the most important aspects of the operation of stereo video cameras is the synchronisation of the image sensors of both cameras so that they image the scene at precisely the same time. This is very important for dynamic scenery because if the left and right images are taken at slightly different time, any object that is moving will have horizontal or vertical disparity (depending on the direction of movement) that will result in depth distortions or an inability to stereoscopically fuse on the object. This phenomenon was observed when stereoscopically filming a busy road with cars traveling horizontally across the field of view: at sufficient speed the disparity would be too great and a double image of the car would be seen. Unfortunately, no way was found to synchronise the cameras to each other.

It should also be noted that the horizontal position of the image sensor cannot be adjusted with respect to the centre of the lens on the cameras supplied for this project, which means that the ideal camera configuration discussed in a previous section cannot be applied. However, initial tests showed that depth is still perceivable despite all these deficiencies. It was also shown that by modifying the vergence distance and camera separation that ranges of distances could be made to give a good stereo depth effect.

4.5 The Camera Mount

A purpose built mount was constructed to fix the two cameras in a particular orientation with respect to each other. Both the separation of the optical centres of the cameras, and the vergence distance of the cameras could be adjusted using this mount. The separation distance could be varied from approximately 110mm up to 200mm, while the the angle of each camera could be oriented $\pm 15^{o}$, giving a maximum vergence angle of 30^{o}. The camera mount was also designed to be fitted to an ordinary tripod mount. The primary application of the camera mount was to determine optimum parameters for the capture of stereo video to give the most realistic perception of depth.

The camera mount, with cameras fixed in position, is shown in Figure 4.5. Some experiments were done with the camera mount for the stereo filming of different types of scenes. The two types of scenes we are mainly concerned with are landscape scenes for use with the control of a mobile robot, and near field scenes where all objects are relatively close to the video cameras.

Increasing the separation of the cameras increases the amount of disparity, and therefore exaggerates the depth of objects. This can be of some utility if the resolution of the screen is such that small disparities are difficult to distinguish, by increasing the camera separation we can enlarge these disparities to ensure that they can be perceived by an observer. However, the disadvantage of increasing the camera separation is that an increase in disparity means that difference between the accommodation distance and vergence distance is larger, and this difference is believed to be the primary cause of viewer discomfort.

The verging of the cameras can be used to alleviate this somewhat. By verging the cam-

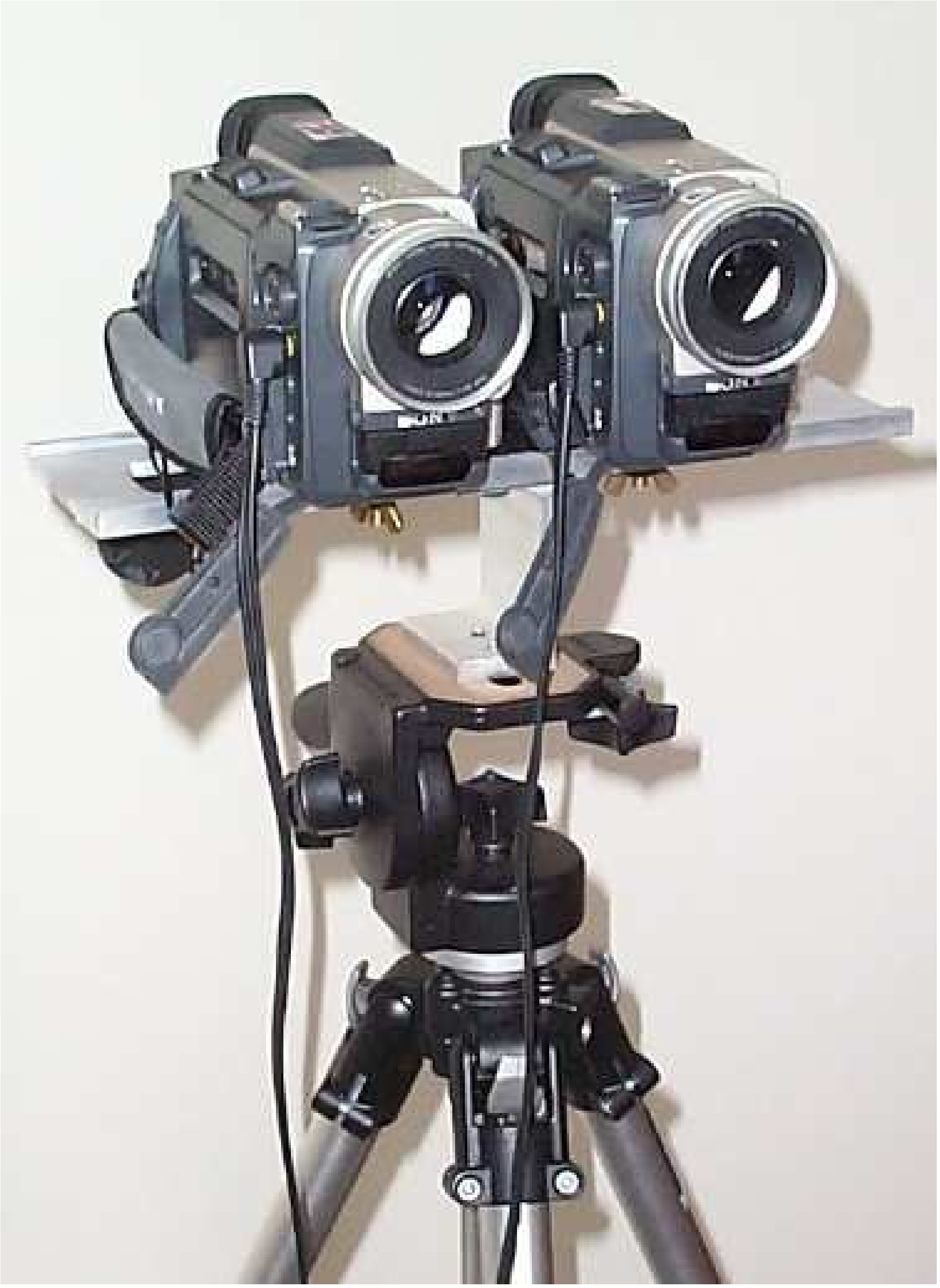


Figure 4.5: The camera mount with cameras fixed

Parameter	Value	Units
Horizontal Width	0.8	metre
Vertical Height	0.6	metre
Viewing Distance (approx.)	1.2	metre
Horizontal Resolution	800	pixel
Vertical Resolution	600	pixel
Horizontal Field of View	36.9	degree
Vertical Field of View	28.1	degree

Table 4.3: Robot Command Station viewing parameters for top screen

eras we reduce the total depth of the scene that can be fused. This is because at a certain distance beyond the vergence distance, the crossed disparity is too great to be fused by an observer. It is important to keep in mind that the display system used in this project is the Robot Command Station, which, if we view the top screen only, has the parameters shown in Table 4.3.

Given the figures for the maximum crossed and uncrossed disparity that can be stereoscopically fused, which are 4.93° and 1.57° respectively, we can determine the range of depths that can be fused given a set of camera configuration. Considering the camera configuration shown in Figure 4.6, and a display system where the positioning of an observer gives a horizontal field of view of ϕ_{fov} degrees, we can obtain approximate limits for the minimum and maximum depths that can be stereoscopically fused, these are shown in equations 4.4, 4.5 and 4.6 which are derived in Appendix C.

$$d_{\mathrm{max}} = \frac{d_c}{2} \tan\left[\tan^{-1}\left(\frac{2d_{\mathrm{v}}}{d_{\mathrm{c}}}\right) - \frac{\phi_{\mathrm{max\,crossed}}\theta_{\mathrm{fov}}}{2\phi_{\mathrm{fov}}}\right] \tag{4.4}$$

$$d_{\mathrm{min}} = \frac{d_{\mathrm{c}}}{2} \tan\left[\frac{\phi_{\mathrm{max\,uncrossed}}\theta_{\mathrm{fov}}}{2\phi_{\mathrm{fov}}} + \tan^{-1}\left(\frac{2d_{\mathrm{v}}}{d_{\mathrm{c}}}\right)\right] \tag{4.5}$$

$$d_{\infty} = \frac{d_{\mathrm{c}}}{2} \tan\left[90^{\mathrm{o}} - \frac{\phi_{\mathrm{max\,crossed}}\theta_{\mathrm{fov}}}{2\phi_{\mathrm{fov}}}\right] \tag{4.6}$$

From these equations we can calculate the range of distances that can be stereoscopically fused by an observer as in Table 4.4. It is important to bear in mind however that since it is desirable to have small disparities for the range of distances we are interested in, for certain types of scene these values are not so important. These figures allow for the verging of the eyes to stereoscopically fuse an image, and as stated before, this is one of the causes of fatigue in stereoscopic displays. The key to choosing a camera configuration is then to minimise or eliminate the need to verge the eyes to observe the scene in stereo.

For example, if a stereo scene is composed of a room sized area where the maximum depth is of the order of five metres, then the cameras should be verged at a distance close to

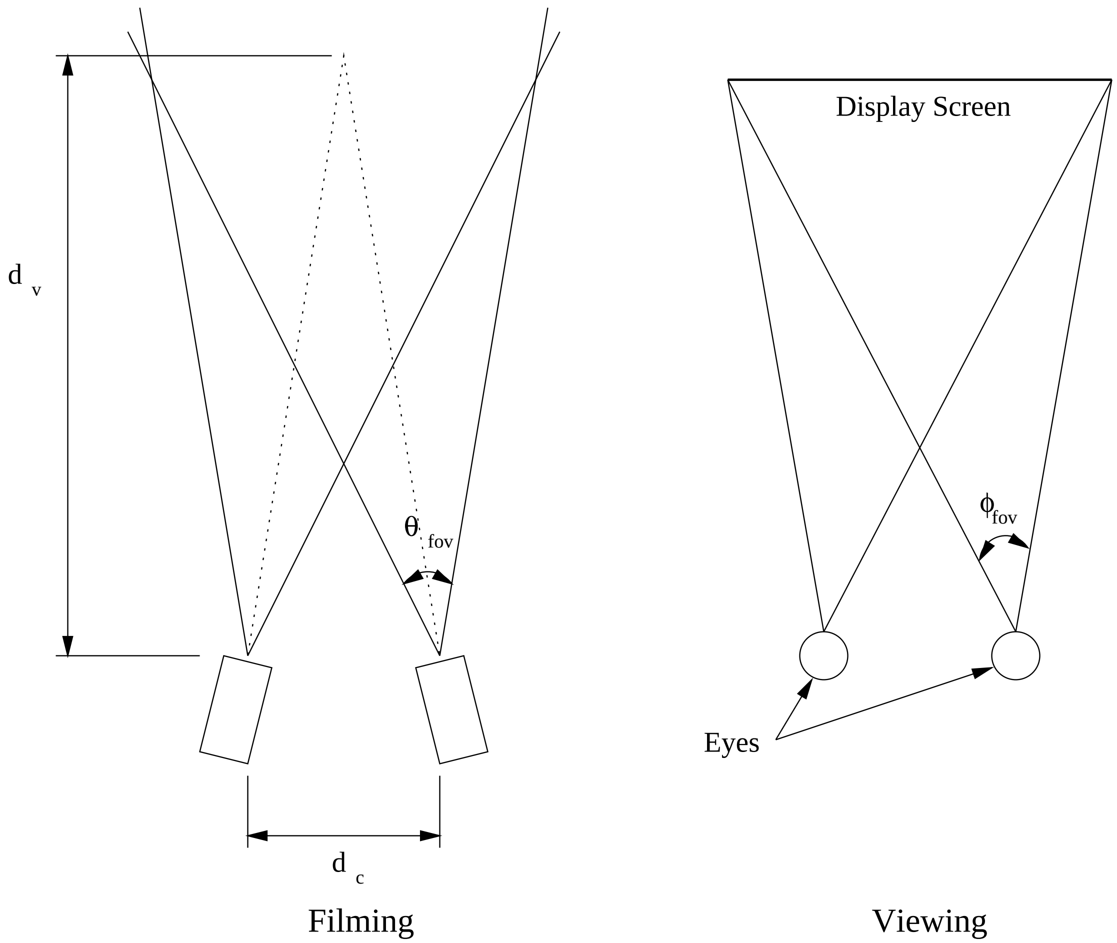


Figure 4.6: A theoretical camera configuration

Vergence	Camera Separation			
Distance	(m)			
(m)	0.050	0.100	0.150	0.200
1	0.34 . . . 2.61	0.51 . . . 1.45	0.61 . . . 1.26	0.67 . . . 1.18
5	0.47 . . . ∞	0.86 . . . ∞	1.18 . . . ∞	1.46 . . . 21.8
10	0.49 . . . ∞	0.94 . . . ∞	1.34 . . . ∞	1.71 . . . ∞
20	0.50 . . . ∞	0.98 . . . ∞	1.44 . . . ∞	1.87 . . . ∞
50	0.51 . . . ∞	1.01 . . . ∞	1.50 . . . ∞	1.98 . . . ∞
100	0.51 . . . ∞	1.02 . . . ∞	1.52 . . . ∞	2.02 . . . ∞
∞	0.51 . . . ∞	1.03 . . . ∞	1.55 . . . ∞	2.06 . . . ∞

Table 4.4: Table of fusable distances with respect to camera separation and vergence

this. An application of this is stereo teleconferencing: if the participants in the teleconference are relatively fixed in position then the stereo cameras should be verged on this position as this will be the most comfortable stereo camera configuration for an observer.

However, for scenery with large distances, the stereo cameras should be aligned parallel, to prevent vergence from limiting the maximum distance that can be stereoscopically fused. The disadvantage of this is that objects close to the cameras will not be able to be fused, but this distance can be reduced by decreasing the horizontal separation of the stereo cameras.

4.6 Capturing to Computer

Due to the high computational cost of decoding two DV video streams, the IEEE1394 connection could not be used for the capture of stereo video. This left us with the option of using analog frame grabbers connected to one of the analog outputs of the cameras. There were two methods for the capture of analog video from two cameras. We could either use two analog frame grabbers, or we could interlace the left and right camera signals into one signal. Hardware devices that interlace two streams are commercially available, and cost around $1000, but the problem with using such a device is that full resolution must be captured by the analog frame grabber in order to obtain both fields of the signal. At lower resolutions, the two fields are blended which results in a loss of stereo disparity information.

The real power in using analog frame grabbers is that we can obtain an image in computer memory of virtually any dimensions in many different colour formats without incurring any CPU overheads. For this reason, a capture system using two analog frame grabbers was chosen. As well as the two analog frame grabbers, the IEEE1394 connector was originally planned to be used for the control of various parameters of the cameras, such as zoom, focus and exposure control, however the AV/C commands are quite poor for the precise control that is required for stereo cameras, and this concept was later dropped.

4.7 Results

One of the earliest problems that was discovered with the use of two analog frame grabbers in an IBM PC, was that dual analog frame grabbers could not be used at the same time using Microsoft Windows NT 4.0 and Microsoft Windows 98. This was a severe problem that was overcome by using the Linux operating system.

It was later found that this particular problem in the Microsoft operating system was caused by using two analog frame grabbers that had the same chip-set [5]. It was found [24] that by using the television capture capability of the Matrox Productiva card and another analog frame grabber, that two simultaneous video signals could be received. However, this fact was found out late in the project and was only of relevance to the use of a Microsoft operating system. There were several other reasons why the choice of the Linux operating system was made, and it was primarly due to the fact that Linux provides a more standard set of programming interfaces to Internet related protocols.

4.7.1 Displaying interlaced stereo

Some initial tests were carried out with the capturing of analog video signals from two separate sources simultaneously. Of primary concern was the time between processing of left and right images as this can create disturbing artifacts for dynamic video footage. The method used was to set the computer monitor into an interlaced mode at a resolution of 768 by 576 pixels. From two video cameras, separate images were captured in full colour at the same resolution as the screen. These images were then interlaced with each other and then displayed so that they filled the entire screen.

To see this stereoscopically, the VREX liquid crystal shutter glass system was used. This system can synchronise to either a normal television signal (which is always interlaced) or an interlaced monitor signal. A piggy-back adaptor is placed in-line with the computer monitor cable and connected to an infra-red transmitter that transmits to the LCS glasses. Odd and even lines of the interlaced display are seen by alternate eyes, and which eye sees which field can be changed with a switch on the infra-red transmitter.

With this system it is possible to see, with the aid of the VREX LCS glasses, stereo video in near real-time. The latency of this system is approximately 300ms, with the bulk of this time spent on memory copy operations to do with the interlacing and display of the video frames. Of greater note was the fact that artifacts caused by the cameras not being in perfect synchronisation were not visible. Such artifacts were apparent in previously recorded video that was interlaced off-line and subsequently displayed on a large screen television. So perhaps, in doing much the same task in near real-time with a computer, which results in

[5] The majority of these adaptors use the Brooktree chipset

a slightly slower frame rate, these problems are reduced. Alternatively, there may be some perceptive differences between the display on the television and the display on the computer monitor relating to image size and/or refresh rate. Regardless, this simple test showed that the lack of a synchronisation mechanism for synchronising the two cameras did not have a serious effect, and did not impede the perception of the stereo effect.

The most positive aspect of this work was that it was made evident that it was possible to capture two individual video sources at the same time with little computational expense, at full frame rate.

4.8 Summary

The two main points that can be gained from this chapter are that:

- an ideal camera configuration is not achievable without construction of a purpose built system, but acceptable results can be obtained using a few simple guidelines, and
- capturing video to computer using an analog frame grabber is cheap, fast, flexible, and reduces quality only marginally.

Chapter 5

Compression

5.1 Overview

The bandwidth available from the mobile robot to the Robot Command Station is limited by the bandwidth of the Radio-Ethernet bridge, which has been previously measured at 6Mb/s. This is an upper limit of the bandwidth available to the stereo video that must be transmitted in near real-time from the mobile robot. A sufficient portion of the total bandwidth must be available for the transmission of control commands from the Robot Command Station to the mobile robot, as well as status information from the mobile robot to the Robot Command Station.

Given the capturing of video is accomplished using analog frame grabbers, the bandwidth that these devices produce from the two stereo video cameras is variable. Factors such as the resolution, and colour space can be altered and these result in a change in the number of bytes of storage required for each frame. Since we are using PAL video cameras, there are 720 active samples per line, 576 active lines and 25 frames per second. The bandwidth requirements for this resolution is shown in Table 5.1, which shows that we do not have sufficient bandwidth to transmit colour stereo video in this format.

While this is a pessimistic view of the situation, a fairly substantial reduction in the amount of data transmitted can be made. This can be in the form of data reduction or data compression. Data reduction is a very simple technique of reducing the amount of data that is transmitted while data compression is generally much more computationally complex and relies on statistical properties of the data to reduce the number of bits necessary to represent a particular block of information.

5.2 Data Reduction

Data reduction is throwing away extra data that can be considered redundant or not important. All data reduction techniques fall into two different types, quantization and sub-sampling. Quantization methods of data reduction reduce the number of bits per sample, thus reducing the signal to noise ratio. Sub-sampling reduces the number of samples temporally or spatially and this results in the loss of higher frequency information and potentially

Colour Space	Bits per Pixel	Stereo Bandwidth (b/s)
Black & White	1	20,736,000
Monochrome	8	165,888,000
YC_BC_R 4:2:0	12 [1]	248,832,000
YC_BC_R 4:2:2	16 [2]	331,776,000
Pseudo Colour	8	165,888,000
Hi Colour	16	331,776,000
True Colour	24	497,664,000

Table 5.1: Bandwidth required for a PAL stereo video for various colour spaces

can cause aliasing.

5.2.1 Quantization

The only quantization method available in stereo video is to reduce the number of bits available to represent each pixel. We know that the source data is captured using DV format which has 12 bits per pixel. However, the compression scheme reduces this to approximately 3 bits per pixel [3]. The only colour representation with less than this is Black and White which uses only 1 bit per pixel.

Some early experiments were done where a stereo image had its quality slowly degraded, down to a 1 bit per pixel black and white image. It was found that at all levels, stereoscopic fusion could occur. However, banding causes edges to be seen which don't fuse well, and dithering which is to some degree necessary for black and white images essentially erases small stereo information so that distant objects don't seem to be stereo. What these experiments showed was that reducing the image to Monochrome did not greatly affect the stereo effect, while Pseudo-colour and Black and White representations had marked effects of the perception of stereo.

Clearly, reducing the number of bits per pixel is not on its own going to result in a sufficiently low bandwidth. Quantization reduction on the raw stereo video results in immediate and costly losses in image quality, with the maximum reduction in data size being 25 percent to retain a reasonably good quality image.

[1] See Figure 4.2, 16 pixels require 16 bytes for the luminance signal and 8 bytes for the two colour signals ($(16 + 8)$ bytes $\times\ 8\frac{\text{bits}}{\text{byte}} \div 16\text{pixels} = 12\frac{\text{bits}}{\text{pixel}}$)

[2] See Figure 4.2, 16 pixels require 16 bytes for the luminance signal and 16 bytes for the two colour signals ($(16 + 16)$ bytes $\times\ 8\frac{\text{bits}}{\text{byte}} \div 16\text{pixels} = 16\frac{\text{bits}}{\text{pixel}}$)

[3] A 720 by 576 pixel frame requires 144,000 bytes. It should be noted that the majority of the 144,000 bytes is video data, but a fair proportion is audio and miscelaneous other data [35].

5.2.2 Sub-sampling

Sub-sampling, as stated before, can be done spatially or temporally. In stereo video spatial sub-sampling means that the number of samples per scan-line and the number of scan-lines is reduced. Whereas temporal sub-sampling results in a reduction of the number of frames per second. We have already stated that in order to convey motion, a minimum of 10 to 15 frames per second is required, restricting the amount of temporal sub-sampling data reduction to approximately 50 percent.

The amount of spatial sub-sampling that can be done is limited to some degree by the size of the display and the viewing distance. If the size of the display is small, then the resolution required is smaller and therefore a higher degree of sub-sampling can be tolerated. Similarly, as the viewer moves farther away from the screen, a higher degree of sub-sampling can be tolerated. The operator of the Robot Command Station will be quite close to a relative large screen, quantitatively, 1.2m away from a 0.8 by 0.6m screen. As such the resolution cannot be degraded too far.

Another form of spatial sub-sampling is to have the left and right channels of the stereo video sub-sampled with respect to each other. That is, if the resolution of the left channel is 720 by 576 pixels, then the right channel can be sub-sampled to 360 by 288 pixels to reduce the data size. Some initial experiments were carried out using this technique. It was found that the results were not satisfactory, as the low resolution of the right image was quite evident. While this technique was not used in subsequent investigations, it should be noted that the method used for displaying the sub-sampled image at a higher resolution was the zero order hold method. This is the most simple interpolation method and therefore it produces mediocre results, and it may be possible to produce a higher quality stereo effect by using a more complex interpolation scheme. Regardless, the main thing to note is that it is possible to stereoscopically fuse left and right stereo images that are at different resolutions.

While spatial and temporal sub-sampling of the stereo video can reduce the amount of data transmitted by a small factor, another benefit is that by sub-sampling, there is a lower computational cost required at a later stage for compressing the stereo video. It is clear however that data reduction on its own will not be sufficient to enable the transmission of stereo video over the Internet at speeds less than 6Mb/s without severely degrading the image and stereo quality.

5.3 Existing Compression Techniques

There are several different ways to compress data. What these compression schemes do is replace the binary sequence contained in a block of data with a more compact representation. Some compression schemes result in a loss of data (which are called lossy compression schemes) and are quite commonly used for compression of speech, audio and video data types as these can accommodate some degree of data loss. The more general compression schemes used for the compression of arbitrary types of data are lossless, in that the reconstructed binary sequence is exactly the same as the original. As can be expected, the compression ratios for lossy compression schemes are usually much higher than those of lossless compression schemes.

In the following sections, the methods underlying some of the common compression standards will be described. Rather than concentrating on the properties of each standard, the focus will be on the fundamental algorithms that are used to reduce the number of bits necessary to represent the original data.

5.3.1 Generic Compression Methods

5.3.1.1 Run Length Encoding

One of the simplest methods for compressing data is known as Run-Length Encoding (RLE). This method takes advantage of repeating data by encoding it as a tuplet consisting of the number of repetitions and the repeating symbol or number. Run-length encoding used exclusively generally only performs well on highly correlated data.

An example of an image that compresses well with run-length encoding is a low colour cartoon style image or a black and white line drawing. Both of these styles of image contain large areas with the same colour pixels. This means that a great deal of information can be represented with a single tuple. However, true colour images typically have subtle changes between adjacent pixels meaning that large runs are rare in such images and therefore compression performance is significantly less.

5.3.1.2 Huffman and Arithmetic Coding

Huffman coding, uses a table of bit sequences which maps a variable length bit sequence to a particular token. For Huffman coding to compress data, shorter variable length codes must be used for tokens with the highest probability of occurring in a particular bit-stream. For this reason, Huffman coding usually requires the bit-stream to be pre-scanned to determine the relative probabilities of tokens in the bit-stream. Alternatively, several samples may be scanned to develop an average histogram of the probability of certain tokens occurring in a bitstream. Arithmetic coding uses a similar principle to Huffman coding, but does not require a lookup table for encoding or decoding.

5.3.1.3 Substitutional Compression

Many of the lossless compression schemes fall into the category of substitutional compressors, such as Ziv-Lempel code and related methods. These styles of compression schemes are also know as dictionary compression schemes. The term “Dictionary Compression” comes from the analogy that a sentence can be made shorter by substituting an appropriate word for a particular phrase.

5.3.2 Spatial Transform Methods

Generic compression methods use a purely statistical method to reduce the number of bits to represent a particular block of information. They rely on the data being composed mostly of a small number of symbols, with other symbols occurring infrequently, to gain high compression ratios. Given the fact that video varies temporally and spatially, there is rarely a case where a particular token is more prevalent than another. For this reason, generic compression methods in general, do not compress video well.

Hence, several different methods have been devised for the compression of digital images and video. These methods involve the extra step of spatial transform, which converts an image or part thereof from a spatial representation to a pseudo-frequency representation. Of the many different transforms available, only two will be discussed as these seem to be the most prevalent for the compression of digital images and video.

5.3.2.1 The Discrete Cosine Transform

The Discrete Cosine Transform (DCT) is a derivative of the Discrete Fourier Transform (DFT), widely used for the temporal analysis of many different types of time varying signals of many dimensions. Given an input block of image intensity values, the DCT converts this into an equivalent pseudo-spatial frequency representation of the same dimensions. This aids in the compression of images because the human visual system response is very dependent on spatial frequency and therefore we can eliminate unnecessary high spatial frequency components while still retaining a high quality image.

There are several compression methods that use the DCT as the basis of the compression scheme. Some notable examples include JPEG (Joint Photographic Experts Group) [41, 57], MPEG(Motion Picture Experts Group) [42, 43] and H261/3 [45, 46].

The DCT, as used in the various image and video compression schemes, is performed on an 8 by 8 array of values, where each value may luminance value or a colour value. For an 8 by 8 block, the formula used to calculate the DCT is shown in equation 5.1, with the inverse transform shown in equation 5.2. Luminance and colour data are processed using the same equations.

$$F\left[i,j\right] = C\left(j\right)C\left(i\right)\Sigma_{y=0}^{7}\Sigma_{x=0}^{7}\left(f\left[x,y\right]\cos\left(\pi j\frac{(2y+1)}{16}\right)\cos\left(\pi i\frac{(2x+1)}{16}\right)\right) \tag{5.1}$$

$$f\left[x,y\right] = \Sigma_{i=0}^{7}\Sigma_{j=0}^{7}\left(C\left(j\right)C\left(i\right)F\left[i,j\right]\cos\left(\pi j\frac{(2y+1)}{16}\right)\cos\left(\pi i\frac{(2x+1)}{16}\right)\right) \tag{5.2}$$

$$\text{where } C\left(x\right) = \begin{cases} \frac{1}{2\sqrt{2}} & x = 0 \\ \frac{1}{2} & x = 1 \ldots 7 \end{cases} \tag{5.3}$$

The property of the DCT that makes it applicable to compression is that it produces

$$\begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

Table 5.2: Visual weighted coefficients for luma

uncorrelated coefficients. In continuous tone images, i.e. real life images, pixels near each other will often have predictable differences between their values. The DCT transforms this into a series of uncorrelated coefficients that can be compressed more effectively, in either a lossless or lossy way.

There are many simple methods that can be used to compress the DCT coefficients with a controllable amount of information loss. For example, setting to zero a specified percentage of coefficients with the lowest magnitude will often result in only a few coefficients with magnitudes not equal to zero and therefore will compress well. However, the most common method used to compress the DCT coefficients is quantization. Quantization reduces the number of bits required to represent each coefficient. Typical images and video use 8 bits to represent each luma or colour component of a pixel, which means that 8 bits are also commonly used for the representation of DCT coefficients. We can reduce the number of bits used for each coefficient using a simple technique such as dividing each coefficient by a power of two. However, a series of visually weighted coefficients have been developed for the JPEG compression standard which weight each coefficient relative to its effect on human perception of the image. This means that certain coefficients are less important for the overall image and can be quantized further in order to achieve greater compression with minimal impact on quality. By quantizing the DCT coefficients, we effectively decrease the signal to noise ratio and therefore introduce errors in the reconstructed image. The most prominent of these errors are the so called blocking artifacts, which, in severe cases, result in the 8 by 8 block of pixels having a distinct border.

The visually weighted coefficients, as standardised in the JPEG standard, are shown in Table 5.2 for luma coefficients. The chroma coefficients follow a similar pattern, but due to the relatively poor acuity of human colour vision the coefficients are de-emphasised further.

As can be seen from these tables, much greater emphasis is placed on the coefficients closer to the [0,0] position. In essence, quantization has the effect of performing a low pass filter on spatial frequency, with the pass band decreasing as the quantization level is increased. In the JPEG standard, a user specified value ranging from 5 — 95 is specified to indicate the quantization level to use. From a user's perspective, this value appears as a quality factor usually expressed as a percentage, i.e. 100% indicates a lossless compression and 0% would presumably erase all information from the image. However, this value is not a percentage at all, it is combined with the visually weighted coefficients to obtain the actual quantization coefficients for encoding and decoding the image, the formula used in the JPEG standard is shown in equation 5.4 with q being the quantization factor with a useful range of 5 — 95.

$$c\,[i,j] = \frac{C\,[i,j] \times \mathcal{M} + 50}{100} \tag{5.4}$$

$$\text{where } \mathcal{M} = \begin{cases} \frac{5000}{q} & q < 50 \\ 200 - 2q & \text{otherwise} \end{cases} \tag{5.5}$$

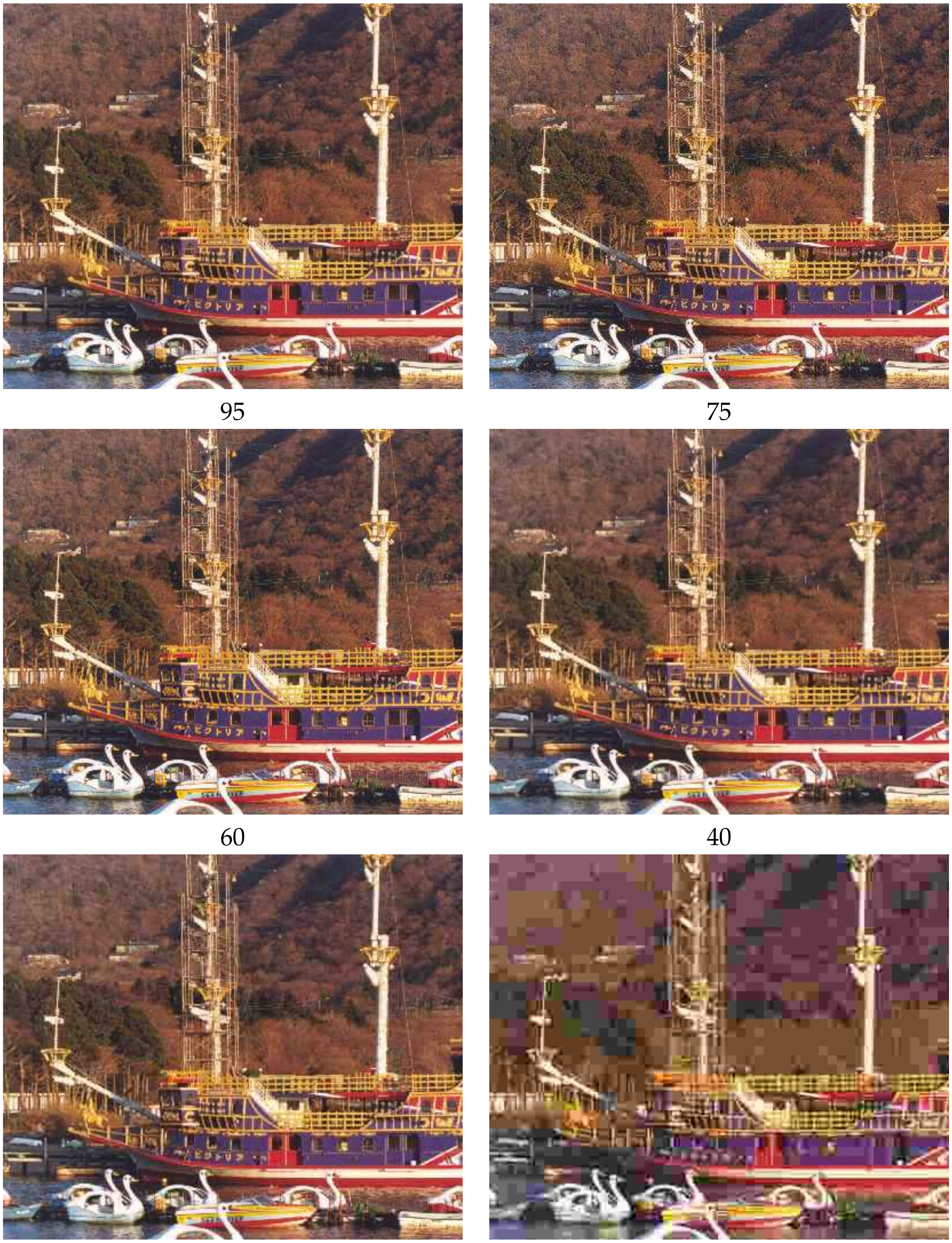
A mosaic of JPEG compressed images is show in Figure 5.1 that illustrates the effect of quantization of the DCT coefficients on the overall quality of the image. Quantization of the coefficients results in many of the coefficients being equal to zero. This allows greater compression to occur. The JPEG standard then transcodes the coefficients into run-length encoded tokens, which are subsequently compressed using either Huffman or Arithmetic encoding[4].

Figure 5.2 shows the performance of the JPEG compression scheme which gives an indication how effective the DCT is in the compression of images and video. These results have been obtained from the compression of a single image, and as the effectiveness of DCT compression varies from picture to picture, these results are indicative, not quantitative.

Since the use of the DCT for the compression of images and video has become widespread, several fast algorithms have been developed for computing both the forward and inverse DCT. The fastest algorithms, such as that described by Arai, Agui, and Nakajima and described in Pennebaker and Mitchell [57], uses fixed point arithmetic. Fixed point arithmetic can be implemented using integer operators which can be performed much faster on computers than equivalent floating point calculations. The only disadvantage with using fixed point arithmetic is that a degree of error is introduced into the calculation. For highly quantized images, this is of only minor concern because error introduced by the quantization will be significantly higher than that introduced by the fixed point arithmetic.

Overall, the computational requirements for the DCT are quite small. Since the computation is limited to an 8 by 8 block of pixels or coefficients, this alleviates the fact that the DCT is an $\mathcal{O}(n^4)$ operation, that is, the time taken to compute the DCT is proportional to the size of the block to the fourth power. Hence, by keeping the block size small, the computation cost is reduced. Of course, as the image size increases, computational cost will subsequently increase. Preliminary tests showed that full colour images could be compressed at 25 images per second at a resolution of 384 by 288 pixels on a Pentium III 600Mhz PC.

[4] The JPEG standard specifies use of either method is acceptable



95

75

60

40

25

5

Figure 5.1: The effect of quantization on image quality

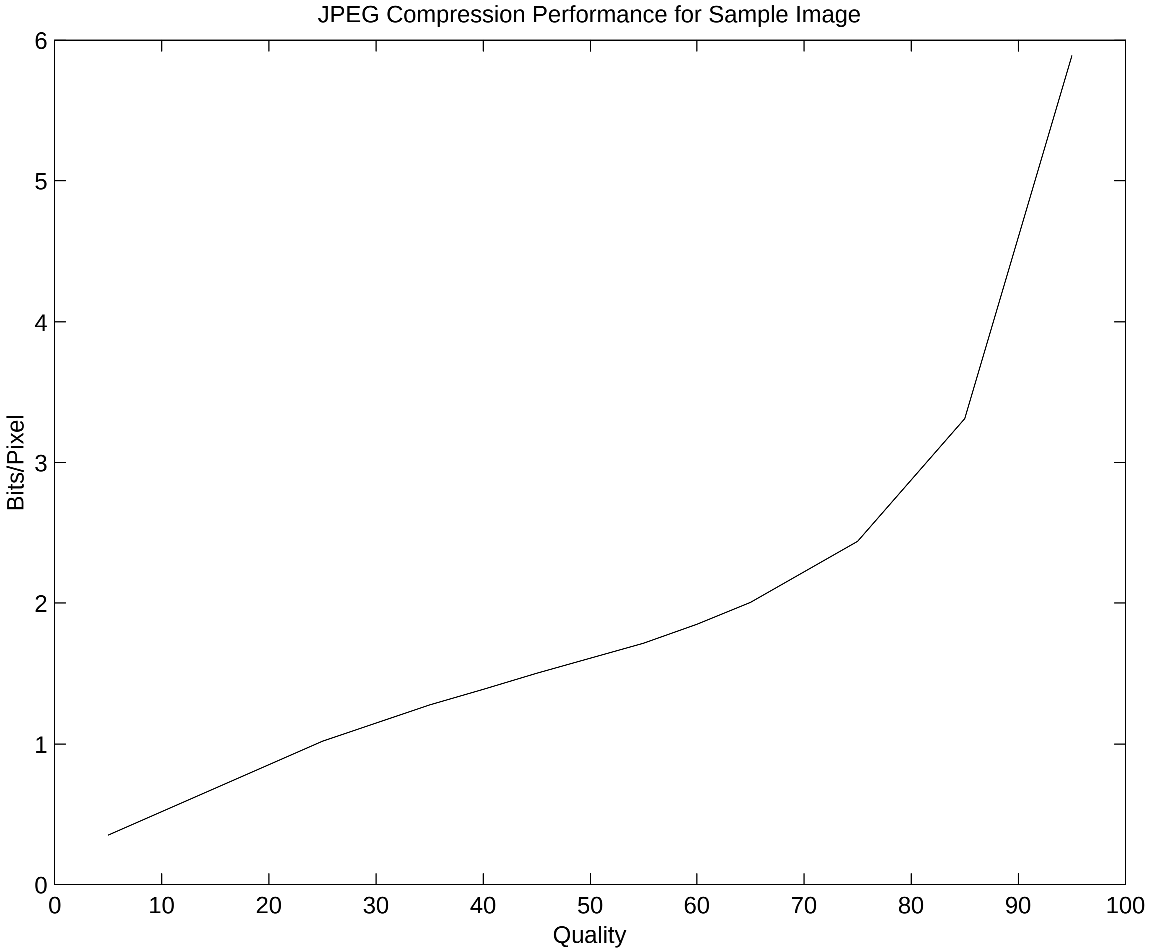


Figure 5.2: JPEG Compression Performance

5.3.2.2 The Wavelet Transform

The Discrete Wavelet Transform [5, 26, 31, 72] is an orthogonal transform which is applied to a set of sampled data and is functionally similar to the Discrete Fourier Transform. The primary difference between the Fourier and the Wavelet transforms is that the Fourier transforms use sinusoidal functions as its basis functions whereas a wavelet basis function is defined by the following recursive difference equation:

$$\sigma\left(x\right) = \Sigma_{k=0}^{M} c_k \sigma\left(2x-k\right) \tag{5.6}$$

Where M specifies the number of non-zero coefficients and is commonly referred to as the order of the wavelet. There are an infinite number of basis functions for wavelets, however, most of these are of limited use. There are several different families of wavelets, usually named after their discoverer, for example, Daubechies, Coiflet, Haar, and Antonini.

The Discrete Wavelet Transform is analogous to the Fast Fourier transform in that it is a fast linear operation that operates on a vector of data whose length is a power of two. The algorithm used for this computation is commonly known as the "Pyramid Algorithm". A newer method for computing the forward and inverse discrete wavelet transform is known as the lifting scheme [13, 14, 15, 17, 20, 28, 70, 69, 68, 71, 75, 76]. This method is much simpler and has the potential to be much faster as it can be implemented using integer operators.

The wavelet transform has the potential to perform much better than the discrete cosine transform in the compression of images. The main reason for this is that the wavelet transform is performed over the entire image, which means that there are no 8 by 8 boundaries and therefore, no blocking artifacts. The greatest disadvantage of the wavelet transform is that it is more computationally expensive compared to the DCT.

5.4 Commercial Solutions

Although there are several commercial packages available that can be used to compress and stream video and audio over Internet based networks, none of these as yet have the capability to specifically handle stereo video. A brief description of some of the commercial solutions examined will be given. It should be noted that the different packages presented here are but a sample of the wide variety of packages available. Also, these commercial packages are capable of storing/streaming several different types of data such as audio and video. They are not individual compression schemes, rather, they are content management systems capable of compressing a wide variety of data.

5.4.1 RealNetworks

RealNetworks [3] provide a range of products that can be used for the streaming and playing of video, audio and other multimedia data. For the compression of video, RealNetworks use a proprietary standard and the exact compression scheme used is unknown. However, when observing RealNetworks video, it does appear to be based on a spatial transform method such as the Discrete Cosine Transform or the Discrete Wavelet Transform. With the RealNetworks software, it is possible to scale the bandwidth used by the transmission from modem speed, 28.8 kb/s, up to a local network speed of 10 Mb/s.

For the streaming of video, RealNetworks uses the Real-Time Streaming Protocol (RTSP) Internet Standard [64]. It also uses the Synchronised Multimedia Integration Language (SMIL) for the integration of different types of real-time data such as video and audio. Although the products that are produced by RealNetworks, namely RealPlayer, RealProducer and RealServer, are perhaps the most popular media streaming technology that is currently available, they do have some shortcomings in their applicability to this project.

Firstly, before a video is displayed, it is buffered to account for fluctuations in the available bandwidth for the transmission of the data which is typical on shared networks. This buffering typically amounts to about 30 seconds of video, which means that the latency of the video is a little more than this. Efforts were made to try and minimise this buffering, however a latency of approximately 10 seconds was the best that was achieved.

It was also not clear how to transmit stereo video with the RealNetworks system, as the software has no provision for such a system. This means that it would be necessary to write software that would interface to the RealServer software to provide stereo video input and processing. Also, the RealPlayer software would have to be altered in order to display the stereo video properly. There is a software development kit that is available so that such alterations can be implemented over the RealNetworks system. Although the RealNetworks software provides a reliable solution to the streaming of video and other real-time data, further investigation was abandoned primarily due to the high latency inherent in the system.

5.4.2 QuickTime

QuickTime [2] software was originally developed by Apple to handle various types of time-sequential data, such as video and audio, in a generic way. The most popular use for this software was for the storage and playback of video and audio data. With increasing bandwidth and lower cost of the Internet, QuickTime was extended in version 4 to allow streaming of movies using the Real-time Transport Protocol (RTP) and the Real Time Streaming Protocol (RTSP).

QuickTime (version 4) supports the streaming of many different types of audio and video. However, the Internet Engineering Taskforce (IETF) has only standardised the streaming of a select few of the video compression methods, namely:

- H.261 [74]
- H.263+ [10]
- DVI
- JPEG [8]

Unfortunately, QuickTime is optimised for the streaming of pre-recorded video. The general method for creating a streaming movie is to create a QuickTime file, and then preprocess it to place a hint track in the file which is later used for the "packetising" of the data as it is transmitted using RTP. QuickTime is the most open of the three commercial solutions presented here, in that it has the most well defined and documented software interface.

5.4.3 Windows Media

Windows Media is software, similar in functionality to QuickTime, developed by Microsoft. It is a far more closed system than QuickTime in that the software libraries used for controlling various aspects of the Windows Media system are not documented or hidden. It supports several different compression methods for audio and video data, and streams these using its own proprietary format.

One of the biggest problems in using Windows Media is that it only operates on a Microsoft Windows operating system. As stated previously, it was not possible in the early stages of this project to successfully operate two analog frame grabbers running a Microsoft Windows operating system. Although this problem was later solved, it all but precluded the use of the Windows Media system from this project.

Again, like QuickTime, Windows Media is optimised for the streaming of pre-processed video and audio. It has tools to convert normal video and audio files such as WAV, AVI, QuickTime, and MP3, into Active Streaming Format (ASF) streams or files. It is possible with Windows Media to produce live transmissions, however, no testing was done with this system to determine the applicability of this to the streaming of stereo video.

5.4.4 Summary

Although there is no reason that prohibits the use of either QuickTime or Windows Media for the compression and streaming of stereo video over the Internet; it is clear that some form of software will need to be written to perform various aspects of the compression that are unique to stereo video. It was decided that rather than tie any software written for this purpose to a particular platform, it would be better in the interim to write stand-alone software independent of any commercial software available.

5.5 Existing stereo image compression techniques

There are some compression schemes, or more correctly some image formats, that already support to some degree, stereo images and video. However, most of these are relatively trivial attempts which do not provide good solutions to the problem of compressing a stereo image pair well by using the similarity between left and right images.

5.5.1 JPS

JPS is a defacto image format that uses the JPEG compression standard to compress a stereo image pair. It is a trivial method, since both left and right images are joined side by side to form a single image, which is then compressed as a normal monoscopic image. This has the advantage of retaining the full resolution of the image at the expense of extra computational cost and a lower compression ratio. JPS files are still valid JPEG files, to the extent that if they are opened with a normal image viewing application, the left and right stereo images are seen side by side.

5.5.2 Interlaced GIF

Another image format that can be used for the storage of stereo images is Interlaced GIF (Graphics Interchange Format). With Interlaced GIF images, a stereo image pair are interlaced together to form an image that is the same dimensions as the original images. GIF was an image format created by Compuserve, which predates the JPEG image format. It is a simple image format that only allows 256 different colours which are specified using a palette. The image is compressed using LZW compression, so the compression performance is quite poor for real-life images with continuous tone changes. Another drawback of this compression format is that half of the vertical resolution is lost in the interlacing process.

5.5.3 Coloured Anaglyph

Another method of compressing a stereo image pair is to convert it into a coloured anaglyph image and then use normal monoscopic image compression methods. The main problem with this is that in converting to anaglyph format there is potentially a large loss of colour information. Also, many image compression formats use the fact that in human vision, intensity is more important than colour, so colour information is often sub-sampled. Such sub-sampling in a coloured anaglyph image would degrade the stereo effect of the image.

5.6 Proposed Stereo Compression Techniques

5.6.1 Disparity

One of the early ideas was to use a correlation algorithm to generate a 3-dimensional map of the environment from the stereo image pairs obtained from the cameras. This map would then be transmitted, along with one of the stereo image pair, and on reception the stereo scene could be reconstructed from the data. The benefit of this type of system is that the viewpoint can be altered, allowing an observer to look at the scene from slightly different angles while retaining true perspective.

The primary problem with such a system is that the correlation algorithm is typically computationally expensive, with the faster algorithms often being quite inaccurate. It is simply not possible to do stereo correlation and video compression at a high enough frame rate.

5.6.2 Differencing

A simple stereo compression technique is to compress one of the stereo pair as a normal monoscopic image, and then compress the difference between the two stereo image pairs. Since the left and right images share much of the same information, there will often be large areas where the difference between the two images is small or zero, which means that the difference should compress well.

5.6.3 Sub-sampling

Using different resolutions for each of the stereo image pairs was attempted and, although the final result could still be stereoscopically fused, the low resolution image was clearly evident. It was decided that sub-sampling one of the stereo image pair does not result in further compression with minimal impact on image quality. Hence, further use of this technique was abandoned.

5.6.4 Shared Chrominance

Since the human eye is more sensitive to brightness than colour, experiments were performed with left and right images sharing the colour information. What this means is that if we use $\mathrm{YC_RC_B}$ 4:2:0 encoding for the image, then we can compress the a stereo image pair by compressing one of the images as a monoscopic colour image, and then compress the other image as a monochrome image. Upon reconstruction, the colour components are shared between both stereo image pairs, meaning that two colour images can be reconstructed.

Some tests showed that the stereoscopically fused image was still of very high quality, and that the fact that one image used the colour information of the fellow image was not perceivable.

5.7 Computational Speed

Since the final compression scheme that is settled upon must be run in real-time and cannot use any specialised hardware devices, the computational speed of the algorithm is an important factor. For the final decision there are several factors to be considered, many of which are detrimental to each other. The main aims of the compression scheme is to:

- minimise the bandwidth required to transmit,
- maximise the image and stereo quality,
- minimise the latency, and

- maximise the number of frames per second.

As can be easily seen, several of these requirements interfere with each other, for example, increasing the image quality will in general increase the bandwidth required for the transmission of the data. In this section we focus on computational cost, as this is critical for both the number of frames per second, and the latency of the transmission system.

5.7.1 JPEG

The primary computational cost in the JPEG compression standard is the forward and inverse Discrete Cosine Transform. Subsequent quantization, entropy coding and colour space conversion are quite fast by comparison. Using the Independent JPEG Group's software, it is possible to compress monoscopic images at quite high frame rates. For example, 25 frames per second was achieved for monoscopic colour images at a resolution of 384 by 288 pixels.

5.7.2 MPEG 2

MPEG 2 is essentially a more generic version of MPEG 1 with certain enhancements. It can compress video at a wide variety of resolutions extremely well, with many different quality levels. In early testing, an MPEG 2 encoder was speed tested on a Pentium III 600MHz computer, and an average speed of 4.43 frames per second was achieved for the encoding of monoscopic colour video with a resolution of 352 by 240 pixels. The main computational cost of MPEG over JPEG is the creation of predicted and bidirectionally predicted frames requiring complex calculation of changes in the video. The predicted frames are used to compress the video further by compressing the changes, or lack thereof, in the scene between successive frames.

It was clear at an early stage in this thesis that compression of stereo video at a reasonable frame rate was not possible if a software encoder was used. It is intended in the near future that MPEG will have extensions to handle stereoscopic video, and as computer performance increases, MPEG will be the compression method chosen for streaming stereo video over the Internet.

5.7.3 Wavelet

It was hoped in this project that a wavelet compression scheme could be used, since it provides many image quality advantages over DCT based methods. However, the computational cost of the forward and inverse wavelet transform is quite high, which is primarily due to the fact that the transform is applied to the whole image, rather than to small blocks of the image. On a 256 by 256 monoscopic monochrome image, a frame rate of 5.7 frames per second was achieved using the standard pyramid algorithm. Using the newer lifting scheme, a frame rate of 19.6 was achieved for the same image.

It was felt that one of the primary concerns of the compression scheme was the speed performance of the algorithm. With the lifting scheme, a reasonable frame rate may be achieved for stereoscopic colour video. Regardless, the JPEG based scheme was still much faster and while the wavelet transform provides better quality images in many respects there is no great difference in compression performance.

5.8 Summary

- The final compression scheme that was decided upon was a JPEG style compression scheme, using shared chrominance to encode the stereo aspect of the video.
- The reason for this was primarily due to the fact that this method was the simplest and fastest, and it was believed that speed was going to be of greatest importance.

Chapter 6

Transmission

A brief description of the Internet Protocol (IP) is given along with the common protocols that rely upon it. A discussion of the options for the transmission of stereo video, is followed by the outline of the concept for stereo video transmission. The results from the implementation are given.

6.1 Overview

For the Robot Command Station project, the intended transmission medium is a radio-ethernet bridge between the mobile robot and a relay station that will be connected to the Internet by a high speed link. Radio-ethernet communication devices are now commonplace and many amateur and professional wide area networks have utilised this technology. The technology is based on the IEEE802.11 Standard[38] which is, in principle, the same as Ethernet that is based on the IEEE802.3 Standard[37], the fundamental difference is that Ethernet is transmitted on cable, whereas 802.11 is transmitted via free space propagation. This results in some distinct differences in terms of hardware behavior, or the physical layer, but the principles of transmission are the same for both standards.

IEEE802.11 in its simplest form is a point to point communication system, so that a transmitter and receiver is required on each of the two computers being connected. To link two computers a special peripheral card (either ISA, PCI or PCMCIA) is required in each. These peripheral cards are relatively cheap, and can cost as low as $150 each.

The performance of radio-ethernets is comparable to that of 10 Mb/s ethernet. There are several increments of performance available in the radio-ethernet range, however the common maximum is 11 Mb/s (claimed by the manufacturer). Such a radio Ethernet system was purchased by the Research School of Information Sciences and Engineering (RSISE) and subsequent testing revealed that approximately 6 Mb/s could be sustained at a range of 2 km. Hence, for this project, the absolute maximum transmission rate was assumed to be 6 Mb/s, with a preference to ensure that the actual transmission of video was substantially less than this to allow transmission of other data as well as the video.

Finally, access to the radio Ethernet devices was not available during the term of this project. It was deemed that simulation on the local network, a 10 Mb/s Ethernet network, was a suitable test of the transmission methods developed. In any case, all testing measured the actual transmission rates used and the results are independent of the underlying hardware used (This is of course not entirely correct, as things such as latency will differ with different hardware).

6.2 The Internet Protocol Suite

What is presented here is a very brief overview of the Internet Protocol that is necessary for an understanding of the material presented in this chapter. For a more thorough review of this protocol, see Stevens [66, 67]. It should be noted that the focus is entirely on version 4 of the Internet Protocol as this is the most commonly available. Most of the principles of the protocol are the same between versions, with the differences being primarily due to differences in addressing nodes on a network.

6.2.1 Network Layers

A common way to categorise the various functions necessary for the networking of computers is to use layers. The ISO Open Systems Interconnection (OSI) [22] model has seven layers which provide a hierarchical separation of the functions that a network must provide. Unfortunately, the Internet Protocol was developed before the creation of this standard model which means that there is not a direct mapping between each OSI layer and a functional group in the Internet Protocol suite.

The OSI layers, in order from lowest to highest, are Physical, Data Link, Network, Transport, Session, Presentation, Application. The following sections will give an explanation of each of the layers and discuss mappings between the layer and the corresponding functional group in the Internet Protocol suite.

6.2.1.1 Physical

The Physical layer defines the physical connection between nodes. This includes the type of connections and cabling, the definition of radio or electrical signals and their interpretation, as well as the procedural elements required for the transmission of a message. In general, this layer is defined by a particular standard for a network such as IEEE 802.3 or IEEE 802.11, although there are many other standards.

6.2.1.2 Data Link

The Data Link layer provides the lowest level software interface to the network. This is most commonly in the form of a device driver for a network adaptor card. The Data Link layer is responsible for the transmission and reception of blocks of data to and from the network.

The Internet Protocol does not have a corresponding functional group for this particular layer.

6.2.1.3 Network

The Network layer is responsible for the addressing of nodes on the network, the end to end transmission of data, and the routing, switching and relaying of information across distant networks. The Internet Protocol is responsible for this layer in the Internet Protocol suite. The Internet Protocol gives each computer a 32 bit address, which is usually written in number dot number form, e.g. `150.203.205.12`.

Protocols such as Address Resolution Protocol (ARP) and Reverse Address Resolution Protocol (RARP) are used as a bridge between the Data Link layer and the Network Layer. The Internet Control Message Protocol (ICMP) and the Internet Group Management Protocol (IGMP) are also classed in the Network layer as these provide control messages that are used between routers and hosts.

6.2.1.4 Transport

The Transport layer provides the ability to transmit data reliably between two communication systems. There are two elementary protocols in the Internet Protocol Suite that can be considered to be in the Transport layer: the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP). There are several other transport protocols but these are supplementary and certainly not as common as TCP and UDP. TCP provides a reliable connection based service while UDP provides an connection-less service that does not guarantee delivery. Both of these protocols will be described in the following sections.

6.2.1.5 Session, Presentation, and Application

The Session, Presentation and Application layers are application specific. As most of the applications that have been developed on the Internet predate the OSI model, they naturally have developed their own particular method of implementing their respective application protocols. Hence the Internet Protocol suite provides only the Network and Transport layers, and Internet applications such as the World Wide Web and e-mail use these directly.

6.2.2 The Sockets Interface

The Sockets interface was first released in 1983 with the 4.2BSD (Berkely Software Distribution) Unix system. Due to the popularity and wide-spread use of the BSD system, the Sockets Application Programming Interface (API) became the defacto standard for interfacing with the Internet Protocol suite.

Work on standardising the Sockets API began in the late 1980s as part of the Portable Operating System Interface (POSIX) set of standards. Recently, the Sockets API has been standardised as an IEEE standard, under the heading of Protocol Independent Interfaces [39]. This standard defines two independent APIs:

- Sockets — based on the 4.4BSD Sockets API
- XTI — based on the X/Open Portability Guide, Issue 4.

The Sockets API is the most prominent standard that is used for writing software for communicating on Internet based networks. For Microsoft operating systems, the Sockets API has been adopted, albeit with some subtle differences.

It should also be noted that the primary source of standards on the Internet is the Internet Engineering Task Force (IETF). This group is a large community of people concerned with the Internet who propose portable methods for functioning on the Internet using informative RFCs (Request For Comment). All of the Internet Protocol suite is defined in RFC form, and is the most accessible form of standards available relevant to the Internet.

6.2.3 The User Datagram Protocol

The User Datagram Protocol (UDP) is the simplest Transport layer protocol in the Internet Protocol suite and it is described in RFC768 [59]. UDP simply transmits a datagram [1] that is encapsulated in an Internet Protocol packet. UDP is no more than a specification of the header information contained at the beginning of each UDP packet.

When data is sent using UDP, it must be packaged into datagrams in a particular fashion. Most UDP implementations provide very little in the way of buffering, which means that it is quite easy to lose many packets by transmitting packets too fast for the receiver to match. This means that the latency of UDP transmission is quite low, that is UDP does not incur much overhead on top of the underlying Internet Protocol network layer.

It is possible on the Internet for sequentially transmitted packets to arrive at the receiver in a different order to which they were sent. This occurs due to changes in the route taken by packets since routing on the Internet is updated dynamically according to usage and bandwidth.

The UDP does not guarantee delivery of the information to the intended recipient. If a particular piece of information must be transmitted reliably, then an application will have to add reliability to UDP. In these cases it is often better to simply use the Transmission Control Protocol (TCP).

6.2.4 The Transmission Control Protocol

The Transmission Control Protocol (TCP), described in RFC793 [60], provides many of the functions that are missing in UDP. In short, these functions are reliability, sequencing and

[1] A datagram is a contiguous sequence of bytes

flow control, all of which are important for providing a consistent transport layer for transmission of data over the Internet.

The first difference, from a software perspective, between TCP and UDP is that TCP is a connection based protocol, whereas UDP is connection-less. Prior to transmitting data to remote host, a connection must be made. The connection is made using an exchange commonly refered to as a three-way handshake. A connecting client sends a SYN (synchronise) segment to the remote host, the remote host will respond to this message with its own SYN segment and an acknowledgment of the received message. Finally, the client sends an acknowledge back to the remote host to acknowledge the received message.

The SYN segments are used to make both ends of the connection aware of the current value of the sequence numbers used in TCP packets to preserve the order of packets in transmission, so each end of the connection maintains a sequence number that corresponds to the expected sequence number of the next packet to be received. This is important to the TCP protocol since large blocks of data that are transmitted are automatically split into segments before being transmitted, with the maximum segment size (MSS) usually negotiated during the three-way handshake. Associated with each transmitted segment is a sequence number, so that if segments arrive out of order at the receiver, they can be reordered before being processed by the application.

The connection is terminated in a similar way to which it is initiated and either of the parties that are connected to each other may terminate the connection. A host signals that it wishes to end the connection by sending a FIN segment. The receiver acknowledges this and sends its own FIN segment. Finally, the host initiating the termination of the connection acknowledges the FIN from the other host.

Flow control is provided in the TCP protocol through the communication between connected hosts of the amount of data that each is able to receive. In implementation terms, each connection has a sending and receiving buffer, whose size is fixed. This changes dynamically as data is received and subsequently processed by an application. It is possible for a host to be in the state where it cannot receive any more data from the remote host, that is, its receive buffer is full. This facility prevents a fast host from transmitting data at too high a rate for a remote host to process. The size of the send and receive buffers can be set prior to connection, with the minimum and maximum buffer sizes dependent on the operating system implementation.

TCP also provides reliability in the use of acknowledgments, time outs and retransmissions. When a packet is sent to a remote host, an acknowledgement is expected within a certain period of time. If an acknowledgment is not received, then the packet is resent and the length of time to wait for an acknowledgment is increased slightly. After a certain amount of transmission attempts, it is assumed that there is some form of network error and transmission of the data will be aborted. So, with the exception of terminal situations, TCP will guarantee the reception of all data that is sent.

It should be noted that the software interface provided by the Sockets API hides much of the detail of the TCP from the user. For example, when opening and closing connections, one is never aware of the SYN and FIN segments.

6.3 Preliminary Experiments

6.3.1 Transmission of JPEG using the Transmission Control Protocol

One of the earliest experiments performed was the transmission of stereo video using a simple technique: analog frame grabbers were used to capture an image at a certain size from the cameras. This size was a defined constant in the software and therefore could not be changed without recompilation. The two stereo images were combined into a single stereo image of the same dimensions by halving the vertical resolution. Therefore, the transmitted image had the left image in the top half of the image, and the right image in the bottom half. This combined image was then compressed with the Independent JPEG Group's JPEG compression software [34], with the quality of the image also defined by a constant in the software.

Two programs were written, one to capture and transmit the stereo video, and the other to receive and display the video. The program that transmits the video, the server, waits for a connection to be received. When a connection is received, the server enters a tight loop in which stereo video frames are continuously sent. Hence, the client simply connects to the server and continually receives the video frames and displays them in turn. All data, as can be implied from the necessity for a connection, is transmitted using TCP and, as such, it was necessary to specify in the transmitted data how big each stereo video frame was. Therefore, preceding each stereo video frame was a 32 bit number which gave the length in bytes of the stereo video frame.

As can be seen, this is a trivial method for the transmission of stereo video, however there were several problems discovered. Firstly, given the trivial nature of the protocol between the client and the server, in order to change a parameter such as the width, height, or quality of the image, both the client and server software had to be modified and recompiled. This became noticeably tedious during testing and so it was determined that specification of various image parameters should be included in the final protocol between client and server.

The entire compressed frame varied in size between 5,000 and 30,000 bytes for a 384 by 288 image, depending on the quality selected. The test network used for the transmission of data was ethernet and which had a hardware imposed maximum transmission unit (MTU) of 1,500 bytes. This meant that fragmentation was required for the transmission of each frame, with a minimum of four fragments required. Such fragmentation increases the latency, since each fragment must be buffered before reassembly of the entire packet can occur.

This system performs quite well on computers that are closely networked. In such a situation, the frame rate was approximately 15 frames per second with a latency of around 0.8 s. Transmission to more remote sites resulted in latencies approaching 5 s, with a substantially reduced frame rate. The reason for this sudden increase in latency is that TCP requires an acknowledgment for each packet that is transmitted. So, as the distance to the remote receiver increases, the latency increases markedly due to the fact that both the transmission of the video and the resulting acknowledgment are affected.

Another problem with this method is that half of the vertical resolution is lost due to the fact that two stereo images are combined into a single image. This can be easily remedied by simply fabricating an image twice as high or wide as the individual image, in the same way that the defacto JPS stereo image file standard stores the data. However, this has the disadvantage of increasing the amount of bandwidth required for the transmission of the stereo video. Also, as this requires greater computational effort for the compression, the frame rate will be somewhat reduced.

6.3.2 Transmission of Red-Blue Anaglyph using Web Pages

Another simple system that was tested was to use an existing piece of software which enabled video to be viewed using a common web browser such as Netscape Navigator or Microsoft Internet Explorer. The software used was called `camserv` [73] and it enabled the creation of a streaming webcam server. It provided the ability to capture images from an analog frame grabber and stream it to a remote client using JPEG compression to reduce the data rate. All options such as image size, quality, and frame rate, can be set in a configuration file.

`Camserv` provided no means for the transmission of stereo video. As a result, this software was modified so that, instead of capturing a single frame from a single analog frame grabber, a single image is created from two left and right frames captured from two analog frame grabbers. The two monochrome images were mixed into the blue and red channels of a colour image to achieve a red-blue anaglyph stereo image which could be viewed by an observer with appropriate glasses on most web browsers.

Initial tests on the local network showed that the performance was quite good. For a 384 by 288 pixel image, the latency was very small between closely located computers, ranging from about 200 ms to 1 s depending on network usage. For a 768 by 576 pixel image, the latency increased significantly to between 2 and 2.5 s.

The primary reason for the development of this system was the desire to transmit stereo video from the Australian National University to a site in Berkeley, California, USA. At such distance, using a specialised system that required particular software at both ends was not practical so it was necessary to rely on existing technology at the receiving end. Therefore, the modification of an existing webcam streaming server seemed to be the simplest and most effective choice. There were, however, some notable problems with this system.

Firstly, as stated in Chapter 5, compression of red-blue anaglyph images with normal lossy image compression techniques such as JPEG is sub-optimal due to the sub-sampling of colour information in which the stereo information is stored. Secondly, the World Wide Web uses TCP for all transmission of information between a server and a client. Therefore, the same problems that were encountered with using TCP in the previous experiments were evident. Of particular concern was the latency of the transmission. For transmission to many different computers in California, the latency was of the order of 5 — 10 seconds. Subsequent tracing of the route to the remote computers revealed that the round-trip-time (RTT) to these computers ranged between 300 and 600 ms. This shows that TCP, by itself, is not suitable for the transmission of large volumes of real-time data over long distances and through several different types of networks [2].

There is the potential to alleviate the latency with the use of the "long fat pipe" extensions to TCP, as described in RFC1323 [47]. However, this is a relatively new feature and not supported by all implementations. As the feature needs to be used by the web browser, parts of the web browser software would need to be rewritten which is not feasible. While using this option alleviates the latency introduced by TCP, it does not remedy it. The problem will still be present and will be most evident when packets are lost as TCP will have a large delay between the timeout and retransmission of the lost data. This will, in turn, result in a pause in the video, the length of the pause being proportional to the round-trip-time. Therefore, TCP should not be used for the transmission of real-time video over long distances or on networks prone to the occasional loss of information, for example, satellite and radio transmission).

6.3.3 Transmission of DCT compressed video using the User Datagram Protocol

Given that TCP could not be used for the transmission of stereo video in real-time, a prototype system was developed using the User Datagram Protocol. The use of this protocol remedies many of the problems associated with the use of TCP due to the fact that there is no reliability or flow control, both of which can be detrimental to the latency of the transmission. These two facilities of TCP are provided for a reason, and, early on there were several problems that were discovered by using UDP.

Firstly, since UDP has no form of flow control it is possible to overflow a receiver with data by transmitting data at too high a speed. When this occurs, a large percentage of the data is lost. Packets that cannot be processed are simply discarded. Initial testing of the developed system highlighted this problem and subsequently a small test program was written to test the severity of this form of packet loss. The software consisted of a server which simply counted the number of UDP packets that were received, and a client which transmitted UDP packets in a tight loop as fast as possible. The client was configured to transmit 10,000 UDP packets, each of which was 512 bytes in size, not including the overhead of the Internet Protocol and User Datagram Protocol headers. Several trials were run in order to reduce the effect of network traffic on the result. For the transmission of 10,000 packets, an average of 280 packets were received, which means that approximately 97% of the transmitted packets were lost at some point. Such a loss rate in the transmission of digital video would result in severely degraded images. Insertion of a small delay between the transmission of the UDP packets resulted in a substantial improvement in the percentage of packets lost. With a small delay, an average of 9,190 packets were received of the 10,000 transmitted, which equates to a packet loss of 8%.

As UDP is unreliable, each packet must be able to be decoded in isolation to ensure that

[2] Tracing showed that the route to California traversed ATM, ADSL, Gigabit Ethernet and various other serial network interfaces

the loss of a single packet does not result in other correctly received packets being wasted. Each individual packet must therefore contain enough information in it to allow complete decoding of the data it contains. The most logical method of doing this is to partition each frame of video into smaller pieces that can be encoded and transmitted separately. Since UDP is datagram based, we must specify a particular size for each data block or partition of the image. If packets are lost or discarded in the transmission of video, this is not a severe problem if such loss is random and each packet can be individually decoded. This is because with video, up to 30 frames a second are transmitted and if a packet is lost then there will only be a period of about 33 ms before the next frame will cover this loss. This assumes that there is a sufficiently high frame rate and that packet loss occurs randomly, that is, evenly distributed over the entire frame area.

Another aspect of the unreliability of UDP is that packets can, because of dynamic routing, be received out of order. Therefore, some scheme needs to be included in the protocol to insure that the order in which the particular frames are displayed is as intended.

With all the specialised requirements for the transmission of video using UDP, it was determined that existing image compression libraries, such as that written by the Independent JPEG Group [34], did not provide the functionality required. For this reason, a simple DCT based system was developed that could compress monochrome images in stereo to varying degrees. This particular compression system split an image into 16 by 16 pixel blocks which were compressed individually using 4 DCT transforms. Each of these blocks was given a number to specify which part of the frame they belong to, where the number of each block increases from left to right and top to bottom.

Initially, each of these 16 by 16 blocks was transmitted in a single UDP packet. This is inefficient since each UDP packet has a transmission overhead of a minimum of 20 bytes of IP header and 8 bytes of UDP header. For a highly compressed 16 by 16 block of data (i.e. low quality), it is possible for less than 28 bytes to be used. So it is possible to be in the situation that the protocol overhead is larger than the data in the packet. For this reason, several contiguous 16 by 16 blocks were transmitted in a single packet whose size was dictated by the path Maximum Transmission Unit (MTU) of the transmission medium. Thus each packet would be filled to the maximum size that can be transmitted without fragmentation. This had the effect of minimising the effect of the UDP/IP overheads on the data transmission rate.

There are some quite obvious trade-offs to be found with this algorithm. If each 16 by 16 block is transmitted in a single packet then data is used inefficiently. Conversely, if several 16 by 16 blocks are transmitted in a single packet, then the loss of a single packet causes greater detriment to the video quality at the receiver. However, of greater concern is the bandwidth usage, which in itself can cause more packets to be lost. For example, if each 16 by 16 block is sent separately then for transmission of a monoscopic video stream with a resolution of 384 by 288 pixels at 25 frames per second, this requires the transmission of 10,800 packets per second which will result in a UDP/IP overhead of nearly 2.5 Mb/s which is excessively high. Considering that a desirable bit-rate target is 1.0 Mb/s, such inefficiency cannot be allowed, and therefore a degree of packet loss tolerance must be sacrificed.

For testing on the local network, a path MTU of 1500 bytes was specified. When streaming the 16 by 16 blocks, each block was added to a packet buffer until the packet was too large. At this point, the packet was transmitted with its collection of blocks, and the final block would be the first in the next packet. This process optimises the use of bandwidth by minimising the overhead of the UDP/IP headers. Any such buffering of the data adds to the overall latency of the system as there is a small amount of time required to fill the buffer prior to sending. Unfortunately, this is a factor that cannot be remedied.

The quantization method used in the compression scheme was a very simple one that divided the coefficients by a power of two. This effectively reduced the number of bits needed to represent each of the DCT coefficients to a whole number of bits. Initial testing of the compression scheme showed that a minimum of 4—5 bits were necessary for the representation of coefficients to produce reasonable quality images upon reconstruction. The difference in quality between images compressed using this compression scheme is shown in Figure 6.1.

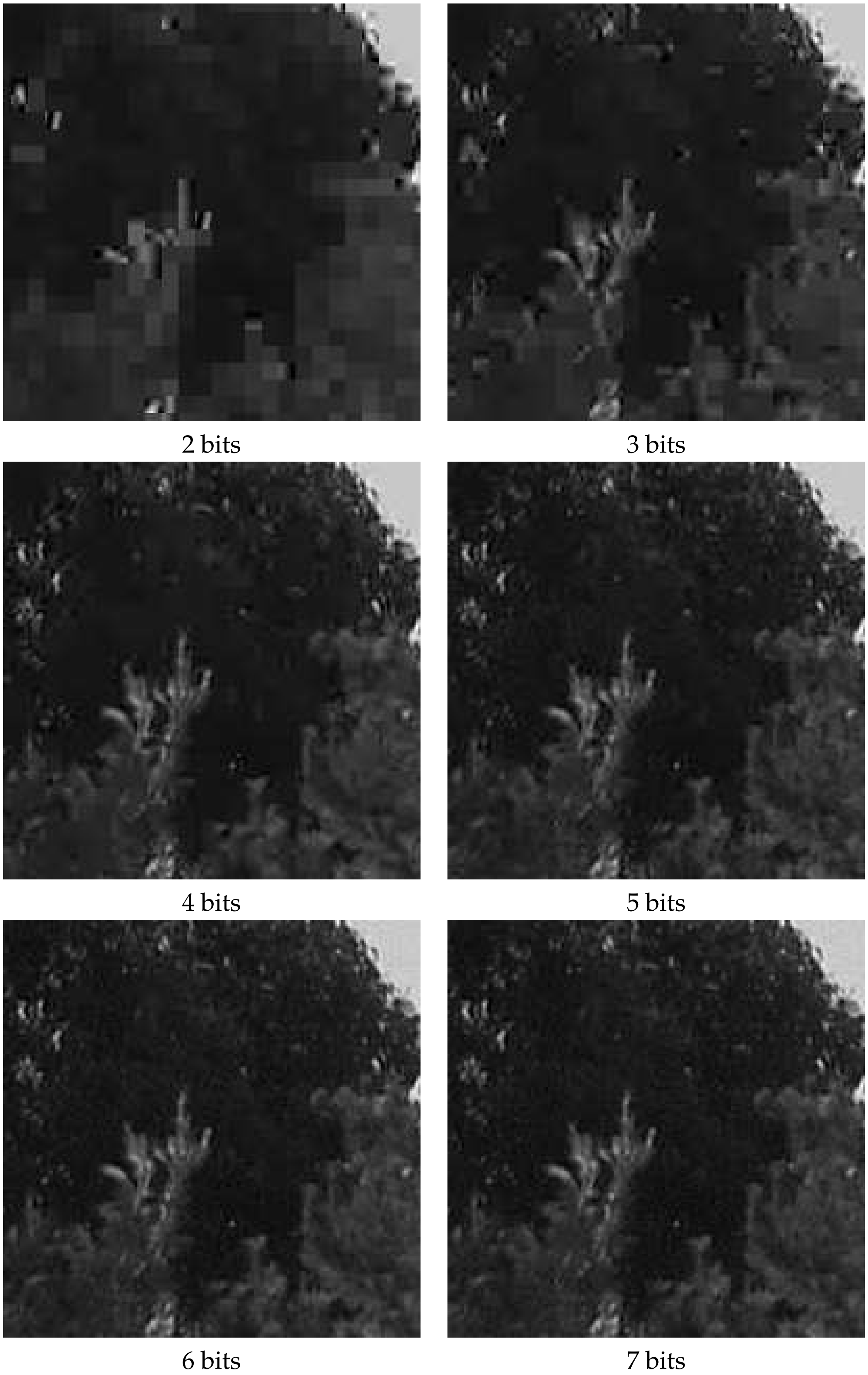
As this system purely used UDP, there was no way of reliably conveying various parameters between the sender and the receiver. The standard setting was for an image size of 384 by 288 pixels, with 32 quantization levels or 5 bits per coefficient. If it was necessary to changed the image size or the number of quantization levels from the default values, then the client and server had to be restarted with the new parameters. The requirement for restarting the software at the smallest change in one of the system parameters was a real inconvenience and one of the greatest drawbacks of this simple test system.

The results obtained from this simple system were reasonably good, in that an acceptable frame rate was achieved and the quality of the stereo was quite good. Plots of some performance measured as a function of image size are shown in Figure 6.2. Of note is the fact that with an image size of 384 by 288 pixels, transmission was achieved at approximately 10 frames per second utilising less than 2 Mb/s. For such a simple experiment, this system came close to the requirements for the project.

There were, however, several problems with this software. Firstly, the lack of any reliable form of communication between the sender and the receiver is a big drawback. For this reason, there should be both a UDP channel and a channel that guarantees delivery (e.g. TCP) between the sender and receiver to provide reliable transmission means. This also allows parameters to be specified on startup or dynamically and also has the added benefit of signaling when one end of the transmission has been stopped or has crashed.

Next, the compression scheme used in this experiment used a simple quantization scheme that does not take into account psycho-visual factors. Namely, visually weighted quantization, as is used in the JPEG compression scheme, was not used. The advantage of this method can be seen from the highly visible edge artifacts visible in Figure 6.1. These are a direct result of using the simple quantization scheme, and such artifacts only appear in JPEG compressed images at much lower quality settings. Another more obvious problem is the fact that the image compression scheme could only cope with monochrome images.

Lastly, the combination of the computationally expensive task of compressing the images and the transmission of the data, which typically requires a lot of waiting, lends itself well



2 bits

3 bits

4 bits

5 bits

6 bits

7 bits

Figure 6.1: Quality of simple compression scheme for varying quantization levels

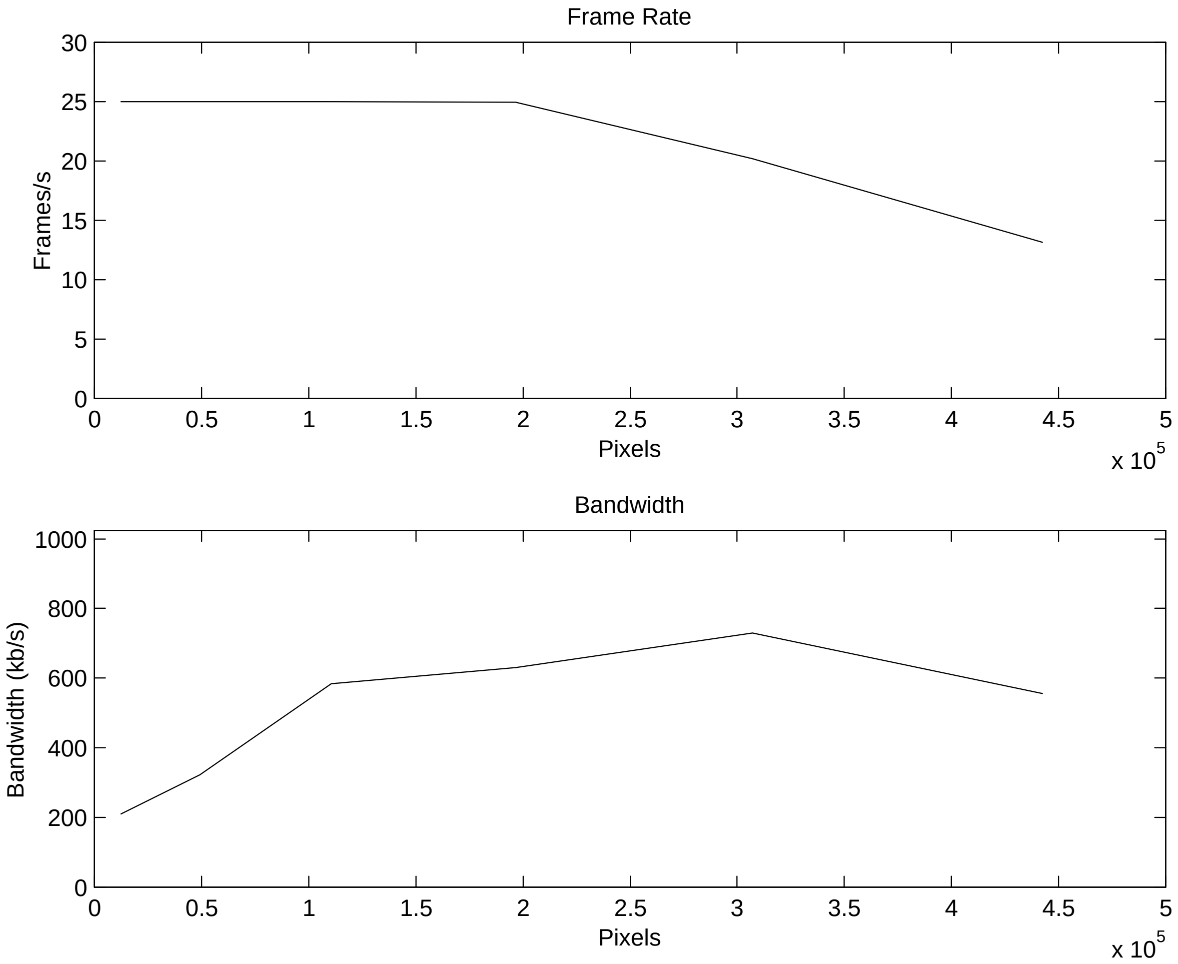


Figure 6.2: Performance of UDP based transmission

to a multi-threaded system.

6.4 Combined TCP/UDP Transmission

Given the compression method that was decided upon was a modified form of the M-JPEG defacto standard, this provided some final restrictions on the method of transmission. Of most importance is the fact that the quality of the JPEG image can be varied to obtain differing compression ratios and, therefore, the number of bytes per frame will vary. This impacts on the transmission since it is desirable to avoid creating a packet larger than the path MTU. Therefore the transmission method must allow for a highly variable number of bytes per frame while still maintaining packet sizes approaching the path MTU.

These problems have already been addressed to some extent by the Real-time Transport Protocol (RTP) [63], the Real-time Transport Control Protocol (RTCP) [63] and the RTP payload format for JPEG compressed video standard [8]. The RTP suite of standards specifies many different features, such as multicasting and session control. It was deemed that implementing the standard, but using the modified M-JPEG compression method, was unnecessary. Instead, it was decided that a system that used TCP and UDP and emulated several of the features of RTP would be implemented.

The JPEG standard provides a method of splitting up the image into smaller image sections. This method consists of inserting restart markers into the JPEG bit stream. Each section of a JPEG bitstream is delineated by a marker, in which the first byte is `0xFF` and the second byte specifies the type of marker. The marker `0x00` is reserved for bit sequences where the byte `0xFF` naturally occurs. The restart marker indicates that the various encoding/decoding parameters should be reset at this point which, in turn, means that encoding/decoding can begin independently of all other sections of the image. This particular mechanism lends itself well to transmission via packets as the restart interval can be adjusted so that each section of the image can fit into a single packet.

One of the first problems discovered using UDP was the prevalence of congestion. The sender would frequently send packets faster than the receiver could receive and decode them. This resulted in quite severe packet loss and therefore degraded the quality of the video quite substantially. Hence, it was obvious that some form of bandwidth throttling was necessary to prevent a sender from transmitting at too great a rate. For this reason, it must be possible for a user to specify the maximum amount of bytes per second that can be transmitted over the network.

As well as being able to specify the maximum amount of bandwidth that can be used for the transmission of video, it must also be possible to specify many of the characteristics of the video being transmitted. For example, things that the user may wish to specify are:

- image width and height,
- quality factor,
- restart interval,

- maximum frame rate, and
- compression scheme to use.

The particular scenario that is being envisaged here is that an operator will specify their particular settings at the receiving end of the transmission, and then inform the sender to start transmitting the video with the particular parameters. Since some form of reliable communication is necessary between the sender and the receiver to perform these tasks, a joint TCP/UDP protocol is necessary for the transmission of the stereo video.

TCP must be used for messages and the like, when transmission to the other participant in the transmission must be guaranteed. For example, any video parameter must be sent using TCP as both ends of the transmission need to ensure that they have the same set of parameters as each other. Similarly, messages from the receiver to start or stop the transmission of video need to be sent using TCP as these messages are critical for the transmission of the video. As the video requires low latency and is less susceptible to loss of packets, it will still be transmitted using UDP.

6.5 Design

The following sections give an overview of how many of the features of the stereo video streaming software were implemented. For a more detailed description with samples of the source code, see Appendix D.

6.5.1 Compression Scheme

Since it was desired to experiment with a variety of different compression schemes, all of which used M-JPEG compression as the underlying method, some flexible software was necessary for the compression of the stereo video. Early experiments with the Independent JPEG groups software library showed that this software was quite fast. However, one of the things that was noted was that an entire image had to be compressed before the bit-stream was available for transmission. Also, the bit-stream that was produced was in the form of a JPEG File Interchange Format (JFIF) file, since this library's purpose was for the compression of images to and from files.

Since we have seen that the entire compressed bit-stream cannot be transmitted in a single packet, compressing an entire image all at once has the effect of increasing the latency of the transmission. This can be best illustrated with the use of an example. Consider the scenario where it is desired to transmit video at a resolution of 384 by 288 pixels at 25 frames per second using an M-JPEG compression scheme at a quality giving a compression of 1 bit per pixel. This means that the transmission rate is 2,764,800 bits per second, which means that a single frame will take 40 ms to transmit. If we assume the compression of a frame of this video takes 40 ms and that the compression time is linear with respect the number of pixels process, i.e. if we compress half of a frame it will take 20 ms [3]. Now, a single frame will consist of approximately 14,000 bytes which will require 10 UDP packets for transmission to the receiver if the path MTU is 1500 bytes, i.e. on an Ethernet network. If it takes 10 ms for a single UDP packet to reach a receiver, then we can work out the difference in latency between compressing the entire image and then sending or compressing each packet as required.

If we compress the entire image and then send it, then the time from when the image is obtained to when it is reconstructed at the receiver is:

$$\begin{aligned} \text{latency} &= 40 + 40 + 10 + 40 \\ &= 130\text{ms} \end{aligned}$$

On the other hand, if we compress one packet of data at a time and send, then we effectively multiplex the process of compressing and transmitting the video, and therefore reduce the latency. Since ten packets are required to transmit the video, compressing a packet of video will take 4 ms. Once this is compressed it can be sent and processing of the next packet can begin immediately. Similarly, if the receiver also multiplexes the receiving and reconstruction of the images, then time is also saved. In this scenario, if both ends of the transmission can maintain the multiplexing of transmission and compression, then the latency can be reduced to:

$$\begin{aligned} \text{latency} &= 4 + 40 + 10 + 4 \\ &= 58\text{ms} \end{aligned}$$

The differences between these two compression and transmission methods is shown diagrammatically in Figure 6.3. By multiplexing the compression with the transmission, we reduce the effect of the computational cost of the compression algorithm on the latency of the system. This also has the added benefit of making the transmission more consistent, that is, data is transmitted at a consistent rate, rather than bursts of data being transmitted after each frame is compressed.

The other problem with the Independent JPEG Group's software is that it is designed to create bit-streams that are to be written to files. For this reason, the bit-stream contains additional information such as the image attributes, type of compression, quality, and Huffman encoding parameters [4], i.e. a percentage of the bits in the file are, to some extent, redundant, and therefore the transmission of files of this type is inefficient.

It is clear to see that although the Independent JPEG Group's software provides a very good library for the compression of images using the JPEG standard, it is not suitable for the real-time transmission of video using the defacto M-JPEG standard. It therefore became necessary to implement a JPEG compression software library that was more specialised to the task of transmitting video over the Internet.

[3] Although the DCT is an $O(n^4)$ operation, JPEG style compression can be considered linear since the DCT is only ever applied on an 8 by 8 pixel block

[4] The Independent JPEG Group's software does not support Arithmetic Encoding

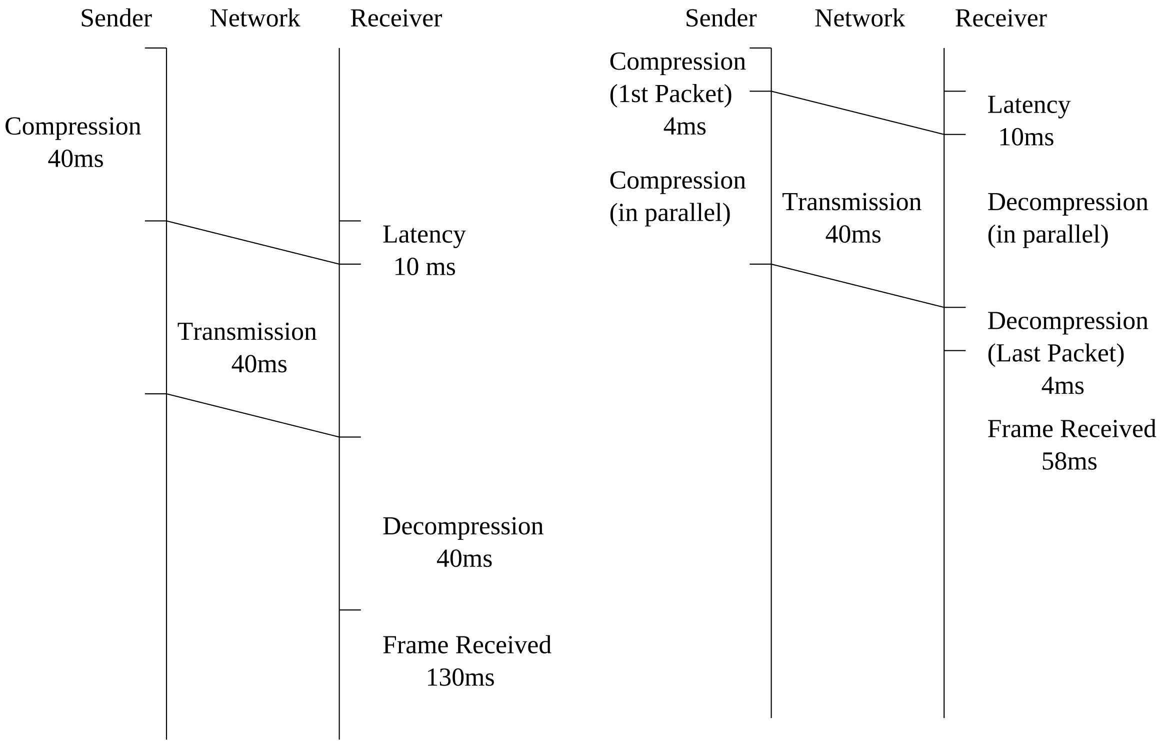


Figure 6.3: Comparison of latency using two different compression-transmission schemes

What the new software library had to do was to compress part of a JPEG image independently of all other parts of the image. Also, the compression and decompression was to use a set of standard Huffman tables in order to eliminate the transmission of this information between the sender and the receiver. This sacrifices some compression performance, as the Huffman tables cannot be optimised for each particular set of images, but it does save the transmission of the Huffman encoding parameters between sender and receiver which saves on bandwidth and a little on the computational cost. Also of importance is the ability to specify the restart interval for the compression and decompression. This is important as increasing the restart interval increases the compression performance slightly while increasing the size of the minimum number of bytes that can be transmitted independently.

6.5.2 Protocol

It was necessary to define a protocol for communication between the sender and the receiver to define the characteristics of the stereo video that was to be transmitted. The necessary information that needs to be transmitted is the:

- compression method,
- image height and width,
- quality,
- restart interval, and
- the maximum amount of bandwidth to use.

As well as the video parameters, the sender and receiver also need to communicate when to start and stop the transmission as well as which UDP port to transmit/receive on.

For this, a simple messaging system was designed which consisted of a 32 bit message that is transmitted between the sender and receiver. This 32 bit message consists of a 16 bit code to represent the type of message, and the remaining 16 bits for an operand if necessary. Even though these messages are transmitted using TCP, each sent message is acknowledged to ensure that the command code and parameter (if required) are valid. The choice was made to use this simple method because of the relative ease of both implementation and of future extension.

6.5.3 Bandwidth Throttling

In order to limit the usage of bandwidth a buffer was required that would store the packets that needed to be sent until the number of bytes written per second was at the desired rate. For this to occur, a separate thread of execution was necessary that sent a packet of the video stream, accumulated the total number of bytes sent, and the total time elapsed, and used these two pieces of information to determine at what time the next packet of video could be sent without exceeding the desired bandwidth limit.

Hence this thread of execution was also most suitable for the task of packaging the contiguous restart intervals into a packet prior to sending. Hence the software responsible for the sending of the stereo video over the Internet can be thought of as containing two distinct components. Firstly, there is the stereo video bit-stream creator which captures the stereo images from the analog frame grabbers and compresses each frame. Then there is the stereo video streaming component which accepts each restart interval from the bit-stream creator and uses this to make up part or the whole of a packet. The coupling between these two components is a finite buffer which can store a number of restart intervals of the compressed stereo video bit-stream.

Using this method, it is possible to control the amount of bandwidth that is being used while still ensuring that each frame that is processed is transmitted in its entirety. This occurs because the buffer that couples the two components of the stereo video streaming software blocks when it becomes full. It can only become full due to the action of the bandwidth throttling. When the sending buffer blocks, whole frames are dropped because the analog frame grabbers will obtain the next frame in the sequence since their input is of a fixed frequency. This fact is important because if the process of bandwidth throttling interferes with part of the frame being transmitted, then portions of the video may not be updated for long periods of time. If we ensure that each frame is entirely transmitted, then the overall frame rate may be reduced, but we greatly improve the chances of entire frames being displayed at the receiver, which then increases the quality of the stereo video.

For testing purposes, two assumptions were made as these particular parameters are difficult to discover dynamically. Firstly, it was assumed that the combination of the IP and UDP header amounted to 28 bytes (20 bytes for the IP header and 8 bytes for the UDP header). This is likely to be correct most of the time, the only time it will differ is when optional aspects of the Internet Protocol are invoked, which increase the size of the IP header. Secondly, for all testing it was assumed that the path MTU was 1500 bytes. All testing was done on an Ethernet network, so this fact was correct. However, a more correct test would have involved dynamically determining the actual path MTU in software. This can be done but involves some tedious software to test the path with different length packets until the path MTU is discovered, hence the path MTU was simply assumed to be that of standard Ethernet networks.

6.5.4 Supported Compression Methods

In the end, only an M-JPEG style compression scheme was used which had several different variants. The different variants that were implemented are:

- Monoscopic Monochrome Video
- Monoscopic Colour Video using $\mathrm{YC_BC_R}$ 4:2:0 encoding
- Monoscopic Colour Video using $\mathrm{YC_BC_R}$ 4:2:2 encoding
- Stereoscopic Monochrome Video

- Stereoscopic Colour Video using YC_BC_R 4:2:0 encoding
- Stereoscopic Colour Video using YC_BC_R 4:2:2 encoding

Of these, the variants that used YC_BC_R 4:2:2 encoding for colour were of little use as these had very little in terms of performance gain over the YC_BC_R 4:2:0 encoded variants.

6.6 Results

6.6.1 Performance

One of the desirable qualities for the transmission of the stereo video was the number of frames per second. It was stated that as a minimum, 10 to 15 frames per second were required. Figure 6.4 shows the frames per second achieved for different bandwidth limits, and the different compression methods.

6.6.2 Bandwidth

Since the target bandwidth was approximately 6 Mb/s, the bandwidth used by the various compression methods with varying quality was also tested. Figure 6.5 shows the bandwidth used for various qualities and different bandwidth limits.

6.6.3 Packet Loss

One of the requirements of the streaming software is to ensure that as many packets as possible are transmitted correctly. Figure 6.6 shows how the percentage of packets lost increases with increasing bandwidth usage.

6.7 Summary

In the transmission of data using the Internet Protocols, there are a few points to keep in mind when the data has a real-time component:

- TCP is not suitable for the transmission of real-time data over long (network) distances.
- UDP is not suitable for the transmission of data that cannot afford any loss of information.
- UDP is suitable for real-time data transmission due to the fact that it has very little computational or protocol overhead.
- Transmission of high bit rate data, such as video, should be made as consistent as possible (i.e. burst of transmission should be avoided), as this will reduce the chance of loosing data.

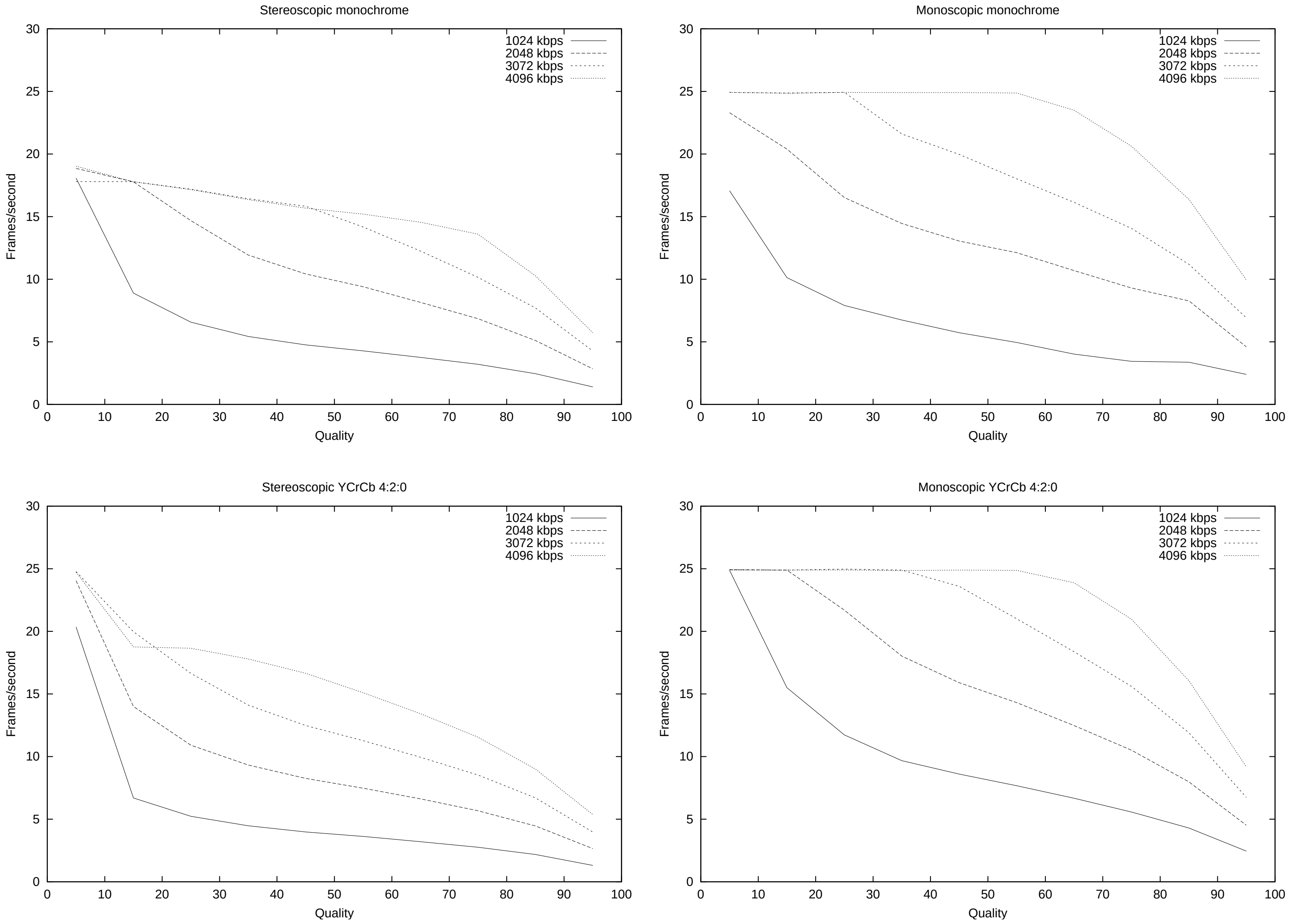


Figure 6.4: Transmitted frames per second with varying parameters

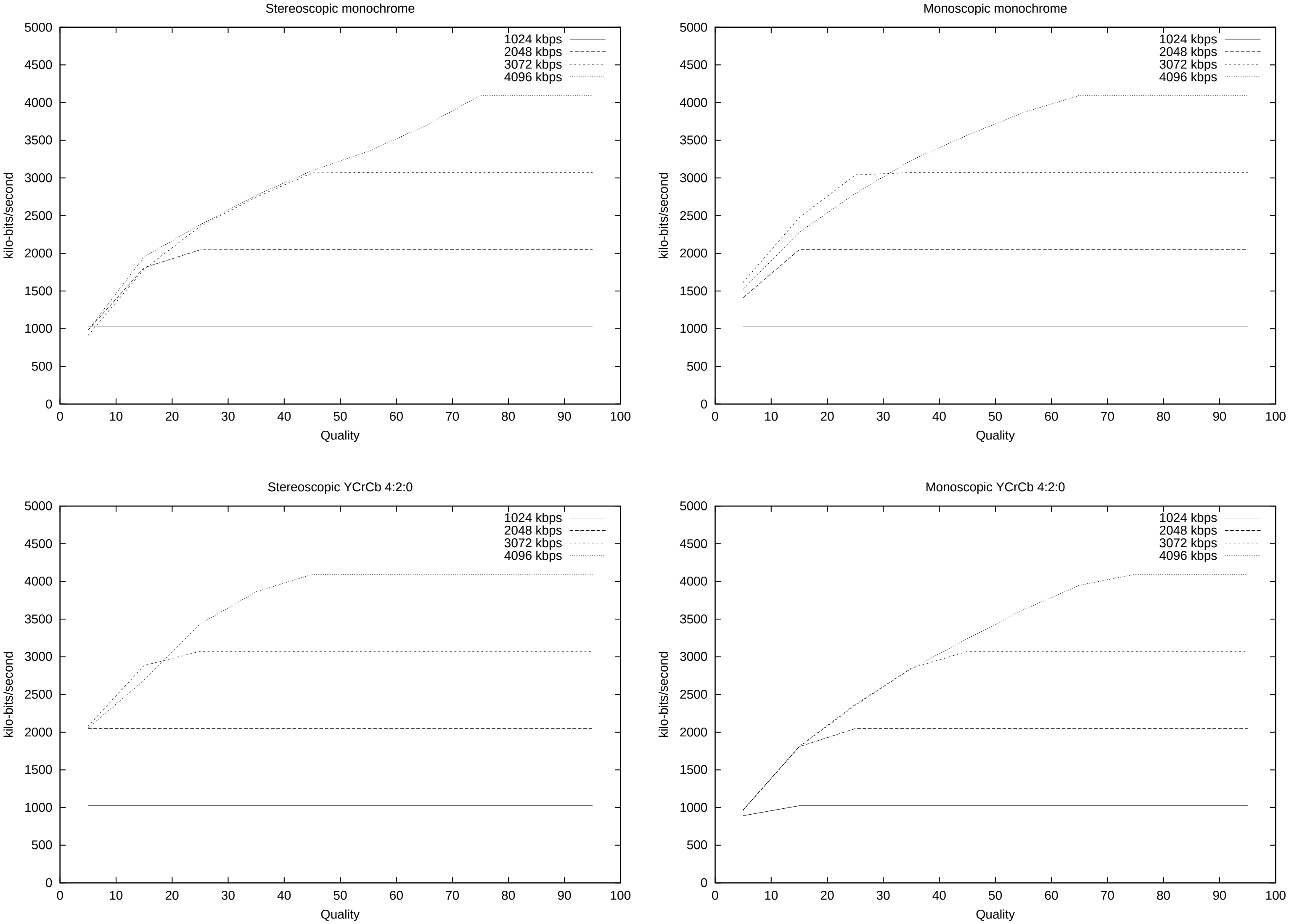


Figure 6.5: Actual bandwidth usage with varying parameters

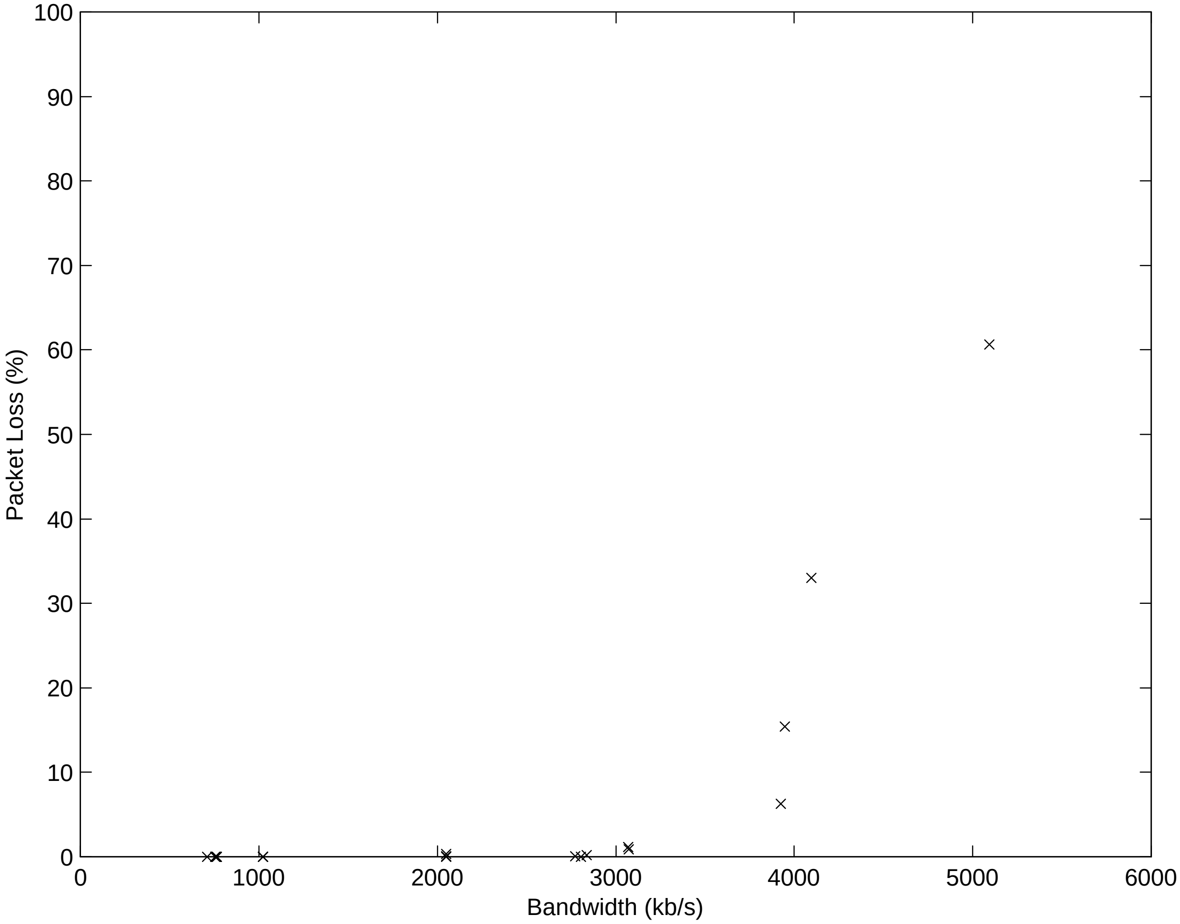


Figure 6.6: Packets lost with increasing bandwidth

Chapter 7

Discussion

A summary of the contributions of this thesis is presented and an outline of future work is given.

7.1 Overview

This thesis has presented a broad coverage of the technology and facts required for the implementation of a digital stereo video transmission system, using the Internet as the transmission medium. The topics covered were:

- the capabilities of the human binocular vision system,
- stereoscopic display techniques,
- capturing stereo video using video cameras,
- compression of stereo images/video, and
- transmission of video over the Internet

We saw that up to 2 minutes of arc (on average) can be stereoscopically fused, without the need for the observer to verge their eyes. One of the key factors in viewer discomfort in stereoscopic displays is caused by verging the eyes on a point not in the plane of the screen, and therefore, not at the distance at which the eyes are focussed. We can adjust various parameters in the system to ensure that the necessity to verge the eyes is kept to a minimum, or eliminated completely and therefore improve viewer comfort. We also saw that some degree of vertical and rotational disparity can be accomodated, however, these too should be eliminated or minimised.

Another important aspect of the human binocular vision system is that monoscopic depth cues will override binocular depth cues if they contain conflicting information. In order to avoid this happening, the visual field (or perpective) seen in the stereoscopic display must closely match the visual field a person would expect to see. Failure to do so results in stereoscopic depth cues being suppressed, which defeats the purpose of having a stereoscopic system.

The Robot Command Station has two screens, and onto each of these screens are projected two images. It was shown that a normal PC did not have the computational capacity to generate the four separate images sufficiently fast to give an acceptable frame rate. This implies that in order to make the Robot Command Station (and similar 2 screen systems) feasible, special computer hardware is required.

In experimenting with circular polarisation, we found that the circular polarising film did not provide sufficient extinction. It was possible to stereoscopically fuse an image using circular polarisation, but there was severe ghosting visible. The cause of this was deemed to be due to the quarter waveplate being tuned to a wavelength in the middle of visible spectrum, and not being broadband enough to produce effective results. Linear polarisation, in the optimum position, produced far superior results.

Aligning two DLP projectors to display on the same screen is problematic. With CRT projectors, there is a great deal of flexibility in the shape and size of the projected image. This means that it is possible to produce a good image on a screen from different projection points. With the DLP projectors used in this project, the only adjustment was the size of the image, and keystoning adjustment which used pixel interpolation rather than any optical means. In order to use DLP projectors on the same screen, a prism system may need to be employed to obtain more satisfatory results.

In terms of capturing, we learnt that the optimum geometry of capturing stereo video is very much related to how it is intended to be displayed. The ideal stereo camera configuration must be deliberately constructed for a given display. Since this is rarely possible, some more general solutions were arrived at. The goal is to maintain correct perspective and to minimise horizontal disparity. This means that the stereo cameras should be verged on the object of interest. For example, if the footage is of a landscape with all objects of interest far away then the cameras should be verged on an object far away (i.e. the cameras should be near parallel). Conversely, if the object of interest is close then the cameras should be verged on the object. The caveat to this is that if there is something in the distance of interest as well then this may not be stereoscopically fusable by an observer and a compromise will have to be made.

Even though the cameras used in this project produced a high quality digital signal, it was determined that the analog signal was superior in a practical sense. This was due to the fact that the digital signal would have to be decompressed before processing. In the future, with higher performance computers and greater acceptance of the IEEE-1394 standard, this digital signal (or a similar standard) will become more practical.

With current computing performance, JPEG style image compression gives the best performance in terms of computational cost. It is not as good as Wavelet image compression or MPEG video compression, but both of these are too computationally expensive to achieve an acceptable frame rate without hardware assistance on current commodity computers.

Through various experiments with the transmission of video/image data over the Internet (locally and overseas), we have discovered several things about the Internet protocols relating to the transmission of real-time data. Firstly, TCP is not suitable for the transmission of real-time data over long distances. This is primarily due to the fact that TCP uses acknowledgements to ensure that data is received correctly. Over a long distance (network wise), this dramatically increases latency which is something that must be minimised in real-time transmission.

UDP should not be used for the transmission of data that is critical (i.e. cannot be lost). For data that can cope with loss of information, such as audio and video, transmission via UDP is acceptable. Transmission via UDP is also preferable for the transmission of real-time data due to the fact that it does not require acknowledgements, and that it has very low protocol overheads and consequently low latency.

Finally, in the transmission of high bit-rate data via UDP, efforts should be made to ensure that the transmission of the data is consistent. Periodic bursts of information should be avoided as this increases the likelihood that information will be lost. Therefore, transmission should be monitored and controlled via buffering to maintain as consistent a flow of data as can be obtained.

7.2 Future Work

The implementation of the hardware and software in this project is at the experimental stage, and more work will be required to shape this into a useful product. There are strong desires to use Wedge style systems for remote collaboration and teaching. Clearly, the transmission of data of all kinds between remotely located Wedges will become increasingly desirable. Before this can be done, the software developed for this project will have to be substantially improved. Improvements such as:

- performance optimisation,
- increasing the robustness of the system,
- bi-directional transmission,
- the addition of synchronised audio, and
- the transmission of other data types, e.g. 3-d models.

Appendix A

Physics of Polarisation

A.1 Overview

This appendix is intended to give the reader an appreciation of the physics of polarisation in order to understand how some of the stereo display technologies work.

A.2 The Electromagnetic Plane Wave

The *electromagnetic field* is described by several related quantities:

- **E** - the electric vector
- **B** - the magnetic induction
- **j** - the electric current density
- **D** - the electric displacement
- **H** - the magnetic vector

These quantities are related by Maxwell's equations as follows:

$$\text{curl } \mathbf{H} - \frac{1}{c}\dot{\mathbf{D}} = \frac{4\pi}{c}\mathbf{j} \tag{A.1}$$

$$\text{curl } \mathbf{E} + \frac{1}{c}\dot{\mathbf{B}} = 0 \tag{A.2}$$

$$\text{div } \mathbf{D} = 4\pi\rho \tag{A.3}$$

$$\text{div } \mathbf{B} = 0 \tag{A.4}$$

For time-harmonic fields, isotropic materials with bodies at rest (or having slow motion relative to each other), the following relations, known as the *Material equations* also hold:

$$\mathbf{j} = \sigma \mathbf{E} \tag{A.5}$$

$$\mathbf{D} = \epsilon \mathbf{E} \tag{A.6}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{A.7}$$

Where:

- σ is the specific conductivity,
- ϵ is the dielectric constant, and
- μ is the magnetic permeability.

Of particular interest is the simplest electromagnetic field, the plane wave. An electromagnetic plane wave has vectors **E** and **H** that are time-harmonic. In this case, given some time-harmonic electric field vector, **E**, the following relationships hold:

$$\mathbf{H} = \sqrt{\frac{\epsilon}{\mu}} \mathbf{s} \times \mathbf{E} \tag{A.8}$$

$$\mathbf{E} \cdot \mathbf{s} = \mathbf{H} \cdot \mathbf{s} = 0 \tag{A.9}$$

Where **s** is the direction of propagation of the electromagnetic wave. If we set **s** to be along the z-axis, we know that **E** will have cartesian components restricted to the x and y axes, and **E** can be written as:

$$\mathbf{E} = \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = \begin{bmatrix} a_1 \cos(\omega t + \delta_1) \\ a_2 \cos(\omega t + \delta_2) \\ 0 \end{bmatrix} \tag{A.10}$$

Where ω is the frequency and δ_1 and δ_2 are the phase components of the electromagnetic wave.

A.3 Polarised Light

Light is polarised, either linearly, circularly or eliptically depending on the phase difference between the x and y components of the electric field vector. If we defined the phase difference as:

$$\delta = \delta_2 - \delta_1 \tag{A.11}$$

and set $a_1 = a_2$ then we can see that a plane wave will be linearly polarised if $\delta = m\pi$ where $m = 0, \pm 1, \pm 2, \ldots$. Similarly if $\delta = \pm\frac{\pi}{2}$ then the plane wave will be circularly polarised. These conditions can be seen in Figure A.1 in which the z-direction is out of the page. All other cases of polarisation are elliptical, although, if $a_1 \ll a_2$ or visa-versa then the plane wave is approximately linearly polarised.

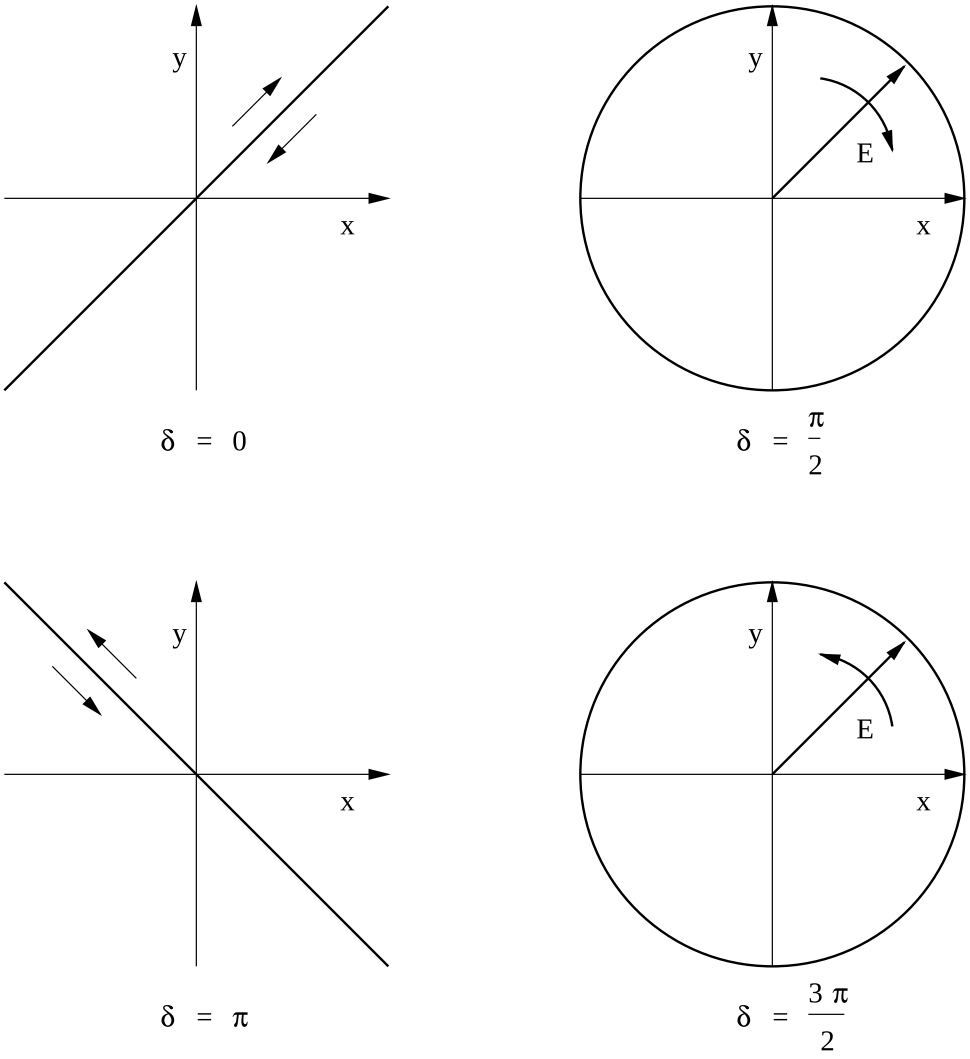
A.4 Polarising Filters

Polarising filters for both linear and circular polarisation use polymer films with different attenuation and refractive index in different directions [1]. These materials are called *birefringent*. In general, light from a light source is a superposition of many plane waves with varying polarisations and polarisation orientations.

Linear polarising film has very high attenuation in one direction, and very low attenuation perpendicular to this direction. This means that the light passing through the polarising film will have its electric field vector attenuated in directions that are not in the plane of polarisation. The result is that light, after passing through the film is approximately plane polarised.

Circularly polarising films contain a linear polarising film and a quarter wave retardation film. The light first passes through a linear polarisation film and becomes linearly polarised. The quarter wave retardation film is oriented at 45 or 135 degrees to the plane of polarisation depending on whether left or right circular polarisation is required. The quarter wave retardation film retards light passing through such that the electric field vector is retarded by $\frac{\lambda}{4}$ in one direction with respect to the perpendicular direction. So when linearly polarised light passes through a quarter wave retardation film at 45 or 135 degrees, the linear polarised light is changed into left or right circularly polarised light. Unfortunately, the quarter wave retardation film can only be tuned to a single wavelength of light, and is therefore only effective over a narrow band of electromagnetic spectrum.

[1] Other methods are available for filtering or creating polarised light, but we restrict ourselves here to cheap methods using commerically available film.



$\delta = 0$

$\delta = \frac{\pi}{2}$

$\delta = \pi$

$\delta = \frac{3\pi}{2}$

Figure A.1: Linear and Circular Polarisation modes

Appendix B

OpenGL and Stereoscopic Displays

B.1 Overview

This section provides a description of a method for calculating perspective parameters for stereoscopic displays.

B.2 Calculating Perspective Parameters

B.2.1 Introduction

This section presents a generic series of steps for calculating the parameters for the `gluLookAt` and `glFrustum` commands of the OpenGL API. The particular application here presents a true perspective view for a viewer looking at an arbitrary screen. This can be applied to systems such as the Wedge and the RCS.

B.2.2 Derivation of parameters for a single display screen

B.2.2.1 The Display Screen

Given an arbitrary, real world reference frame, we can define the geometry of a screen using a position vector for each corner of the screen as in Figure B.1.

Where the position vector:

$\mathbf{s}_{tl}$ is the top left corner of the display screen
$\mathbf{s}_{tr}$ is the top right corner of the display screen
$\mathbf{s}_{bl}$ is the bottom left corner of the display screen
$\mathbf{s}_{br}$ is the bottom right corner of the display screen

There are two restrictions on the geometry of the screen, these are:

1. the screen is planar, and

2. the screen is rectangular.

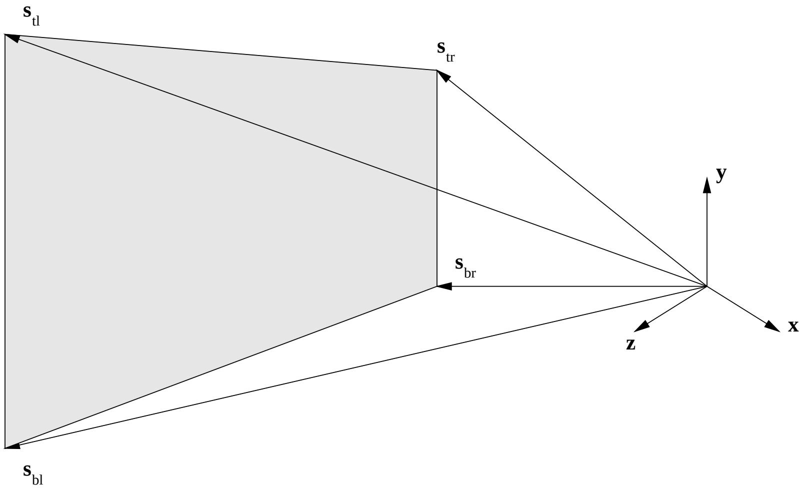


Figure B.1: The mathematical definition of a display screen

Or more mathematically:

$$\begin{gathered}\mathbf{s}_{br} - \mathbf{s}_{bl} + \mathbf{s}_{tl} - \mathbf{s}_{tr} = 0 \\ \text{and} \\ (\mathbf{s}_{tl} - \mathbf{s}_{tr}) \cdot (\mathbf{s}_{br} - \mathbf{s}_{tr}) = 0\end{gathered} \tag{B.1}$$

B.2.2.2 The Viewer

We wish to specify the position of the viewer in the same reference frame as the display screen for both mono and stereo-scopic viewing. For mono-scopic viewing, we simply need a position vector for the position of the viewer, which we define as $\mathbf{e}_o$ (eye origin).

For stereo-scopic viewing we need two extra variables to define the position of the eyes:

$\hat{\mathbf{e}}_a$ is the direction of the line passing through both eyes
e_s is the distance between the eyes

Hence, the position of the left and right eyes may be calculated using the following equations:

$$\begin{aligned}\mathbf{e}_l &= \mathbf{e}_o - e_s \cdot \hat{\mathbf{e}}_a \\ \mathbf{e}_r &= \mathbf{e}_o + e_s \cdot \hat{\mathbf{e}}_a\end{aligned} \tag{B.2}$$

The derivations in the following sections assume that the viewer is positioned in front of the screen.

B.2.3 The Model Transformation

Now that we have defined our system, we need to calculate the paramters for the `gluLookAt` and `glFrustum` functions. Since the `glFrustum` function expects a rectangular shaped frustum, this implies that the near-clipping plane must be parallel to the plane of the screen. This further implies that the line from the eye to the centre of vision must be normal to the plane of the screen.

The normal of the screen can be found using the cross-product:

$$\mathbf{s}_n = (\mathbf{s}_{tl} - \mathbf{s}_{tr}) \times (\mathbf{s}_{br} - \mathbf{s}_{tr}) \tag{B.3}$$

Given the assumption that the view is in front of the screen, $\hat{\mathbf{s}}_n$ is in the direction from the screen to the viewer.

The general equation of a plane is:

$$ax + by + cz + d = 0 \tag{B.4}$$

where the homogenous normal vector of this plane is;

$$\begin{bmatrix} a \\ b \\ c \\ 1 \end{bmatrix}$$

Hence, to obtain the equation of the plane of the screen, we need to solve Equation B.4 for d using the normal of the plane of the screen, $\mathbf{s}_n$, and a point on the plane, $\mathbf{s}_{tr}$. Hence

$$\begin{aligned} s_{n_x}s_{tr_x} + s_{n_y}s_{tr_y} + s_{n_z}s_{tr_z} + d &= 0 \\ \mathbf{s}_n \cdot \mathbf{s}_{tr} + d &= 0 \end{aligned}$$

Hence, the equation of the plane of the display screen is;

$$s_{n_x}x + s_{n_x}y + s_{n_x}z - \mathbf{s}_n \cdot \mathbf{s}_{tr} = 0 \tag{B.5}$$

For a particular eye e_α, i.e. e_o for monoscopic viewing or e_l and e_r for stereoscopic viewing, we must find the centre of vision. Since the line from the eye to the centre of vision must be normal to the display screen, the centre of vision can be written as:

$$\mathbf{c}_\alpha = \mathbf{e}_\alpha + k \cdot \hat{\mathbf{s}}_n = \begin{bmatrix} e_{\alpha_x} + k \cdot \hat{s}_{n_x} \\ e_{\alpha_x} + k \cdot \hat{s}_{n_x} \\ e_{\alpha_x} + k \cdot \hat{s}_{n_x} \\ 1 \end{bmatrix} \tag{B.6}$$

A robust method of specifying the centre of vision is to add the further constraint that the centre of vision must lie in the plane of the screen. Hence the position $\mathbf{c}_\alpha$ must satisfy the equation of the screen plane given in Equation B.5. Hence, substituting Eq. B.6 into Eq. B.5 and solving for k;

$$\begin{aligned}\hat{s}_{n_x}\left(e_{\alpha_x}+k\hat{s}_{n_x}\right)+\hat{s}_{n_y}\left(e_{\alpha_y}+k\hat{s}_{n_y}\right)+\hat{s}_{n_z}\left(e_{\alpha_z}+k\hat{s}_{n_z}\right)-\hat{\mathbf{s}}_n\cdot\mathbf{s}_{tr} &= 0\\ \hat{s}_{n_x}e_{\alpha_x}+\hat{s}_{n_y}e_{\alpha_y}+\hat{s}_{n_z}e_{\alpha_z}+k\left(\hat{s}^2_{n_x}+\hat{s}^2_{n_y}+\hat{s}^2_{n_z}\right)-\hat{\mathbf{s}}_n\cdot\mathbf{s}_{tr} &= 0\\ \mathbf{e}_\alpha\cdot\hat{\mathbf{s}}_n+k-\hat{\mathbf{s}}_n\cdot\mathbf{s}_{tr} &= 0\end{aligned}$$

Therefore;

$$k=\hat{\mathbf{s}}_n\cdot\mathbf{s}_{tr}-\mathbf{e}_\alpha\cdot\hat{\mathbf{s}}_n \qquad \text{(B.7)}$$

Hence, the centre of view for a particular eye can be found using the following equation;

$$\mathbf{c}_\alpha=\mathbf{e}_\alpha+\left(\hat{\mathbf{s}}_n\cdot\mathbf{s}_{tr}-\mathbf{e}_\alpha\cdot\hat{\mathbf{s}}_n\right)\hat{\mathbf{s}}_n \qquad \text{(B.8)}$$

B.2.4 The Perspective Transformation

In the previous section, we computed the normal of the plane of the wall and this normal was directed from the wall to the viewer. For convenience, we wish the normal of the near-clipping plane to be directed in the opposite direction, therefore:

$$\hat{\mathbf{n}}_n=-\hat{\mathbf{s}}_n \qquad \text{(B.9)}$$

If n_d is the near clipping distance, then a point on the near-clipping plane is;

$$\mathbf{n}_o=\mathbf{e}_\alpha+n_d\hat{\mathbf{n}}_n \qquad \text{(B.10)}$$

The coordinate, $\mathbf{n}_o$, is the reference point from which the (top,left) and (bottom,right) parameters of the `glFrustum` function are measured.

The equation of the near-clipping plane can be obtained using the same method in the previous section to obtain;

$$\hat{n}_{n_x}x+\hat{n}_{n_y}y+\hat{n}_{n_z}z-\mathbf{n}_o\cdot\hat{\mathbf{n}}_n=0 \qquad \text{(B.11)}$$

For each corner of the display screen, we project a line from the eye to the corner and find where this line intersects the near-clipping plane.

Firstly, we need to find a parametric equation of the lines passing through $\mathbf{e}_\alpha$ and the corners $\mathbf{s}_{tl}$, $\mathbf{s}_{tr}$, $\mathbf{s}_{bl}$, and $\mathbf{s}_{br}$.

For a particular corner of the display screen, $\mathbf{s}_\alpha$, the parametric equation of the line passing through the eye and this corner is;

$$\begin{aligned} x &= e_{\alpha_x} + \left(s_{\alpha_x} - e_{\alpha_x}\right) t \\ y &= e_{\alpha_y} + \left(s_{\alpha_y} - e_{\alpha_y}\right) t \\ z &= e_{\alpha_z} + \left(s_{\alpha_z} - e_{\alpha_z}\right) t \end{aligned} \tag{B.12}$$

Where t is the parametric variable.

Next, we substitute this parametric equation into the equation of the near-clipping plane and solve for t.

$$\begin{array}{rcl} \hat{n}_{n_x}\left(e_{\alpha_x} + \left(s_{\alpha_x} - e_{\alpha_x}\right) t\right) + & & \\ \hat{n}_{n_y}\left(e_{\alpha_y} + \left(s_{\alpha_y} - e_{\alpha_y}\right) t\right) + & & \\ \hat{n}_{n_z}\left(e_{\alpha_z} + \left(s_{\alpha_z} - e_{\alpha_z}\right) t\right) - & & \\ \mathbf{n}_o \cdot \hat{\mathbf{n}}_n & = & 0 \\ \hat{\mathbf{n}}_n \cdot \mathbf{e}_\alpha + \hat{\mathbf{n}}_n \cdot \left(\mathbf{s}_\alpha - \mathbf{e}_\alpha\right) t - \mathbf{n}_o \cdot \hat{\mathbf{n}}_n & = & 0 \end{array}$$

Therefore:

$$t = \frac{\mathbf{n}_o \cdot \hat{\mathbf{n}}_n - \hat{\mathbf{n}}_n \cdot \mathbf{e}_\alpha}{\hat{\mathbf{n}}_n \cdot \left(\mathbf{s}_\alpha - \mathbf{e}_\alpha\right)} \tag{B.13}$$

Substituting this value of t into Equation B.12 we can obtain the point at which the near-clipping plane is intersected. For each of the corners, we can obtain the intersection, i.e.:

$$\begin{array}{lll} \text{line} & \mathbf{e}_\alpha \rightarrow \mathbf{s}_{tl} & \text{intersects the near clipping plane at} & \mathbf{n}_{tl} \\ \text{line} & \mathbf{e}_\alpha \rightarrow \mathbf{s}_{tr} & \text{intersects the near clipping plane at} & \mathbf{n}_{tr} \\ \text{line} & \mathbf{e}_\alpha \rightarrow \mathbf{s}_{bl} & \text{intersects the near clipping plane at} & \mathbf{n}_{bl} \\ \text{line} & \mathbf{e}_\alpha \rightarrow \mathbf{s}_{br} & \text{intersects the near clipping plane at} & \mathbf{n}_{br} \end{array}$$

Next we define a reference frame with its origin at the point $\mathbf{n}_o$ as follows:

$$\begin{aligned} \mathbf{x}' &= \mathbf{n}_{tr} - \mathbf{n}_{tl} \\ \mathbf{y}' &= \mathbf{n}_{tr} - \mathbf{n}_{br} \\ \mathbf{z}' &= -\hat{\mathbf{n}}_n \end{aligned} \tag{B.14}$$

So the rotation matrix to transform directions to this new reference frame is:

$$\mathbf{R} = \begin{bmatrix} \hat{x}'_x & \hat{y}'_x & \hat{z}'_x \\ \hat{x}'_y & \hat{y}'_y & \hat{z}'_y \\ \hat{x}'_z & \hat{y}'_z & \hat{z}'_z \end{bmatrix} \tag{B.15}$$

And the homogenous transformation to transform coordinates specified in the base reference frame to this new reference frame is;

$$\mathbf{T} = \begin{bmatrix} & \mathbf{R} & & \mathbf{n}_o \\ & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{B.16}$$

And since this is an homogenous transform, the inverse of this is;

$$\mathbf{T}^{-1} = \begin{bmatrix} & \mathbf{R}^T & & -\mathbf{R}^T\mathbf{n}_o \\ & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{B.17}$$

The (top, left) and (bottom,right) coordinates can then be found using the following formulae;

$$\begin{bmatrix} \text{left} \\ \text{top} \\ 0 \\ 1 \end{bmatrix} = \mathbf{T}^{-1}\mathbf{n}_{tl} \quad \text{and} \quad \begin{bmatrix} \text{right} \\ \text{bottom} \\ 0 \\ 1 \end{bmatrix} = \mathbf{T}^{-1}\mathbf{n}_{br} \tag{B.18}$$

Appendix C

Stereo Camera and Display Geometry

C.1 Overview

This appendix gives a derivation for the calculation of various parameters related to the capture of stereo video and subsequent display using a stereoscopic display system. The primary use of these derivations is in the design of stereo camera systems for capturing stereoscopic video footage.

C.2 The Parameters

For simplicity and to be more appropriate for the cameras used in this project, we make the following assumptions:

- the image sensor has no offset, that is, the centre of the image sensor is aligned to the centre of view,
- the field of view is symmetric about the centre and the same for both cameras,
- the objects of interest are located along the perpendicular bisector of the two cameras,
- resolution is sufficient so that small angular disparities can be seen,
- the observer is located so that the distance from the screen is large compared to the distance between the eyes, and
- the observer is located so that the perpendicular bisector of the eyes intersects the centre of the screen.

From Figure C.1, that the important parameters of the stereo camera configuration are:

- d_c — the camera separation,
- d_v — the vergence distance, and

- θ_{fov} — the field of view of the cameras.

Similarly, the important parameter for the stereo display system is the horizontal field of view that an observer sees, ϕ_{fov}. As well as these parameters, the limits of stereoscopic fusion are also important and these are denoted $\phi_{max\ crossed}$ and $\phi_{max\ uncrossed}$, and represent the maximum crossed and uncrossed angular disparities that can be stereoscopically fused.

C.3 Derivations

C.3.1 The range of distances imaged by both cameras

The vergence angle, θ_v, can be found easily from the camera separation and vergence distance.

$$\theta_v = \tan^{-1}\left(\frac{2d_v}{d_c}\right) \tag{C.1}$$

A special case is that when the cameras have their optical axes parallel, then the vergence distance is infinity and the vergence angle is 90°.

If we define two angles, θ_{max} and θ_{min}, as follows:

$$\theta_{max} = \theta_v + \frac{1}{2}\theta_{fov} \tag{C.2}$$

$$\theta_{min} = \theta_v - \frac{1}{2}\theta_{fov} \tag{C.3}$$

Then the maximum and minimum distances that can be imaged by both cameras can be determined by the following equations.

$$d_{max} = \begin{array}{ll} \infty & \theta_{max} \geq 90^{\mathrm{o}} \\ \frac{1}{2}d_c \tan\theta_{max} & \text{otherwise} \end{array} \tag{C.4}$$

$$d_{min} = \begin{array}{ll} 0 & \theta_{max} \leq 90^{\mathrm{o}} \\ \frac{1}{2}d_c \tan\theta_{min} & \text{otherwise} \end{array} \tag{C.5}$$

C.4 The range of distances that can be stereoscopically fused

In order to determine the range of distances that can be stereoscopically fused by an observer for a given stereo camera configuration, we first need to find an expression for the disparity caused by an object at a specified distance. For a verged camera configuration, i.e. the optical axes of the cameras are not parallel, two distances must be calculated. One for crossed disparity for distances less than the vergence distance and the other for uncrossed disparity for distances more than the vergence distance.

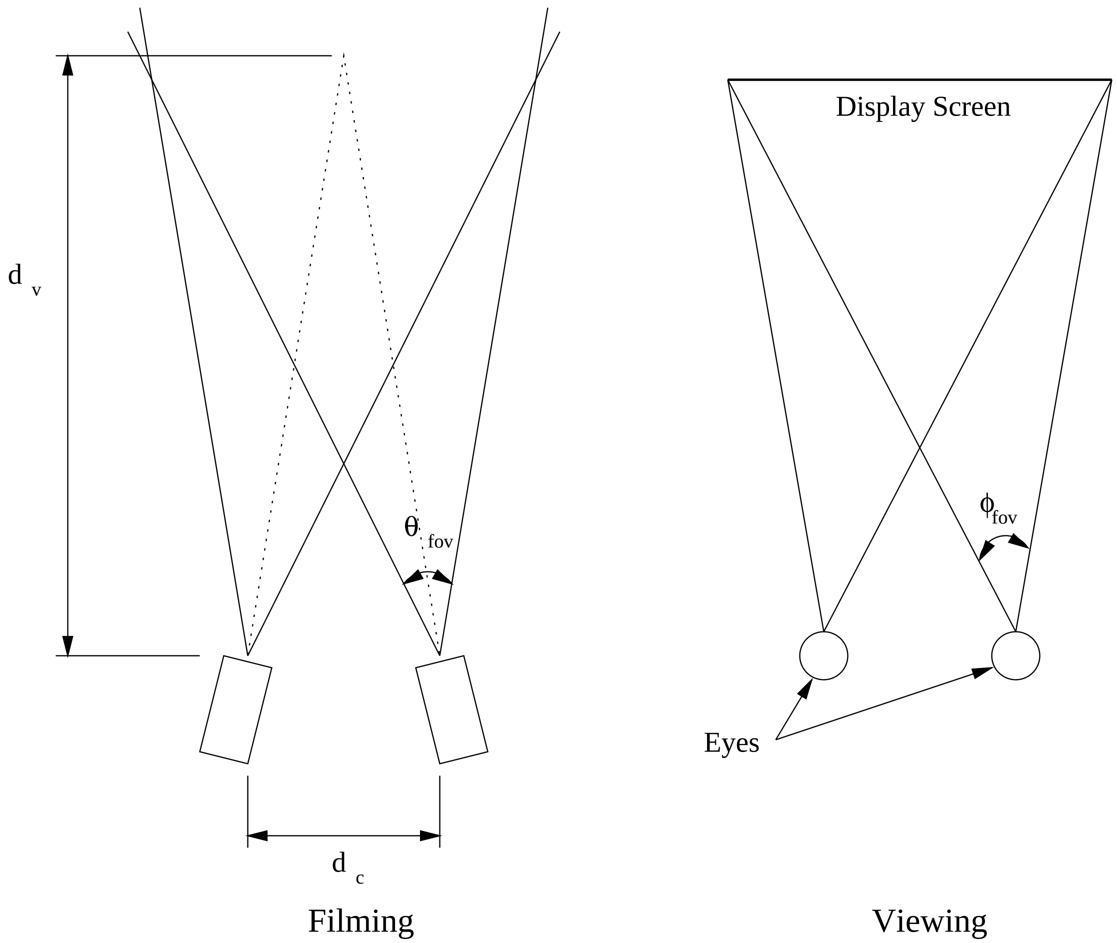


Figure C.1: The stereo camera and display parameters

Using the expression for the vergence angle given above, and the equation for the angle for an object at a distance, d_o, along the perpendicular bisector of the two cameras:

$$\theta_o = 2\tan^{-1}\left(\frac{2d_o}{d_c}\right) \tag{C.6}$$

The crossed angular disparity that is imaged by the stereo video cameras is:

$$\theta_{crossed} = 2\left(\theta_v - \theta_o\right) \tag{C.7}$$

This angular disparity is varied by the ratio of the field of view of the camera with respect to the field of view of the display, so that the disparity an observer will see is:

$$\phi_{crossed} = \theta_d \times \frac{\phi_{fov}}{\theta_{fov}} \tag{C.8}$$

To obtain an expression that is purely in terms of the parameters of the stereo camera configuration and the stereo display system, we substitute the various variables to obtain the following expression for the crossed angular disparity caused by an object at a distance of d_o.

$$\phi_{crossed} = 2\frac{\phi_{fov}}{\theta_{fov}}\left[\tan^{-1}\left(\frac{2d_v}{d_c}\right) - \tan^{-1}\left(\frac{2d_o}{d_c}\right)\right] \tag{C.9}$$

Using the same method, we can obtain a similar expression for the uncrossed angular disparity caused by an object at a distance of d_o.

$$\phi_{uncrossed} = 2\frac{\phi_{fov}}{\theta_{fov}}\left[\tan^{-1}\left(\frac{2d_o}{d_c}\right) - \tan^{-1}\left(\frac{2d_v}{d_c}\right)\right] \tag{C.10}$$

We can rearrange these two equations by substituting the maximum crossed and uncrossed disparities that can be seen, to obtain expressions for the minimum and maximum fusable distances as follows:

$$d_{min} = \frac{d_c}{2}\tan\left[\tan^{-1}\left(\frac{2d_v}{d_c}\right) - \frac{\phi_{max\,crossed}\phi_{fov}}{2\theta_{fov}}\right] \tag{C.11}$$

$$d_{max} = \frac{d_c}{2}\tan\left[\frac{\phi_{max\,uncrossed}\phi_{fov}}{2\theta_{fov}} + \tan^{-1}\left(\frac{2d_v}{d_c}\right)\right] \tag{C.12}$$

It can be seen geometrically that for the special case where the two stereo cameras are parallel, that the angular disparity approaches zero as the distance of an object increases. Therefore, the maximum distance that can be fused with parallel cameras is infinity, and the minimum distance can be calculated using a modified version of the equation above for the minimum fusable distance.

$$d_{\infty} = \frac{d_c}{2}\tan\left[90^{o} - \frac{\phi_{max\,crossed}\phi_{fov}}{2\theta_{fov}}\right] \tag{C.13}$$

Hence, for verged cameras the range of fusable distances is $[d_{min} \cdots d_{max}]$ and similarly for parallel cameras the range of fusable distances is $[d_{\infty} \cdots \infty]$.

Appendix D

Description of the Stereo Video Streaming Software

D.1 Introduction

This appendix provides details of the operation of the Stereo Video Streaming software that was implemented as part of this thesis.

D.2 Compression Library

The compression library, *MJPEG*, uses a structure for the initialisation of compression and decompression settings called `mjpeg_options_t`. This structure specifies the image width and height, and the input, output and encoding formats for the image. The library transcodes the image from the input format to the encoding format during compression, and likewise, transcodes the image from the encoding format to the output format during decompression. There is also an option to specify the Discrete Cosine Transform algorithm, however only one of these is implemented so this option is not operational.

To initialise the *MJPEG* library, the `mjpeg_init` function is used. An example of initialisation of the library follows:

```
mjpeg_options_t options;

options.dctmethod = MJPEG_DCT_FASTEST;
options.width = 768;
options.height = 576;
options.inputformat = MJPEG_IMGFMT_RGB24;
options.encodeformat = MJPEG_ENCFMT_YUV420;
options.outputformat = MJPEG_IMGFMT_RGB24;

if ( mjpeg_init( &options ) < 0 ) {
  /* error handling code here */
}

/* compress/decompress images */

mjpeg_free();
```

D.2.1 Compression

To compress an image requires the use of two functions, `mjpeg_encode_start` to specify the images to compress, and `mjpeg_encode_next` to encode the next restart intervals worth of image. An example of encoding a stereo image follows:

```
int startbit;
int intervals_remaining;

/* Memory must be allocated for these variables */
char *left_image;
char *right_image;
char *bitstream;

if ( mjpeg_encode_start( left_image,
                         right_image,
                         75, /* quantization level */
                         1 /* restart interval */ ) < 0 ) {
  /* error handling code here */
}

startbit = 0;
do {
  intervals_remaining =
    mjpeg_encode_next( &startbit,
                       bitstream,
                       1024 /* bytes allocated for bitstream */ );
} while( intervals_remaining > 0 )
```

D.2.2 Decompression

Decompression is similar to compression, in that two functions are used, `mjpeg_decode_start` and `mjpeg_decode_interval`.

```
int startbit;
int interval;
int totalIntervals;
```

```
/* Memory must be allocated for these variables */
char *left_image;
char *right_image;
char *bitstream;

if ( mjpeg_decode_start( left_image,
                         right_image,
                         75,
                         1 ) < 0 ) {
  /* error handling code here */
}

startbit = 0;
for ( interval = 0; interval < totalIntervals; interval ++ ) {
  mjpeg_decode_interval( interval,
                         &startbit,
                         bitstream );
}
```

D.3 The Streaming Server

The streaming server uses threads to multiplex the compression and transmission of the stereo video. These threads are coupled by a data queue which is used to pass video data from the compressor to the transmitter. The compression threads loop consists of the following steps:

1. Capture the image(s) from the analog frame grabbers
2. For each restart interval in the image(s):
   (a) compress the restart interval
   (b) add the compressed bit stream to the queue

Similarly, the transmission thread consists of the following series of steps:

1. Retrieve the compressed bit stream from the queue
2. Pack the bit stream into a transmission packet
3. If the transmission packet has reached the target size
   (a) transmit the packet
   (b) empty the packet

Full source code for the compression and streaming software is included on the attached CD.

Bibliography

1. Pathfinder mission homepage. World Wide Web. http://mars.jpl.nasa.gov/.

2. Quicktime webpage. http://www.apple.com/quicktime.

3. Real networks webpage. http://www.real.com/.

4. 1394 Trade Association. *AV/C Camera Subunit Specification*, January 1999. Version 1.0.

5. Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. Image coding using wavelet transform. *IEEE Transactions on image processing 1*, 2 (April 1992), 205–220.

6. Arakawa, Y. Quantitative measurements of visual fields for colors. *American Journal of Ophthamology 36* (1953), 1594—1601.

7. Barbe, D. F., Ed. *Charged-Coupled Devices*, vol. 39 of *Topics in Applied Physics*. Springer-Verlag, 1980.

8. Berc, L., Fenner, W., Frederick, R., McCanne, S., and P.Stewart. Rtp payload format for jpeg compressed video. Tech. rep., Internet Engineering Task-force, 1998. RFC-2435.

9. Board, O. A. R. *OpenGL Reference Manual*. Addison Wesley, 1996.

10. Borman, C., Cline, L., G.Deisher, Gardos, T., Maciocco, C., Newell, D., Ott, J., Sullivan, G., Wenger, S., and Zhu, C. Rtp payload format for the 1998 version of itu-t rec. h.263 video (h.263+). Tech. rep., Internet Engineering Task-force, 1998. RFC-2429.

11. Boswell, R., and Gardner, H. The wedge homepage. World Wide Web. http://wedge.anu.edu.au.

12. Brindley, G. S. *Physiology of the Retina and the Visual Pathway*. Edward Arnold, 1960.

13. Calderbank, A. R., Daubechies, I., Sweldens, W., and Yeo, B. Wavelet transforms that map integers to integers. In *Proceedings of the IEEE Conference on Image Processing* (1997).

14. Calderbank, A. R., Debauchies, I., Sweldens, W., and Yeo, B. Lossless image compression using integer to integer wavelet transforms. preprint.

15. CHAO, H., AND FISHER, P. An approach to fast integer reversible wavelet transforms for image compression. preprint.

16. CHARNWOOD, L. *An essay on binocular vision*. Hafner Publishing Company, 1950.

17. CLAYPOOLE, R., DAVIS, G., SWELDENS, W., AND BARANIUK, R. Nonlinear wavelet transforms for image coding. *Asilomar conference on signals, systems, and computers* (1997). preprint.

18. COMMISSION INTERNATIONALE DE L'ECLAIRAGE. *CIE Colorimetric Standard Observer Model*, 1931.

19. CRUZ-NEIRA, C., SANDIN, D., DEFANTI, T., KENYON, R., AND HART, J. The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM 35*, 6 (June 1992), 65–72.

20. DAUBECHIES, I., AND SWELDENS, W. Factoring wavelet transforms into lifting steps. Tech. rep., Bell Laboratories, Lucent Technologies, 1996.

21. DAVSON, H. *Physiology of the Eye*. Churchill, 1949.

22. DICKSON, G., AND LLOYD, A. *Open Systems Interconnection*. Prentice Hall, 1991.

23. DRASCIC, D. An investigation of monoscopic and stereoscopic video for teleoperation. Master's thesis, University of Toronto, 1991. Department of Industrial Engineering.

24. DROUET, T. Stereo video transmission. Tech. rep., Ecole polytechnique de l'universite de Nantes, 2000. Training at the Australian National University.

25. DUWAER, A. L., AND VAN DEN BRINK, G. What is the diplopia threshold. *Perception and Psychophysics 29* (1981), 295–309.

26. EDWARDS, T. Discrete wavelet transforms: Theory and implementation. *preprint* (1992).

27. EVANS, R. M. *The Perception of Color*. John Wiley and Sons, 1974.

28. FERNANDEZ, G., PERIASWAMY, S., AND SWELDENS, W. Liftpack: A software package for wavelet transforms using lifting. In *Wavelet Applications in Signal and Image Processing IV* (1996), M. Unser, A. Aldroubi, and A. F. Laine, Eds., Proceedings of the SPIE 2825.

29. GÄRTNER, K., AND SCHNEIDER, F. Teleoperation with compressed motion picture sequences. In *27^{th} International Symposium on Industrial Robots* (October 1996), International Federation of Robotics.

30. GOODMAN, J. W. *Introduction to Fourier Optics*, 2nd ed. McGraw-Hill, 1996.

31. GRAPS, A. An introduction to wavelets. *IEEE Computational Science and Engineering 2*, 2 (1995).

32. GRINBERG, V., PODNAR, G., AND SIEGEL, M. Geometry of binocular imaging. In *Stereoscopic Displays and Applications V* (February 1994), SPIE/IS&T, pp. 56–65.

33. GRINBERG, V., PODNAR, G., AND SIEGEL, M. Geometry of binocular imaging ii: The augmented eye. In *Stereoscopic Displays and Applications VI* (February 1995), SPIE/IS&T, pp. 142–149.

34. GROUP, T. I. J. JPEG Software. ftp://ftp.uu.net/graphics/jpeg/jpegsrc.v6b.tar.gz.

35. HD DIGITAL VCR CONFERENCE. *Specifications of Consumer-Use Digital VCRs using 6.3mm magnetic tape* (December 1994).

36. INSTITUTION OF ELECTRICAL ENGINEERS. *IEEE Standard for a High Performance Serial Bus*, 1995. IEEE Std 1394:1995.

37. INSTITUTION OF ELECTRICAL ENGINEERS. *Information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements–Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications*, 1998. IEEE Std 802.3:1998.

38. INSTITUTION OF ELECTRICAL ENGINEERS. *Information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, 1999. IEEE Std 802.11:2000.

39. INSTITUTION OF ELECTRICAL ENGINEERS. *IEEE Standard for Information Technology – Portable Operating System Interface (POSIX)–Part xx:Protocol Independant Interfaces*, 2000. IEEE Std 1003.1g:2000.

40. INTERNATIONAL ELECTROTECHNICAL COMMISSION. *Recording — Helical-scan digital video cassette recording system using 6.35mm magnetic tape for consumer use (525-60, 625-50, 1125-60 and 1250-50 systems) — Part 2: SD format for 525-60 and 625-50 systems*, August 1998. IEC 61834-2.

41. INTERNATIONAL ORGANISATION FOR STANDARDISATION. *Digital compression and Coding of Continuous-tone still images: Part 1: Requirements and Guidelines*, 1992. ISO/IEC 10918-1:1992.

42. INTERNATIONAL ORGANISATION FOR STANDARDISATION. *Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 2: Video*, 1993. ISO/IEC 11172-2:1993.

43. INTERNATIONAL ORGANISATION FOR STANDARDISATION. *Information technology – Generic coding of moving pictures and associated audio information: Video*, 1996. ISO/IEC 13818-2:1996.

44. INTERNATIONAL TELECOMMUNICATION UNION. *ITU-R Recommendation 601: Encoding parameters of digital television for studios*, 1993.

45. INTERNATIONAL TELECOMMUNICATION UNION. *Line transmission of non-telephone signals: Video codec for audiovisual services at $p \times 64$ kbits*, 1993. ITU-T Recommendation H.261.

46. INTERNATIONAL TELECOMMUNICATION UNION. *Video coding for low bit rate communications*, 1998. ITU-T Recommendation H.263.

47. JACOBSON, V., BRADEN, R., AND ZHANG, L. TCP extensions for high-speed paths. Tech. rep., Internet Engineering Task-force, 1992. RFC-1323.

48. KERNIGHAN, B. W., AND RITCHIE, D. M. *The C Programming Language*, second ed. Prentice Hall, 1988.

49. KIM, W. S., ELLIS, S. R., TYLER, M. E., HANNAFORD, B., AND STARK, L. Quantitative evaluation of perspective and stereoscopic displays in three-axis manual tracking tasks. *IEEE Transactions on Systems, Man, and Cybernetics* (1987), 61–72.

50. MAIMONE, M., MATTHIES, L., OSBORN, J., ROLLINS, E., TEZA, J., AND THAYER, S. A photo-realistic 3-D mapping system for extreme nuclear environments: Chernobyl. In *International Conference on Intelligent Robotic Systems* (1998), IEEE/RSJ. To appear.

51. MATSUNAGA, K., SHIDOJI, K., AND MATSUBARA, K. Comparison of operation efficiency for the insert task when using stereoscopic images with additional lines, stereoscopic images, and a manipulator with force feeback. In *Stereoscopic Displays and Virtual Reality Systems VI* (1999), J. O. Merritt, M. T. Bolas, and S. S. Fisher, Eds., vol. 3639, SPIE, pp. 50–56.

52. MINAMOTO, M., AND MATSUNAGA, K. Evaluation of stereoscopic video cameras synchronized with the movement of an operator's head on the teleoperation of the actual backhoe shovel. In *Stereoscopic Displays and Virtual Reality Systems VI* (1999), J. O. Merritt, M. T. Bolas, and S. S. Fisher, Eds., vol. 3639, SPIE, pp. 44–49.

53. MOTOKAWA, K. *Physiology of Color and Pattern Vision*. Springer-Verlag, 1970.

54. MUHLBACH, L., BOCKER, M., AND PRUSSOG, A. Telepresence in videocommunications: A study on stereoscopy and individual eye contact. *Human Factors 37* (1995), 290–305.

55. PADGHAM, C. A., AND SAUNDERS, J. E. *The Perception of Light and Colour*. G. Bell & Sons, 1975.

56. PASTOOR, S. Human factors in 3d imaging. Tech. rep., Heinrich-Hertz Institute, 2000.

57. Pennebaker, W. B., and Mitchell, J. L. *JPEG still image compression standard*. Van Nostrand Reinhold, 1992.

58. Pepper, R. L., Smith, D. C., and Cole, R. E. Stereo TV improves operator performance under degraded visibility conditions. *Optical Engineering 20* (1981), 579–585.

59. Postel, J. User datagram protocol. Tech. rep., Internet Engineering Task-force, 1980. RFC-768.

60. Postel, J. Transmission control protocol. Tech. rep., Internet Engineering Task-force, 1981. RFC-793.

61. Poynton, C. A. *A technical introduction to digital video*. John Wiley and Sons, 1996.

62. Regan, D., Ed. *Binocular Vision*, vol. 9 of *Vision and Visual Dysfunction*. Macmillan Press, 1991.

63. Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V. Rtp: A transport protocol for real-time applications. Tech. rep., Internet Engineering Task-force, 1996. RFC-1889.

64. Schulzrinne, H., Rao, A., and Lanphier, R. Real time streaming protocol (rtsp). Tech. rep., Internet Engineering Task-force, 1998. RFC-2326.

65. Society of Motion Picture and Television Engineers. *Television—Data structure for DV-based audio, data and compressed video — 25 and 50 Mb/s*, 1999. SMPTE 314M-1999.

66. Stevens, W. R. *TCP/IP Illustated, Volume 1: The Protocols*. Addison-Wesley, 1994.

67. Stevens, W. R. *Networking APIs: Sockets and XTI*, second ed., vol. 1 of *Unix Network Programming*. Prentice Hall, 1997.

68. Sweldens, W. The lifting scheme: a new philosophy in biorthogonal wavelet constructions. In *Wavelet Applications in Signal and Image Processing III* (1995), A. F. Laine and M. Unser, Eds., Proceedings of the SPIE 2569, pp. 68–79.

69. Sweldens, W. The lifting scheme: a custom-design construction of biorthogonal wavelets. *Journal of Applied and Computational Harmonic Analysis 3* (1996), 186–200.

70. Sweldens, W. The lifting scheme: a construction of second generation wavelets. *SIAM Journal on Mathematical Analysis 29*, 2 (1997).

71. Sweldens, W. Wavelets and the lifting scheme: A 5 minute tour. Tech. rep., Katholieke Universiteit Leuven, 1997.

72. Sweldens, W., and Schroder, P. Building your own wavelets at home. In *Wavelets in Computer Graphics* (1997), ACM SIGGRAPH Course Notes.

73. TRAVIS, J. Camserv Software. http://cserv.sourceforge.net/.

74. TURLETTI, T., AND HUITEMA, C. Rtp payload format for h.261 video streams. Tech. rep., Internet Engineering Task-force, 1996. RFC-2032.

75. UYTTERHOEVEN, G., ROOSE, D., AND BULTHEEL, A. Wavelet transforms using the lifting scheme. Tech. rep., Katholieke Universiteit Leuven, April 1997. Department of Computer Science.

76. VALENS, C. *The fast lifting wavelet transform*, 1999. c.valens@mindless.com.

77. WETTERGREEN, D., BAPNA, D., MAIMONE, M., AND THOMAS, G. Developing Nomad for robotic exploration of the Atacama desert. preprint.

78. WETTERGREEN, D., GASKETT, C., AND ZELINSKY, A. Development of a visually-guided autonomous underwater vehicle. preprint.

79. WOO, M., NEIDER, J., AND DAVIS, T. *OpenGL Programming Guide*, 2nd ed. Addison Wesley, 1996.

80. WOODS, A., DOCHERTY, T., AND KOCH, R. Image distortions in stereoscopic video systems. In *Stereoscopic Displays and Applications IV* (1993), SPIE/IS&T, pp. 36–48.

81. WOODS, A., DOCHERTY, T., AND KOCH, R. Experiences of using stereoscopic video with an underwater remote operated vehicle. In *Underwater Intervention* (1994). http://info.curtin.edu.au:8080/∼iwoodsa/ui94.html.

82. YARIV, A. *Optical waves in crystals - propagation and control of laser radiation*. John Wiley and Sons, 1984.

83. YU YEH, Y., AND SILVERSTEIN, L. D. Limits of fusion and depth judgement in stereoscopy color displays. *Human Factors 32* (1990), 45–60.

84. ZWIMPFER, M. *Color Light Sight Sense*. Schiffer Publishing, 1988.