This is to certify that the
thesis entitled

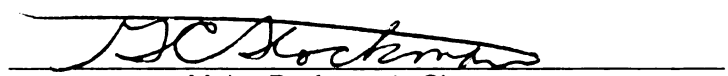
A NON-OBTRUSIVE HEAD MOUNTED FACE CAPTURE SYSTEM

presented by

CHANDAN REDDY

has been accepted towards fulfillment
of the requirements for the

M.S. degree in COMPUTER SCIENCE AND ENGINEERING



Major Professor's Signature

8 August 2003

Date

A Non-obtrusive Head Mounted Face Capture System

By

Chandan Reddy

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Computer Science and Engineering

2003

Abstract

A Non-obtrusive Head Mounted Face Capture System

By

Chandan Reddy

Capturing a face in a multi-user collaborative environment has been a problem of concern for several years. The problem becomes more severe in mobile applications where the capturing system may occlude the user's field of view. In this thesis, a system is proposed that captures two side views of the face simultaneously and generates a frontal view. Applications of facial capture system include tele-conferencing, wearable computing, collaborative work and other mobile applications. The system is designed to produce in real-time a stable, quality frontal view of an HMD user whose face is captured with little obstruction of the field of view. The frontal views are generated by warping and blending of the side views after a calibration step. In tests, the generated views compared well with real video based on both normalized cross correlation and the Euclidean distance between some of the prominent feature points. Preliminary qualitative assessment of these generated views also concludes that the generated video is adequate to support the intended applications. A 3D face model that can support the generation of arbitrary views was also constructed.

To my parents

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Dr. George Stockman, my main advisor for his expert guidance and mentorship. I am grateful to my co-advisor, Dr. Frank Biocca for his moral and financial support throughout my stay at MSU. I would like to express my sincere gratitude to Dr. Jannick Rolland, my external faculty member for her discussions related to optics. I would also like to thank Dr. Charles Owen for providing me some of the hardware components required for this project.

Special thanks to Zena Biocca and Joy Mulvaney for their help during my stay in the MIND Lab. I would like to express my gratitude towards the graduate secretaries Kim Bassa and Linda Moore. Help from my colleagues play a major role in the success of this thesis. In this regard, I would like to thank the students of MIND Lab, MET Lab and PRIP Lab at Michigan State University and the ODA Lab at the University of Central Florida for their help and valuable technical discussions. I would also like to thank my friends Prasanna, Raja, Badri and Shankar for their continuous support. Finally, I thank my family for being with me and supporting me all the time.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

The overall aim of this work is to design an augmented reality based face-to-face teleconferencing system. The main advantages of such a system will be the ability to produce stable video-based images of all the remote participants whose faces are captured without obstructing the users' field of view. This system can be used in the fields of augmented reality, wearable computing, and other mobile applications. This thesis advances the work in face-to-face telecommunication by creating a non-obtrusive real-time Face Capture System (FCS).

1.1 Motivation

One key motivation for the creation of advanced collaborative environments is the increase in computer-mediated communications. Recent concerns over terrorism, highway gridlock and delays at airports dramatically increased the demand for social presence technologies. Consider this evidence of how social disruptions, such as ter-

rorism, increased demand for telepresence technologies. Due to the fear of flying, after the Sept. 11, 2001 attack:

- The use of the collaborative servers increased 300% immediately [Ham01].
- According to CNN's report, "National Business Travel Association showed that 88% of companies planned to increase use of videoconferencing" [Lin01].
- American suppliers reported an increase of 140% in videoconferencing bookings [Bor01].
- Based on a report from British Telecomm, there was an increase of 85% in video conferencing and 30% in audio conferencing in the world.
- A poll by Osterman Research found 60% of business organizations had greater interest in teleconferencing and 41% reported drops in air travel [Ost01].

Although we may have face-to-face interactions with workmates or others, many of the social interactions include an increasing number of purely virtual interactions with others; we rarely or never meet face-to-face. When it comes to communications from remote places, the human face is one of the most important representative parts of the human body because it has great expressive ability that can provide clues to the personality and emotions of a person. The facial expressions of a remote collaborative or a mobile user can convey a sense of urgency, lack of understanding or confidence in action and other non-verbal elements of communications. According to Gary Faigin, "There is no landscape that we know as well as the human face. The

twenty-five-odd square inches containing the features is the most intimately scrutinized piece of territory in existence, examined constantly, and carefully, with far more than an intellectual interest"[Fai90]. With an increase in telecommunication systems, research is being brought in terms of making advanced capture technologies. Even the most advanced teleconferencing and telepresence systems transmit frames of video. These frames are nothing but 2D images. In order to get additional views, the systems are using either a panoramic system and/or interpolate between a set of views [CW93, SD96]. 3D teleconferencing systems are still in the research development stage. Efforts are made in this direction at Michigan State University in collaboration with the University of Central Florida to develop a Tele-collaborative environment called the "Teleportal System".

1.2 Broad Objectives

The overall goal of the teleportal system is to allow multiple users to enter a room-sized display and use a broadband telecommunication link to engage in face-to-face interaction with other remote users in a 3D augmented reality environment. The most fundamental challenge of capture and display technologies is creating a compelling sense of interacting with spaces and people that are not directly present physically in our proximity. Advanced media technologies are designed to give a strong sense of telepresence [OYTY98], defined as the sense of "being there" in a remote location, or social presence[WBL+96], the sense of "being with others" who are not in the same room or place as the user.

The teleportal system will provide remote communication between at least two users. A facial capture system and a projection display provide high quality video of remote participant's faces to the user. The teleportal system will provide a channel for the transmission of the non-verbal cues through an unobtrusive capture of the face. The remote presentation module will display quality frontal views of the remote user in an augmented reality based environment.

A principal feature of this teleportal system is that single or multiple users at a local site and a remote site use a telecommunication link to engage in face-to-face interaction with other users in a 3D augmented reality environment. Each user utilizes a system that includes a display such as a Head Mounted Projection Display (HMPD) and a facial expression video capture system. The video capture system allows the participants to view an image of the face of all remote participants and hear their voices, view the local participants, and view a room that blends physical with virtual objects with which users can interact. The HMPD projects high quality graphics in real-time towards a screen that is covered by a fine grain retro-reflective fabric. The HMPD and video capture system do not occlude vision of the physical environment in which the user is located. This system allows users to see both virtual and physical objects, so that the objects appear to occupy the same space.

1.3 Problem Definition

In the preferred embodiment of the teleportal system, multiple local and remote users can interact in a room-sized space draped in a fine-grained retro-reflective fabric. The Teleportal Face-to-Face System allows individuals to see 3D stereoscopic images of remote participants in an augmented reality environment. The teleportal system will accomplish this through the following key components or sub-systems:

- Face capture system.
- Virtual view synthesis.
- Head Mounted Projection Display.
- 3D virtual and tele-immersive environments.
- High bandwidth network connections.

1.3.1 Face Capture System (FCS)

Capturing a face in a multi-user collaborative environment has been a topic of research for several years. The problem becomes more severe in mobile applications where the capturing system must be designed in such a way that a user's field of view is not occluded. The FCS is responsible for obtaining and transmitting a quality frontal face video of a remote user involved in the communication. The FCS proposed here captures the two side views of the face simultaneously and generates the frontal view. This face capture equipment consists of two miniature video cameras and convex mirrors [BR00]. A flexible bracket attached to the ear rests of the system wraps over

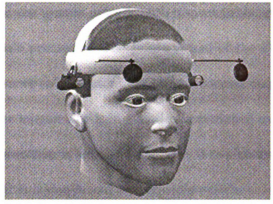


Figure 1.1: Face capture system with two convex mirrors and two lipstick cameras

the top of the users head. Two mirrors are attached to the cap assembly and point down towards the user's face. The two cameras are placed near the ears and they focus the side view of the face images through the convex mirrors.

Figure 1.1 illustrates the face-capture cameras and the mirrors with respect to the user's head. Each of the cameras is pointed towards the respective convex mirror which is angled to reflect an image of one side of the face. The convex mirrors produce a slight distortion of the side view of the face. The left and right video cameras capture the corresponding side views of the human face in real-time. Optics issues concerned with the mirrors and camera lenses are studied in Section 3.2. The side view captured from the cameras will introduce some additional distortion to the images that must be removed in the virtual frontal video.

Advantages

In contrast with the conventional capturing techniques, where either the capture system is static with the environments or the capture system is huge and consumes

enormous horse power, our system is static with respect to the user's head movements and works on any basic processor. The primary goal of the FCS is to increase the telepresence of two remote collaborators. This thesis focuses on providing quality face video in real-time for a human who is in a modestly equipped environment. Since the proposed face capture system is light and portable, it can be used more effectively in mobile applications. It is a simple and user-friendly system that captures the human face in real-time in a mobile environment without obstructing the field of view of the user. A long-term goal of such a system is to achieve a video see through view simultaneously by flipping the mirror and changing the focus of the camera lenses.

Applications

Telephones and teleconferencing systems are widely used communication facilities. High quality video of participants' faces will significantly improve collaborative communication. The current and projected applications of our system will improve the means of communications between two people located at remote places. The main application areas include multi-user collaborative work, mobile environments and teleconferencing. This system can also yield benefits in other areas such as biometrics and e-business. Using the system, a remote expert can view a surgical procedure while observing the surgeon actually performing the procedure, allowing the expert to assess not only the performance, but also how well instructions are being comprehended. For mobile communications, even the latest cell phone technology forces the user to have a small web camera in front of his face and to carry it in his hand showing his face to the camera at all times. This is an over burden to the user and

this is not an effective way of capturing because it might fail in some cases where the user turns around suddenly and when he fails to place the camera in front of him. Our proposed system will be more effective in such an application.

This system can be used effectively in medical and military applications. In these kinds of applications, it is most common to have a system that can capture the person's face as well as the area the person is viewing. If we flip the mirror, the camera can view the area that the person is viewing; so, FCS solves the problem of changing the camera positions every time to view the area of concern. It is clear that the cameras will move with respect to the head motion and we can view the face by just flipping the mirrors.

1.3.2 Virtual View Synthesis

This thesis focuses mainly on vision-based algorithms that are applied to construct the frontal view from the two side views captured by the cameras through the mirrors and also in developing techniques for camera calibration to extract information. A 3D head model of the face is constructed so that the face may be rendered from many frontal view points.

1.3.3 Head Mounted Projection Display

Most of the advanced 3D environments are based on either a "CAVE" based technology [CNSD+92] or head-mounted displays [BC94]. A new type of head-mounted display, which combines the advantages of both technologies, has been designed by

Figure 1.2: A prototype of the Head Mounted Projection Display.

Figure 1.3: Complete integrated system of the head mounted display with the face capture unit.

Rolland et. al [HGGR00]. The HMPD consists of a pair of miniature projection lenses and displays mounted on the helmet and retro-reflective sheeting materials placed strategically in the environment. The novel properties of this technology suggest solutions to some of the problems of state-of-art visualization devices and make it suitable for multiple-user collaborative environments.

1.3.4 3D Virtual and Tele-immersive environments

One essential aspect of the teleportal system is to design a tele-immersive environment that can accommodate spatially matched volumetric datasets. These datasets are transmitted between two (or more) remote sites and are projected by a Head Mounted Projection Display in an Augmented Reality environment. The projection display system requires a retro reflective surface to bounce the projected light directly to the user's eyes. Spatially distributed volumetric datasets such as medical, engineering, and scientific data can be shared across different sites. The retro-reflective material optical properties allow to reflect the light back to its source with little diffusion whatever computer-generated image is being projected onto it. The configuration of the display surfaces is the most flexible part of the system [1]. Display surfaces can be created using forms of the retro reflective material such as wallpaper forms, cloth-like forms, etc. Hence, the surfaces of all types can be prepared as display surfaces for the system.

1.3.5 Broadband Network Connections

The Internet2 test bed has been implemented and tested using MPEG 2 video streams between the MIND Lab at Michigan State University and the ODA Lab at the University of Central Florida. The test bed has been formed to test the reliability and performance for sharing of stereoscopic video signals. These Internet2 [web] connec-

[1] A cylindrical volumetric display has been implemented and tested to project and track a medical dataset of a virtual skeleton. A tabletop display is also tested to project an architectural model of the Beaumont Tower, a landmark of the MSU Campus. Several other displays are being conceptualized to research other volumetric datasets.

tions are capable of transmitting full broadcast quality video streams using MPEG2 video encoding and decoding technology between remote collaborative sites. This Internet2 test bed aims to support real time remote collaboration by allowing the sharing of bi-directional video streams capable of carrying enormous amounts of data that can be used for medical visualization, tele-conferencing, or other applications that make use of large bandwidth data transmission. Two teleportal rooms are being developed, one at Michigan State University and the other at the University of Central Florida. These will be linked using an Internet 2 connection.

1.4 Thesis Contributions

The scientific contributions of this thesis include the

1. Establishment of the basic hardware required for the FCS.

2. Estimation of camera-mirror parameters for the optimal FCS configuration.

3. Algorithms for generating a quality frontal video from two side videos.

4. 3D Face model construction based on two side views.

The overall framework for the teleportal project is to design some innovative advanced 3D interfaces among people sharing a common space as well as people located far apart. The combination of quality facial capture, communications, and graphical display technologies allows for full interaction between remote and local users. The best application where this new technology can be used effectively is a face-to-face distributed collaboration system. There is always an acute need for

more effective collaboration and more effective knowledge sharing systems among geographically scattered people. This thesis forms a stepping-stone for the creation of a complete 3D augmented reality based face-to-face communication system that can produce stereoscopic views of the users via a real-time augmented reality display. Even though the overall framework of this proposed system is to support distributed collaborative work, the major focus of this thesis is towards capturing the human face and providing quality face views in real-time for a human who is in an unobtrusive instrumented environment.

1.5 Organization of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 describes the relevant background for the FCS and the display environments. Chapter 3 describes the system analysis. The experimental setup is detailed and the optics issues that are involved in the configuration of the face capture system are discussed. The algorithms and methods used in the FCS are explained in Chapter 4. Chapter 5 illustrates some results of using the prototype developed. Assessment of the generated frontal videos and discussion of some issues regarding the portability of the system is discussed in Chapter 6. Finally, Chapter 7 presents conclusions of the work with FCS and suggests ideas for future work.

Chapter 2

Relevant Background

This chapter describes the related work that has been done in the sub parts of the FCS. Section 2.1 describes the existing face capture systems and explains the different ways of capturing a human face using the state-of-art devices. Background for synthesizing novel views without reconstructing the complete 3D model is discussed in Section 2.2. Section 2.3 describes some of the traditional methods for depth extraction and face modeling. Section 2.4 describes the relevant work that was done in the areas related to HMDs and tele-immersive environments

2.1 Face Capture Systems

Extensive research work has been done in the areas of face modeling [PW96, DMS98, GGW+98] and face recognition [BP93]. However, capturing a human face in real-time has been a topic of less concern when compared to other areas of face analysis. Face capture can be done in a collaborative environment using a sea of cameras in a highly

equipped environment [FBA+94]. Recently the National Tele-immersion project used many cameras that are statically placed in the environment to develop the 3D head model of a user in a collaborative scenario (see Figure 2.1). Sometimes even such a highly calibrated environment might fail to produce quality frontal views mainly because of the orientation of the head.

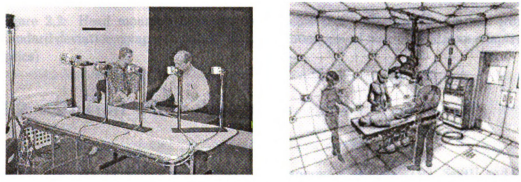


Figure 2.1: A collaborative room with many cameras. (Left:National tele-immersion project. Right: Sea of cameras)

In a mobile scenario, it will be quite difficult to capture the human face, even though some solutions exist (see Figure 2.2). The user wears a head mounted setup and hence the capturing device is assumed to be static with the head movements. The main problem that arises by using the head mounted face capture system shown in Figure 2.2 is that it obstructs the field of view of the user who is wearing the head mounted set. These devices are used mostly for character animation and are not being used for day-to-day tele-collaboration scenarios. In this thesis, the design and construction of *"a non-obtrusive head mounted face capture system"* is described.

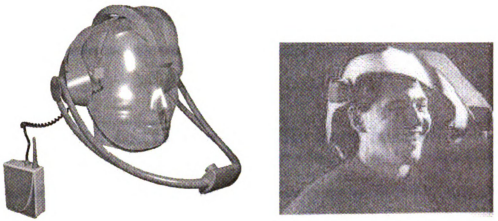


Figure 2.2: Head mounted face capture systems (Left: Facecap3D product from standard deviation company. Right: Head mounted optical face tracker from adaptive optics)

2.2 Virtual View Synthesis

Novel views can be synthesized either by a panoramic system[ZT92, Sei01] and/or by interpolating between a set of views [CW93, SD96]. Using image-based rendering techniques, one need not explicitly derive the 3D information of an object [GGSC96]. These techniques concentrate on rendering novel views without actually reconstructing the 3D structure. In recent years, the techniques in computer graphics and computer vision are being combined to produce interesting results [Len98].

Producing novel views in a dynamic scenario was successfully shown for a highly rigid motion [MD99]. These techniques extend the interpolation techniques to the temporal domain and do not strictly restrict them to the spatial domain. A novel view at a new time period was generated by interpolating views at nearby time intervals using spatio-temporal view interpolation[VBK02], where a dynamic 3-D scene is modelled and novel views are generated at intermediate time intervals.

In this thesis, we study, develop, and test techniques for capturing the human face and reconstructing novel views in a real-time video sequence.

2.3 Depth Extraction and 3D Face Modeling

Structured light is a commonly used technique in computer vision [DT96, PGD98]. Grid patterns have been used successfully in the past for reconstruction and pose estimation of 3D surfaces [HS89, SS89]. Color coding of a structured pattern was first described in[BK87]. Structure from motion [TK92] estimates the structure of a 3D object from an image sequence. The binocular stereo method [BF82, OK93] is the most commonly used method for estimating the 3D coordinates of points on the object and the depth is determined using two images taken by two cameras from different angles. For depth estimation, even though a silicon range finder measures 3D coordinates quickly, it requires expensive hardware for scanning purposes [Sat94]. The shape from shading method [HB86] and photometric stereo method [Woo80] can be used for measuring the normal vectors of objects. These measurements are based on photometric properties. Shape from shading techniques use a single camera and two or more images are taken of the face in a fixed position but under different lighting conditions. Even though these methods can provide accurate 3D information about complicated shapes, they are not recommended for our application because the accuracy of the measured 3D data is highly influenced by the various photometric properties. In our case, the traditional stereo method between the two cameras will not work because of the occlusion of the facial features. A Light projection method

has also been used for face shape estimation[SN97] by stereo vision algorithm. In this technique, multiple light stripes are projected onto the face using the slit pattern projector [Jar83, RK75].

A 3D individualized head model can be constructed from two orthogonal views [IY96]. A three-dimensional face model has also been created from a video sequence of face images[LC01]. Realistic expressions have been synthesized from photographs [PHL+98]. Various vision techniques have been effectively used for visualizing the 3D world in a collaborative application [XLH02].

2.4 Head Mounted Displays and Tele-immersion

The main aim of 3D visualization devices is to visualize computer-generated objects and make them appear as real objects. Interest in 3D visualization devices has endured and permeated various virtual and augmented reality domains. The first attempt to create a virtual reality (VR) environment was the production of an augmented reality (AR) navigational aid for helicopter pilots [Sut65]. Since the first head-mounted display (HMD) originated by Ivan Sutherland in the 1960's [Sut65], 3D visualization devices most commonly used in virtual and augmented reality domains have evolved into three typical formats: standard monitors accompanied with shutter glasses, head-mounted displays (HMDs), and projection-based displays. HMDs provide a fine balance of affordability and unique capabilities such as spanning the continuum proposed by [MK94] from reality, via mixed reality, to immersive environments, creating mobile displays [Fei02] and enabling teleportal capability with

face-to-face interaction [BR00]. There are two major categories of HMDs: immersive and see-through [BC94]. Immersive HMDs present a user with a view that is under full control of computers at the expense of the physical view. These systems require a virtual representation of a user's hand to manipulate the virtual world and avatars of collaborative team members in multi-user environments [BF98]. See-through HMDs superimpose virtual objects on an existing scene to enhance rather than replace the real scene. Video and optical fusion are two basic approaches to combine real and virtual images. The main trade-offs include the resolution, the field of view (FOV), the presence of large distortion for wide FOV designs, the inaccurate eye point representation, the conflict of accommodation and convergence and the occlusion contradiction between virtual and real objects.

The VIDEOPLACE system used vision algorithms to track users within the environment to generate visual and auditory responses [KGH85]. The most popularly used 3D teleconferencing system is the CAVES system. It uses multiple screens arranged in a room configuration to display virtual information. It is often implemented as a cube of approximately 12 feet on each side [CNSD+92] and it uses four CRT projection systems and crystal shutter glasses. As a viewer moves within its display boundaries, the correct perspective and stereo projections of the environment are updated, and the image moves with the viewer surrounding him. In CAVES systems, there is only one correct viewpoint, all other local users have a distorted perspective on the virtual scene. Scenes in the CAVES are only projected onto a wall. So two local users can view a scene on the wall, but an object cannot be presented in the space between users. These systems also use multiple rear screen projectors, and therefore are bulky

and expensive.

An alternative to a cave is to create an immersive virtual environment through a head-mounted display (HMD). While conventional types of head-mounted displays employ eyepiece optics to create the virtual images [RF00], an emerging technology known as a head-mounted projection display (HMPD) has fairly recently been demonstrated to yield 3D visualization capability [KO97, PR98]. The concept of head-mounted projection displays (HMPDs) was initially patented by Fisher in 1996 [Fis96] and was proposed as an alternative to remote displays, head-mounted displays and stereo projection systems for 3D visualization applications. Potentially, the HMPD concept provides solutions to some of the problems existing in state-of-art visualization devices. FCS will be integrated with HMPD to provide a tele-immersive environment.

A thorough discussion of traditional video mediated communication is described in [FSW97]. A number of teleconferencing technologies support collaborative virtual environments that allow interaction between individuals in local and remote sites. For example, video-teleconferencing systems use simple video screens and wide screen displays to allow interaction between individuals in local and remote sites. However, wide screen displays are disadvantageous because virtual 3D objects presented on the screen are not blended into the environment of the room of the users. A mixed reality computer supported collaborative environments, which enable transitions along the virtuality continuum was first illustrated in the Magic Book [BKP01]. The Magic Book provides an experience where users have the capability and the incentive to travel from real to fully immersive environments within the same application. A

mean to travel along the virtuality continuum is also discussed in [DRHL+03].

The most fundamental challenge of face capture and display technology is to create a compelling sense of interacting with environment and remote people who are not directly present physically in our proximity. Quality presentation of the remote users and an effective user interface is critical for all media communications. The combination of quality facial capture, high bandwidth network communication channel, and graphical display technology will provide a complete neat interaction between local and remote participants thus emphasizing the concept of "*people need not physically be there*".

Chapter 3

System Overview

This chapter describes the equipment required by the FCS designed and details the design issues concerned with the cameras and mirrors that are used. Section 3.3 explains the experimental setup and the procedure followed during the face capture.

3.1 Equipment required

The FCS requires some hardware, software and network connections. The network connections are required to transmit the videos to remote locations.

3.1.1 Hardware

Our prototype uses an Intel Pentium III processor running at 746 MHz with 384 MB RAM. It is installed with two Matrox Meteor II standard cards [1]. These cards are

[1] Matrox Meteor II Multi Channel can capture multiple video streams simultaneously in real-time. However, this can be done only using monochrome videos. Hence, two Matrox Meteor II standard cards were chosen for implementing FCS.

connected to the control units of the lipstick cameras through a general cable. The camera that is used is a Sony DXC LS1 NTSC camera with 12 mm focal length lenses. We use Matrox Meteor II Standard that supports both multiple composite and s-video inputs. The video is digitized by a Matrox Meteor II standard capture card, yielding interlaced 320 X 240 video fields at 60 Hz. During the off-line calibration stage, the system also used an Infocus LP350 projector to project a grid onto the user's face. A calibration sphere is used in the process of extracting the depth information. Voice is recorded in the same system using a microphone.

3.1.2 Software

The API required to do the programming for controlling this hardware is MIL-LITE 7.0[2]. The standard Windows based sound recording software is used to record the voice of the user during the conversation. The sound file is appended to the .avi file using Adobe Premiere 6.0, which is a popularly used video editing software.
Using this hardware and software, two videos are captured simultaneously at the rate of 30 frames per second.

3.1.3 Network Connections

The Internet2 testbed has been implemented and tested encoding MPEG 2 video streams at 3 Mbps and decoding video streams at 4 Mbps between the MIND Lab in

[2]MIL is another API provided by MAtrox. The main difference between MIL and MIL-LITE is that MIL is used for high-level programming. In our application, we need to read each frame and applu our own algorithms. This is treated as low-level programming in which we need to have the control of the data at each frame. Hence, MIL-LITE 7.0 is appropriate for our application.

Michigan State University and the ODA Lab in the University of Central Florida.

3.2 Optics Design Issues

3.2.1 General System Layout

The general layout of the system is shown in Figure 3.1. The calculations for

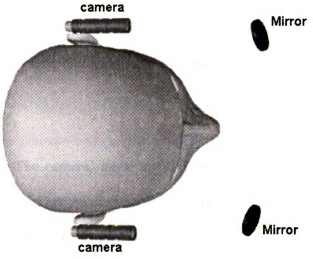


Figure 3.1: General layout of the face capture system.

estimating the variable parameters are simplified by unfolding the overall system (see Figure 3.2). When the system is unfolded, the mirror can be represented as a negative lens. The face is being imaged through the mirror and the camera is actually focusing on the face through the mirror. We shall now consider various parameters for each of the components that are involved in the system. The various components of this system will be the (a) human face, (b) camera and (c) mirror.

(a) Human face: The main parameters of the face that will affect the geometry

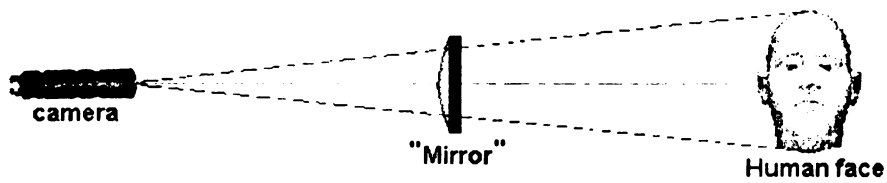


Figure 3.2: Unfolded layout of the face capture system

of the system will be the height and the width of the face. Even though some other factors such as the skin color and illumination might affect the performance of the system, they will not have any effect on the geometry of the system.

(b) Camera: The camera, made up of a sensor and a focusing lens, requires careful study. The various parameters that significantly affect the geometry of the system are the sensing area, the field of view, the pixel dimensions, the focal length, the *f*-number or lens diameter, the minimum working distance and, the depth of field.

(c) Mirror: This is the most flexible component of the system. Hence, all the parameters of this component are estimated and the component is manufactured based on the estimated values of the parameters. The various parameters of this component are the focal length, the *f*-number or mirror diameter, the radius of curvature, and the magnification factor.

3.2.2 Specification Parameters

The various parameters that are involved in the calculations are as follows.

(a) Human face: Without loss of generality, let us assume that we can take the dimensions of an average face for further calculations.

- H - Height of the head to be captured ($\approx 250mm$)
- W - Width of the head to be capture ($\approx 175mm$)

(b) Camera: There are several parameters that significantly affect the geometry of the system. For this application, the main parameters that are to be taken into consideration are the miniature size, the lightweight, the minimum working distance, and the field of view. Based on the approximate values of these parameters, we have obtained the off-the-shelf lipstick camera, SONY DXC LS1. For this camera, two focal length lenses are available, one is of 4mm and the other is 12 mm lens. Because of the wide field of view (45^o X 33^o), the 4mm lens is not well suited for our application when compared to the 12mm lens. The parameters of camera 12mm lens are the

- Sensing area: The sensing area is 1/4", or equivalently 3.2 mm(y) X 2.4 mm(x).
- Pixel Dimensions: According to the specifications, the image sensed has a resolution of 768 X 494. However, when this image is captured using a Matrox

Meteor II standard frame grabbing card, the image is digitized into 320 X 240. For any further evaluations of other parameters (e.g. depth of field) the resolution of the image is considered to be 320 X 240. Even though higher resolution images (640 X 480) can be captured, restrictions of the RAM size force us to capture low resolution images.

- Focal Length(F_c): The focal length of the lens that was selected is 12 mm (VCL - 12UVM).

- Field of View (FOV): The field of view of the camera with the above mentioned lens is $15.2^o X 11.4^o$.

- Diameter (D_c): The diameter of the lens and the camera is 12mm.

- *f*-number (N_c): The *f*-number for this camera lens is 1. While in practice, we will adjust the iris to satisfy optimum illumination of the face provided external room illumination. We shall consider a *f*-number of 1 in the estimation of the other parameters.

- Minimum Working Distance (MWD): The minimum working distance for the selected lens is 200 mm.

- Depth of Field (DOF): This parameter is dependent on all the above mentioned parameter values. This helps in making the system portable. If the system has large depth of field then it will be more portable and can accommodate many users without much changes in the position and focus of the cameras.

(c) Mirror: This part of the system can be customized. The various parameters of the mirror that will affect the geometry of the system are the

- Diameter (D_m) / f-number (N_m)
- Focal Length (F_m) or equivalently the Radius of Curvature (R_m)
- Magnification Factor (M_m)

(d) Distances: There are basically two distances that can be adjusted within a visible range. From Figure 3.1, we can approximate the following distances

- D_{cm} - Distance between the camera and the mirror ($\approx 150mm$)
- D_{mf} - Distance between the mirror and the face. ($\approx 200mm$)

3.2.3 Estimation of the Variable Parameters D_{mf} and D_m

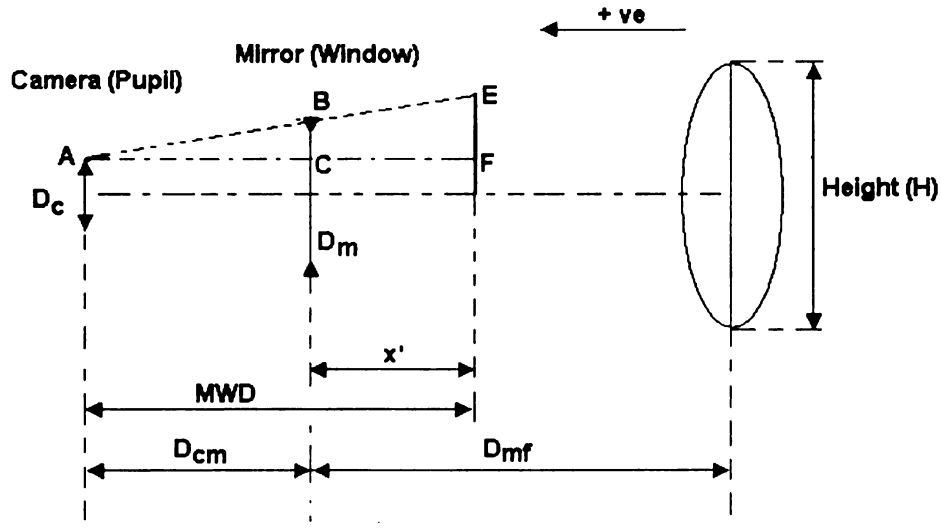


Figure 3.3: Estimation of the variable parameters D_{mf} and D_m

From the theory of pupils and windows [Goo95], the camera is the limiting aperture from the intermediary image plane located behind the mirror. Hence, the camera acts as the pupil of the system and the mirror is the window.

In the unfolded configuration, the mirror is represented as a negative lens with image focal length f'_m equal in magnitude to that of the mirror with an opposite sign. The imaging equation [MM97] for the equivalent lens to the mirror yields

$$\frac{1}{x'} = \frac{1}{D_{mf}} + \frac{1}{f'_m} \tag{3.1}$$

where x' is negative because the values D_{mf} and f'_m are negative. Hence, the image in the unfolded case is virtual and thus it is always between the lens and the human face. In the case of the mirror, the image will be optically located behind the mirror.

Let the FOV of the lens be $\theta_y \times \theta_z$. To maximize the FOV of the capture, we shall image the height (H) onto θ_y and the width (W) onto θ_z. Let H' be the maximum size of the intermediary image formed through the mirror. H' is given by

$$H' = 2 \times tan(\theta_y/2) \times MWD$$

Also, the magnification factor of the mirror is given by

$$M_{mirror} = \frac{H'}{H}$$

and,

$$\frac{x'}{D_{mf}} = M_{mirror} \tag{3.2}$$

Substituting Equation 3.2 in Equation 3.1, we get,

$$D_{mf} = (\frac{1}{M_{mirror}} - 1) * f'_m \tag{3.3}$$

Also, from the definition of f-number,

$$f'_m = N_m \times D_m$$

Based on the similar triangles ABC and AEF shown in Figure 3.3, we can write

$$\frac{D_m/2 - D_c/2}{MWD + x'} = \frac{H'/2 - D_c/2}{MWD}$$

where, D_m must be written as a function of N.

The FOV of the lens is approximately $15^o \times 11.5^o$. Hence

$$M_{mirror} = 52.6 \ / \ 250 = 0.21$$

Taking MWD equal to 200mm and D_c equal to 12mm, D_m can be written in terms of N as shown below

$$D_m = \frac{26.3 \times 2}{(1 + 0.16 \times N)}$$

Table 3.1 presents a summary of estimated values for D_m as a function of the f-number.

A similar computation for a 4mm focal length lens, which was our other off-the-shelf option for the camera considered, is summarized in Table 3.2.

Table 3.1: Estimated values of the variable parameters obtained by varying the *f*-number(Fc =12mm). All dimensions are in mm.

f-number(N_m)	Diameter of the mirror (D_m)	Distance from Mirror to Face (D_{mf})	X'	Focal Length of the Mirror (F_m)	Radius of Curvature of the Mirror (R_m)
1	45.34	-170	-35.80	-45.34	-90.69
1.5	42.42	-238.6	-50.23	-63.63	-127.26
2	39.85	-298.9	-62.92	-79.70	-159.39
2.5	37.57	-352.2	-74.15	-93.93	-187.86
3	35.54	-399.8	-84.17	-106.62	-213.24
3.5	33.72	-442.5	-93.17	-118.01	-236.03
4	32.07	-481.1	-101.28	-128.29	-256.59
4.5	30.58	-516.1	-108.64	-137.62	-275.23
5	29.22	-547.9	-115.35	-146.11	-292.22
5.5	27.98	-577.1	-121.49	-153.88	-307.77
6	26.84	-603.8	-127.12	-161.02	-322.04
6.5	25.78	-628.5	-132.31	-167.60	-335.20
7	24.81	-651.3	-137.12	-173.68	-347.36

Table 3.2: Estimated values of the variable parameters obtained by varying the *f*-number(Fc =4mm). All dimensions are in mm.

f-number(N_m)	Diameter of the mirror (D_m)	Distance from Mirror to Face (D_{mf})	X'	Focal Length of the Mirror (F_m)	Radius of Curvature of the Mirror (R_m)
1	29.59	-139.1	-24.40	-29.59	-59.19
1.5	25.47	-179.5	-31.50	-38.20	-76.40
2	22.35	-210.1	-36.85	-44.69	-89.39
2.5	19.91	-233.9	-41.04	-49.77	-99.55
3	17.95	-253.1	-44.40	-53.85	-107.70
3.5	16.34	-268.8	-47.17	-57.20	-114.40
4	15.00	-282	-49.47	-60.00	-120.00
4.5	13.86	-293.2	-51.43	-62.37	-124.75
5	12.88	-302.7	-53.11	-64.41	-128.82

Based on the practical values that are optimal for the size of the mirror(D_m) and the distances (D_{mf} and D_{cm}), the third row (corresponding to f-number = 2) in Table 3.1 represents the most suitable values for the parameters. The parameters of the mirror are customized using these values.

3.2.4 Customization of the Cameras and Mirrors

The mirrors were manufactured according to the specification table for the 12 mm focal length camera, and available off-the-shelf components. A convex mirror of radius of curvature 155.04 mm was selected, corresponding to a f-number of 2. The convex side of the mirror was coated for the visible light spectrum.

3.3 Experimental Prototype

Even though the overall goal of the FCS is to achieve quality frontal views from cameras and mirrors placed on a headset, as a start, we have simplified the problem into environment based face capture. The general problem is *to generate a virtual frontal view from two side views.* Hence, in our further discussions, we will have two kinds of Face Capture Systems

- Environment Static Face Capture System (ESFCS)
- Head Mounted Face Capture System (HMFCS)

We used a structured light approach to synthesize novel frontal views. The basic idea is to project a structured pattern onto a human face and capture the corre-

sponding grid points on the face from the side views. Based on the distortions of the grid pattern on the face, transform functions are generated to reconstruct the virtual frontal view.

In order to generate the 3D model, a stereo method is applied. According to the design of our system, there is not much overlap of the face region in the two side views. Hence, it is not possible to have stereo between the two cameras. Hence, stereo computation is made between a projector (which is used to project the grid) and each of the cameras. We shall now discuss the procedure of the overall experiment.

The experimental procedure that is followed during the face capture using environment static cameras is discussed in Section 3.3.1. Section 3.3.2 discusses the issues regarding the design of the HMFCS.

3.3.1 Environment-static Camera Face Capture

The experimental bench shown in Figure 3.4 is designed to maintain the same virtual geometry between the subject's head and the cameras/projector. This prototype fixes FCS relative to the user's head but is not a head mounted set. Hence, it has the basic requirement that the user should not move his head. Once, the algorithms are tested thoroughly, we will apply them to a head mounted FCS.

Experimental Procedure

1. The user is asked to sit in a chair. The chair is adjusted according to the convenience of the user and workspace of the cameras.

2. The cameras are focused on the face of the user.

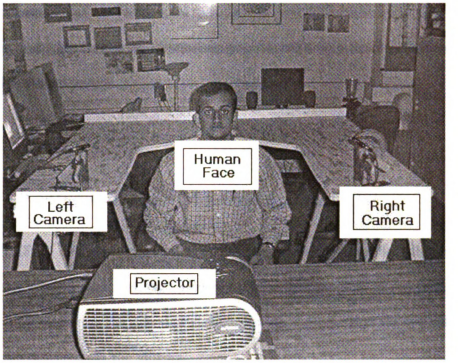


Figure 3.4: Experimental bench prototype of the FCS

3. A microphone is used to record the user's voice.

4. The projector is switched on and a grid is projected onto the face. The grid is projected in such a way that a vertical line passes through the center of the face and bisects the face into two halves.

5. After these initial setup procedures, a "start" command is issued and the cameras start recording the user's face.

7. After 1-2 seconds, the projector is switched off and the grid is no longer projected onto the face.

8. The user is asked to say "The quick brown fox jumped over the lazy dog" continuously for 10 seconds.

9. For the evaluation of the results, a camcorder records the entire user's face during

the experiment.

10. The user's head must be as static as possible with only facial movements and without any head motion.

3.3.2 Porting to a Head Mounted System

In the HMFCS, delicate mirrors off to the side of the user's face reflect the side face image to a camera near each ear. However, the overall problem to reconstruct the novel frontal view from the side view remains the same except that the two side view images are captured via the mirrors.

Placing the cameras and mirrors in appropriate positions is crucial to obtain quality results. Since, this prototype is still in development, it is important for us to have flexibility in the design of the system. Placement of the mirrors and the cameras must be flexible enough to capture quality side views.

Flexible design can be achieved by translation and rotation of the cameras and the mirrors. Basically, this design helps us to estimate effective positions of the cameras and mirrors. Ideally, one would like to have a system that can be used by a wide range of human head sizes. Placement of these cameras and mirrors will significantly affect the quality of the results that are obtained. Most importantly, features of the face must be viewed completely through the mirrors and should be in focus.

The experimental procedure followed to capture the face remains similar to the procedure discussed in the previous section. The main advantage is the freedom

of head motion for the user wearing HMFCS. In both cases, the calibration of the projector and the cameras has to be done only for one frame in the video sequence. Details of the calibration procedure will be discussed in Section 4.2.2.

Chapter 4

Methodology

This chapter explains the methods that are used during calibration and generation of the virtual frontal views and the 3D face model. Section 4.1 describes the overall system and introduces the notations and the coordinate systems that are being used in the rest of the chapter. Steps followed during the calibration procedure are discussed in Section 4.2. Synthesizing virtual videos primarily consists of two phases, namely (1) calibration phase (discussed in Section 4.2.2) and (2) operational phase (discussed in Section 4.3). Similarly, the steps required during the calibration for the 3D face model are discussed in Section 4.2.3 and Section 4.4 describes the various steps followed during the generation of a texture mapped 3D face model. Various implementation issues, including the details about the video capture and processing, are discussed in Section 4.5.

4.1 Description

We shall now discuss the notations that are followed in our algorithms.

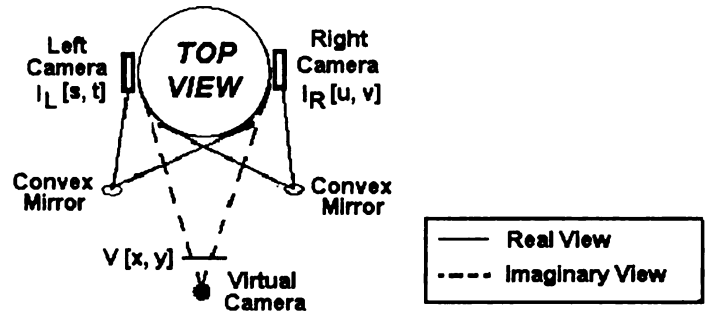


Figure 4.1: Top view of the face capture system

There are basically five coordinate systems that are involved in our system:

1. World Coordinate System (WCS)
2. Head Coordinate System (HCS)
3. Left Camera Coordinate System (LCS)
4. Right Camera Coordinate System (RCS)
5. Projector Coordinate System (PCS)

WCS and HCS are 3D coordinate systems and the rest are 2D coordinate systems. Generating the 2D virtual frontal video, mapping is done between 2D coordinate systems. In this case, the 3D coordinate systems are not used. However, to construct the 3D face model, the mapping is done with respect to the WCS and HCS as shown in later sections. The origin and the coordinate axes in the HCS are shown in Figure 4.9. The coordinate axes in the WCS are aligned in the same manner as the HCS.

The origin of HCS is defined to be the center of the calibration sphere shown in Figure 4.5.

From Figure 4.1, we can see that there are two real images (side views) and a virtual image V has to be generated. The coordinates in these images are described as follows:

V[x,y] - Virtual Image with x,y coordinates (defined in PCS)

$I_L[s,t]$ - Left Image with s,t coordinates (defined in LCS)

$I_R[u,v]$ - Right Image with u,v coordinates (defined in RCS)

4.2 Off-line Calibration

Section 4.2.1 deals with the color balancing technique that is used in our application. The calibration procedure for synthesizing the virtual video is described in Section 4.2.2. Section 4.2.3 discusses the calibration procedures required for the generation of a texture mapped 3D face model.

4.2.1 Color Balancing

Before calibrating the cameras and the projector geometrically, one has to make sure that the cameras are color balanced. Even though, several software based approaches for color balancing can be taken, the color balancing in our work is done at the hardware level. Before the cameras are used for calibration, they are balanced using the white balancing technique. A single white paper is shown to both cameras and

cameras are white balanced instantly. This hardware solution is more reliable than a software based solution. This solution has an upper hand over the software based approach in that it will be much faster, it can handle varying lighting conditions in a more effective manner, nothing is assumed about the color distribution, and this method will give more natural colors. The software based approach might reveal that the virtual frontal view has been synthesized when it is used in varying lighting conditions because some pre-knowledge about the color of the images might have to be stored if the software based approach is used. When the lighting is changed the color balancing will not be handled more effectively [YLW98].

The main reasons for the change in the skin color are

- variation in the lighting condition
- change in the input video camera
- change in the white-balance of the camera

In our case the video cameras remain the same throughout the process and the white balance of the camera is not changed at all. Hence, the only variation is in the lighting conditions. In such a case, if we have some skin color model predefined, it will become problematic. Hence, a hardware based solution is more reliable than a software based solution.

4.2.2 Calibration for Virtual Video Synthesis

During the calibration phase, the transformation tables are generated using the grid pattern coordinates. To get the transformation, a rectangular grid is projected onto

the face and the two side views are captured as shown in Figure 4.2. To generate the virtual video, the cameras and the projector are to be calibrated relative to each other. In essence, the transformation has to be done between PCS and LCS and between PCS and RCS. Since the 3D information is not required to generate the novel frontal view, this calibration will not consider the HCS and WCS. The grid will enable the transformation of corresponding points between the coordinate systems mentioned above. Using these transformation tables, one can map every pixel in the front view to the side view.

Figure 4.2: Demonstration of the behaviour of the grid pattern

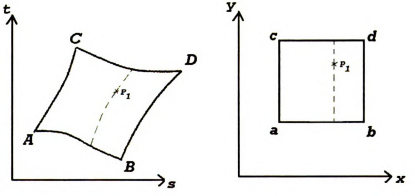


Figure 4.3: Illustration of bilinear interpolation technique.

The behaviour of a single gridded cell in the original side view and the virtual

frontal view is demonstrated in Figure 4.2. A grid cell in the frontal image will map to a quadrilateral with curved edges in the side image. Bilinear interpolation is used to reconstruct the original frontal grid pattern. The bilinear interpolation (see Figure 4.3) is the traditional way of warping a quadrilateral into a square or a rectangle. The number of pixels in the side image in a single grid cell might be less, equal or more than the number of pixels in the corresponding frontal grid cell. This distortion is because of the 3D shape of the human face.

$$s = f_l(x, y) \text{ and } t = g_l(x, y) \tag{4.1}$$

$$u = f_r(x, y) \text{ and } v = g_r(x, y) \tag{4.2}$$

Equations 4.1 and 4.2 represent four functions that are to be determined during the calibration stage (off-line). Once the four functions are obtained the transformation tables can be generated. These transformation tables will be used in the operational stage described in 4.3. A calibration procedure is used to define the mapping of the grid cells.

The procedure followed during the calibration is as follows:

1. Capture the two side views (with a grid projected on the face) from the two cameras and store them in the corresponding images ($I_L[s,t]$ and $I_R[u,v]$) .

2. Take some grid intersection points and define transform functions for determining the (s,t) coordinates in the left image (I_L) and (u,v) coordinates in the right image (I_R)(see equations 4.1 and 4.2) .

3. Apply bilinear interpolation to map any points inside the grid coordinates.

4. To implement the transformation functions, construct two transformation tables (one for the left image and one for the right) which have index as (x,y) and gives a corresponding (s,t) of I_L and (u,v) of I_R.

5. These transformation tables define the mapping ($M_p[x, y]$) of each frontal pixel in the virtual view to the corresponding pixel in the side views (I_L or I_R).

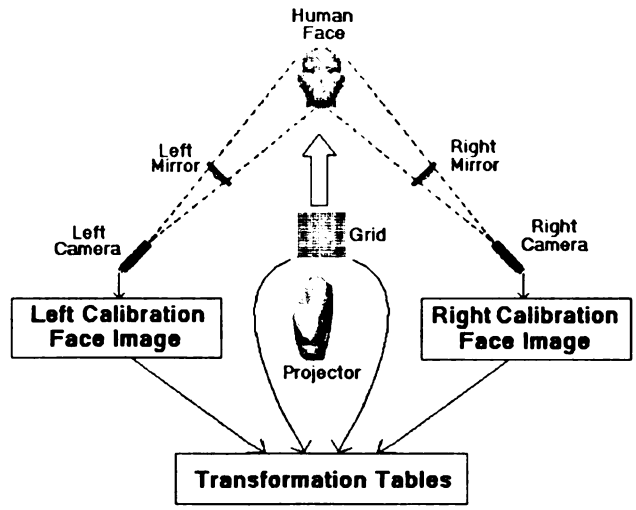


Figure 4.4: The off-line calibration stage during the synthesis of the virtual frontal view.

As shown in Figure 4.4, a projector is used to project a grid pattern onto the human face. The face with gridded pattern is captured from the two side cameras. Using the coordinates of the original grid pattern and the corresponding coordinates in the side views, transformation tables are generated.

4.2.3 Calibration for 3D Face Modeling

For generating the 3D model of the face, some depth information of prominent facial features is to be estimated. The technique that is used here for depth extraction is

structure from stereo. For extracting the depth information, the three components (namely LCS, RCS and PCS) have to be calibrated with respect to the HCS. A calibration sphere (see Figure 4.5) is used for calibrating the system. The origin of this calibration sphere is considered to be the origin of the WCS. A detailed discussion about the conversion of the spherical coordinates to cartesian coordinates is given in appendix A.

Camera Calibration

In order to estimate the depth information, the system has to be calibrated with respect to the WCS. To calibrate the system, a calibration sphere is used. The origin of this calibration sphere is the origin of the WCS. The depth values are first estimated in the WCS and then they will be transformed into the HCS. There are 17 calibration points (A-Q) shown in Figure 4.5

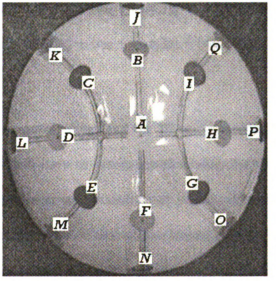


Figure 4.5: The calibration sphere with labeled calibration points

In the Figure 4.5, the calibration points are chosen in such a way that the azimuthal angle is varied in steps of 30^o and the polar angle is varied in steps of 45^o.

The equations for camera calibration [SS01] are explained below.

$$\begin{bmatrix} sL_{Pr} \\ sL_{Pc} \\ s \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & 1 \end{bmatrix} \begin{bmatrix} W_{Px} \\ W_{Py} \\ W_{Pz} \\ 1 \end{bmatrix} \tag{4.3}$$

Eliminating the homogeneous coordinate s from the Equation 4.6, we get

$$u_j = (c_{11} - c_{31}u_j)x_j + (c_{12} - c_{32}u_j)y_j + (c_{13} - c_{33}u_j)z_j + c_{14} \tag{4.4}$$

$$v_j = (c_{21} - c_{31}v_j)x_j + (c_{22} - c_{32}v_j)y_j + (c_{23} - c_{33}v_j)z_j + c_{24} \tag{4.5}$$

During the calibration, the 2D image and the 3D world coordinates of the calibration points are given. We will have to determine the values for the calibration matrix. Hence, we will have two linear equations for each of the calibration points. Equations 4.4 and 4.5 can be rearranged such that all the known values are placed one side and unknown values on the other. The new representation of Equations 4.4 and 4.5 is shown in Equation 4.6

$$\begin{bmatrix} x_j & y_j & z_j & 1 & 0 & 0 & 0 & 0 & -x_j u_j & -y_j u_j & -z_j u_j \\ 0 & 0 & 0 & 0 & x_j & y_j & z_j & 1 & -x_j v_j & -y_j v_j & -z_j v_j \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{12} \\ c_{13} \\ c_{14} \\ c_{21} \\ c_{22} \\ c_{23} \\ c_{24} \\ c_{31} \\ c_{32} \\ c_{33} \end{bmatrix} = \begin{bmatrix} u_j \\ v_j \end{bmatrix} \tag{4.6}$$

In the above matrix representation, all the entities on the left are known from the calibration tuples, while the calibration matrix (c_{ij}) values are unknown. Using n calibration points, we can obtain 2n linear equations. In our case, we need to have at least 6 calibration points to estimate the values of the 11 unknown calibration parameters. As discussed earlier, we have 12 points for calibrating the cameras and 17 points for calibrating the projector. In both of the cases, we will have more than 10 calibration points, so this is an over determined system. This system can be solved using the least squares approach.

Calibration of the Projector

The calibration of the projector is done in the same way as that of the cameras. The basic difference is that the image coordinates of the calibration points on the sphere are not obtained directly from the 2D image. Instead, the image coordinates are obtained by projecting a "blank image" onto the calibration sphere. All the labeled points on the calibration sphere will be "seen" from the projector. The points in

the projected image of all these points are noted by clicking on the 2D-screen image coordinates in the PCS while it is projected onto the calibration sphere in WCS.

4.3 Virtual Video Synthesis

The transformation tables that are generated in the off-line calibration phase are used in the operational phase to generate each virtual frontal frame in the video. The algorithm is described as follows:

1. Get the two side views without a grid projected on the face from the two cameras(I_L and I_R).

2. Reconstruct each (x,y) coordinate in the virtual view by accessing the corresponding location in the transformation table and retrieve the pixel in I_L (or I_R) using the mapping ($M_p[x, y]$).

3. Smooth the geometrical and lighting variations across the vertical midline in V by applying a linear (one-dimensional) filter.

4. Continue this reconstruction of V[x,y] for every frame of the videos to produce the final virtual frontal video.

Figure 4.6 shows the complete block diagram of the operational phase. The operational stage can be split into mainly three steps:

- Face Warping
- Face Mosaicking
- Post-processing

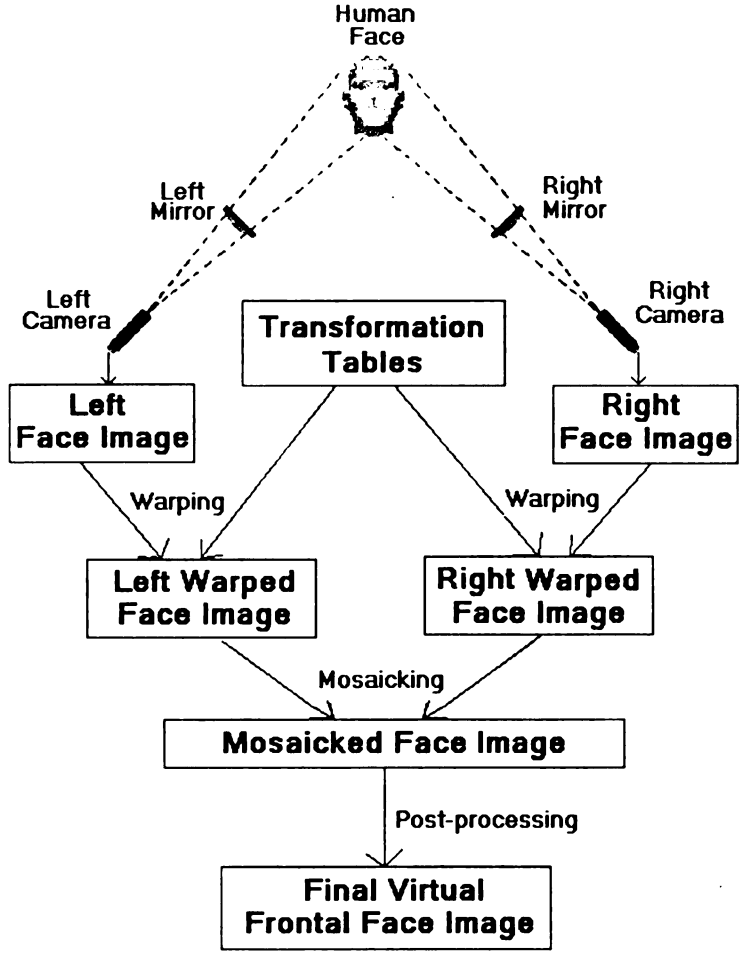


Figure 4.6: Operational stage during the synthesis of the virtual frontal view

4.3.1 Face Warping

Each grid in the side views is warped into the frontal rectangular grid. The transformation tables will define the mapping that is required at the pixel level. These tables are responsible to generate each frontal pixel in the virtual view. Since the transformation is based on the bilinear interpolation technique, each pixel can be generated only when it is inside four grid coordinate points whose transformation is defined by the transformation tables. This is the main reason for our algorithm being unable to generate the ears and and hair portion of the face.

4.3.2 Face Mosaicking

The two warped side views are placed adjacent to each other. The reference points will be the horizontal intersection points that are present on the vertical line passing through the center of the face, thus bisecting the face into two halves. During the calibration stage, if the transformation tables are not created with the vertical line passing through the center of the face, it will significantly affect the mosaicking stage. After these two views are placed side-by-side, a smoothing algorithm is applied at the edge. This is important because this smoothing algorithm will smooth any variation in intensity and geometry. The geometrical variations, especially in the lip region, will become more pronounced when the person speaks fast with large lip movements.

4.3.3 Post-Processing

This is the video editing stage. The video that is obtained will contain the grid pattern in the first few frames. After the grid pattern is stopped there will be a color transform of the skin. These frames with the gridded pattern are to be deleted from the final output. A microphone records the voice of the user and is stored in a separate .wav file. This file is appended to the video file and the final output is obtained.

4.4 3D Face Model Construction

In order to construct a 3D face model, a generic 3D mesh model is used. Once the vertices and the edges of the mesh are defined, it can be customized to any individual user based on that particular user's facial features [IY96]. Hence, for constructing the 3D model from the 2D frontal image, one has to know the depth information of some of the facial features. This section deals with the following subtopics:

- Depth estimation using the stereo algorithm
- Customization of the generic head model
- Texture mapping the 3D mesh model

4.4.1 Stereo Computations

Since a projector is being used to project a pattern onto facial surface, stereo can be established between the projector and each camera. As per our earlier discussion

the two side views do not share many common face features between them, so it is required to establish stereo between the cameras and projector.

Using two calibrated cameras, an unknown 3D point [x,y,z] can be computed from its two images. Let [x,y,z] be the 3D point whose coordinates are to be found. Let this point be projected at $[r_1,c_1]$ and $[r_2,c_2]$ correspondingly in the two images. Using the above discussed camera model,we get

$$\begin{bmatrix} sr_1 \\ sc_1 \\ s \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{4.7}$$

$$\begin{bmatrix} tr_2 \\ tc_2 \\ t \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{4.8}$$

Eliminating the homogeneous coordinates from equations 4.7 and 4.8, we get

$$r_1 = (b_{11} - b_{31}r_1)x + (b_{12} - b_{32}r_1)y + (b_{13} - b_{33}r_1)z + b_{14} \tag{4.9}$$

$$c_1 = (b_{21} - b_{31}c_1)x + (b_{22} - b_{32}c_1)y + (b_{23} - b_{33}c_1)z + b_{24} \tag{4.10}$$

$$r_2 = (c_{11} - c_{31}r_2)x + (c_{12} - c_{32}r_2)y + (c_{13} - c_{33}r_2)z + c_{14} \tag{4.11}$$

$$c_2 = (c_{21} - c_{31}c_2)x + (c_{22} - c_{32}c_2)y + (c_{23} - c_{33}c_2)z + c_{24} \tag{4.12}$$

Equations 4.9, 4.10, 4.11 and 4.12 can be solved for three unknowns (i.e. x, y and z). However, any three equations will yield slightly different results. This is because of the approximations that are made during the calibration stage. A better way of solving for the unknown values is to use the "closest-approach" algorithm

which is described as follows:

Let P_1 and P_2 be points on a line and Q_1 and Q_2 be two points on another line. Let u_1 and u_2 be the two unit vectors in the direction of the line joining the points as shown in the Figure 4.7.

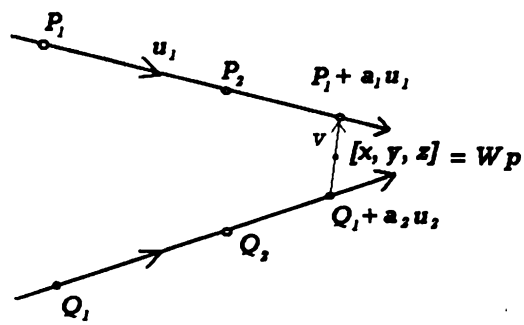


Figure 4.7: The closest approach method

Let v be the shortest vector connecting the two lines P_1P_2 and Q_1Q_2. Hence, v can be written as

$$v = P_1 + a_1u_1 - (Q_1 + a_2u_2) \tag{4.13}$$

Now, a_1 and a_2 can be represented in terms of the known values.

$$((P_1 + a_1u_1) - (Q_1 + a_2u_2)) \cdot u_1 = 0 \tag{4.14}$$

$$((P_1 + a_1u_1) - (Q_1 + a_2u_2)) \cdot u_2 = 0 \tag{4.15}$$

Solving Equations 4.14 and 4.15, we get

$$a_1 = \frac{(Q_1 - P_1) \cdot u_1 - ((Q_1 - P_1) \cdot u_2) \times (u_1 \cdot u_2)}{1 - (u_1 \cdot u_2)^2} \tag{4.16}$$

$$a_2 = \frac{((Q_1 - P_1) \cdot u_1) \cdot (u_1 \cdot u_2) - (Q_1 - P_1) \cdot u_2}{1 - (u_1 \cdot u_2)^2} \tag{4.17}$$

Solving Equations 4.16 and 4.17, the values of a_1 and a_2 are obtained and hence the vector v is estimated. if $|V|$is less than some threshold value then the intersection of the two rays is reported as shown in Equation 4.18

$$[x, y, z]^t = (1/2)[(P_1 + a_1 u_1) + (Q_1 + a_1 u1)] \tag{4.18}$$

4.4.2 3D Model Generation

A generic 3D face mesh, shown in Figure 4.8, was obtained from the University of Sheffield's 3D Computer Graphics Research Lab. In this head model, there are 395 vertices and 818 triangles.

The HCS is defined in Figure 4.9. The three axes and the origin are shown in the figure. To construct an individualized texture mapped 3D head model, the depth z of the facial parameters estimated in the WCS is converted into the HCS. The frontal view is texture mapped onto the 3D mesh in the HCS. The coordinate axes of the calibration sphere are decided by the orientation of the sphere. During the calibration stage, the sphere is positioned in such a way that the z - axis in both the WCS and HCS are parallel to each other.

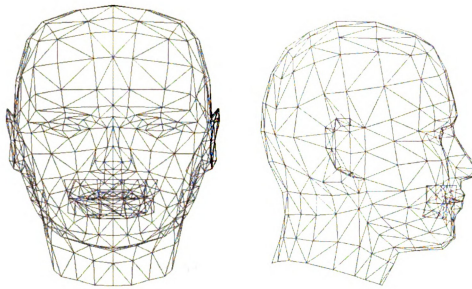


Figure 4.8: Front view and side view of the 3D generic mesh model of the face

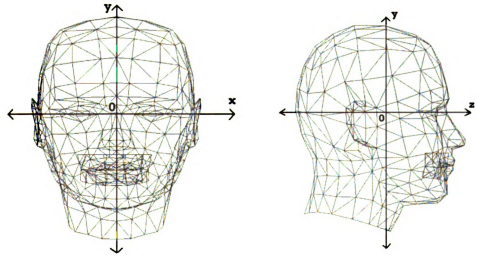


Figure 4.9: HCS with the origin (0) and three perpendicular axes (x, y and z)

4.5 Implementation Details

File Sizes: As described earlier, two videos are captured in real-time. The image sequences are captured without any compression for effective CPU usage. Moreover, any processing of the video will be much easier in an uncompressed form. Each video is captured for 10 seconds which means 300 frames. Each frame is of resolution 320 X 240 and contains 3 channels (RGB). Hence the total size of each file = 300 * 320 * 240 * 3 / (1024 * 1024) = 65.9 MB.

Data Processing: Currently, the data is being processed off-line. This is due to the restrictions of the hardware and specifically the hard disk writing speed. The process of creating the calibration table takes a couple of minutes. The practical writing speed of the hard disk is measured to be 9MB/s. For every second, an uncompressed .avi (audio-video interlace) file of size 6.59 MB is generated and hence the hard disk and the data bus must be capable of transmitting the data at the rate of 14MB/s . If the resolution of the video is to be 640 X 480, then the total required write speed rate must be at least 54 MB/S. The voice is captured using a microphone and is stored in a .wav file. The voice is synchronized with the video file in the post-processing stage using Adobe Premiere 6.0 software.

Chapter 5

Experimental Results

This chapter shows some results obtained from our system. Section 5.1 gives the results of the calibration process. Frames from synthesized virtual videos are shown in Section 5.2. Some preliminary results of the face modeling are shown in Section 5.3.2

5.1 Calibration

Section 5.1.1 describes the results obtained during the calibration of the virtual video synthesis. The calibration results for the 3D face modeling are described in Section 5.1.2

5.1.1 Virtual Video Synthesis

During the calibration, a rectangular grid of dimension 400 x 400 is projected onto the face. The grid is made by repeating three colored lines. Colored lines were used

because it is easy to distinguish the lines (see Figure 5.1)on the captured side views. We used white, green and cyan colors for this purpose. These colors were chosen because of their bright appearance over the skin color. The first few frames will have the grid projected onto the face and then the grid is turned off. One of the frames with the grid is taken and the transformation tables are generated. We calculate the transform functions represented by Equations 4.1 and 4.2 from the grid coordinates in both side views and the original grid coordinates. Each pixel in the frontal virtual grid is then obtained from the corresponding inverse transform functions.

The size of the grid pattern that is projected in the calibration stage plays a significant role in the quality of the video. This size is decided based on the trade-off between the quality of the video and the time taken. An appropriate grid size has be chosen based on trial and error. We started by projecting a sparse grid pattern onto the face and then increasing the density of the grid pattern. At one point, the increase in the density does not significantly improve the quality of the face image but starts consuming more time. At that point, the grid is finalized. Also, this grid is decided based upon the average face size. In our experiments, we settled on a grid cell size of row-width of 24 pixels and column-width of 18 pixels.

Figure 5.2 shows the frames that are captured during the calibration stage of the experiment.

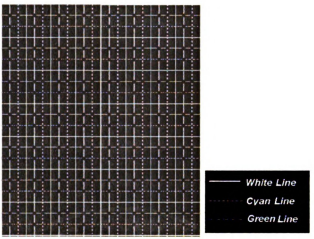


Figure 5.1: A square grid with alternating three colors is projected onto the face. Each grid cell has a row-width =24 pixels, col-width=18 pixels.

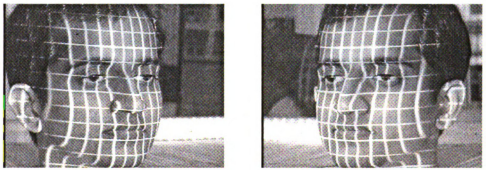


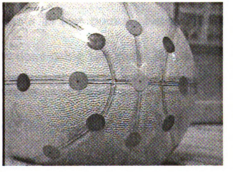
Figure 5.2: Face images captured during the calibration stage using environment-static FCS

5.1.2 Face Modeling

As discussed earlier (Section 4.2.3), to apply the structure from stereo method, we need to calibrate both cameras and the projector with respect to WCS. We shall now discuss the results of the cameras and projector calibration.

Camera Calibration

This section shows some of the sample results of the camera and projector calibration procedures that were discussed in Section 4.2.3. Figure 5.3 shows some of the images of the calibration sphere that were captured from the left camera and the right camera during the calibration stage.



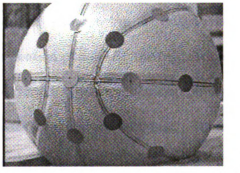


Figure 5.3: The images that are captured from the left camera and the right camera during the camera calibration.

Equations 5.1 and 5.2 shows the camera matrices obtained from the 12 calibration points shown in Figure 5.3.

$$\begin{bmatrix} 32.150 & 0.921 & 8.699 & 135.927 \\ 2.809 & -32.033 & -4.347 & 124.153 \\ 0.015 & 0.003 & -0.023 & 1.000 \end{bmatrix} \tag{5.1}$$

$$\begin{bmatrix} 28.197 & 1.474 & -16.748 & 164.733 \\ -0.509 & -32.143 & -4.232 & 118.646 \\ -0.011 & 0.002 & -0.023 & 1.000 \end{bmatrix} \tag{5.2}$$

After obtaining the calibration matrices using the methods described in the previous chapter, one can evaluate the accuracy of these matrices using the calibration points as test points. Estimation of these calibration points in the 2D image can be done using the given 3D coordinates of the calibration points and the calibration matrix. The difference in the actual 2D image coordinates of these calibration points and the estimated values of the 2D image coordinates is termed the "residual". Table 5.1 shows the residuals obtained while calibrating the left camera.

Calibration of the Projector

The Table 5.2 shows the errors in the calibration of the projector. The errors are slightly higher when compared to both cameras. Perhaps, these errors might have been due to the radial distortion of the projector. Also in the case of the projector,

Table 5.1: Results of left camera calibration for points on the calibration sphere.

INPUT DATA						OUTPUT DATA				
Point	2D Image		3D Coordinates			2D Fit Data		Residuals		
	u	v	x	y	z	u^1	v^1	u_{error}	v_{error}	$Total_{error}$
A	198	116	0.00	0.00	4.65	198.0	116.7	0.0	-0.7	0.7
B	190	36	0.00	2.33	4.03	189.8	35.3	0.2	0.7	0.9
C	135	56	-1.64	1.64	4.03	135.0	55.7	-0.0	0.3	0.3
D	109	115	-2.33	0.00	4.03	110.4	114.9	-1.4	0.1	1.5
E	134	176	-1.64	-1.64	4.03	132.9	176.4	1.1	-0.4	1.5
F	187	202	0.00	-2.33	4.03	187.7	201.4	-0.7	0.6	1.3
G	240	177	1.64	-1.64	4.03	240.3	177.2	-0.3	-0.2	0.5
H	262	121	2.33	0.00	4.03	261.6	120.5	0.4	0.5	0.9
I	241	62	1.64	1.64	4.03	241.3	62.7	-0.3	-0.7	1.0
K	74	16	-2.85	2.85	2.33	73.8	16.3	0.2	-0.3	0.5
L	30	116	-4.03	0.00	2.33	30.1	115.9	-0.1	0.1	0.2
M	69	220	-2.85	-2.85	2.33	69.2	220.1	-0.2	-0.1	0.3

all the 17 calibration points (shown in Figure 4.5) are chosen for calibration. The main difference in the case of the projector when compared with the previously discussed camera calibration method is that the (u,v) coordinates are obtained by projecting an image with known image points onto the 3D calibration points. The mouse was used to click on the 2D image that is projected onto the 3D world points. Equation 5.3 shows the projector calibration matrix.

$$\begin{bmatrix} 35.434 & 0.613 & -9.550 & 512.584 \\ 0.256 & -35.130 & -11.145 & 449.176 \\ 0.000 & 0.000 & -0.018 & 1.000 \end{bmatrix} \tag{5.3}$$

Table 5.2: Results of projector calibration for points on the calibration sphere.

INPUT DATA						OUTPUT DATA				
Point	2D Image		3D Coordinates			2D Fit Data		Residuals		
	u	v	x	y	z	u^1	v^1	u_{error}	v_{error}	$Total_{error}$
A	511	432	0.00	0.00	4.65	510.3	433.1	0.7	-1.1	1.8
B	512	348	0.00	2.33	4.03	511.9	347.3	0.1	0.7	0.8
C	449	373	-1.64	1.64	4.03	448.9	372.7	0.1	0.3	0.4
D	420	435	-2.33	0.00	4.03	422.0	434.8	-2.0	0.2	2.2
E	448	496	-1.64	-1.64	4.03	447.0	497.4	1.0	-1.4	2.4
F	509	524	0.00	-2.33	4.03	509.4	523.7	-0.4	0.3	0.7
G	573	498	1.64	-1.64	4.03	572.4	498.2	0.6	-0.2	0.8
H	600	437	2.33	-0.00	4.03	599.3	436.0	0.7	1.0	1.7
I	573	374	1.64	1.64	4.03	574.2	373.5	-1.2	0.5	1.7
J	513	292	0.00	4.03	2.33	513.6	293.7	-0.6	-1.7	2.3
K	409	336	-2.85	2.85	2.33	407.9	336.2	1.1	-0.2	1.3
L	363	441	-4.03	0.00	2.33	362.7	440.5	0.3	0.5	0.8
M	405	547	-2.85	-2.85	2.33	404.8	545.5	0.2	1.5	1.7
N	508	588	0.00	-4.03	2.33	509.4	589.6	-1.4	-1.6	3.0
O	616	548	2.85	-2.85	2.33	615.2	546.8	0.8	1.2	2.0
P	660	441	4.03	0.00	2.33	660.2	442.4	-0.2	-1.4	1.6
Q	618	339	2.85	2.85	2.33	618.1	337.6	-0.1	1.4	1.5

5.2 Virtual Video Synthesis

This section discusses the results of the virtual video synthesis algorithm that is described in Section 4.3. It also discusses the synchronization issues that might affect the quality of the novel virtual views.

5.2.1 Face Warping

The results [1] of the warping during the calibration and the operation stage is shown in Figure 5.4.

[1] If this thesis work was accessed through the University Microfilm (UMI) database, the images will not be in color. Even though the color information is lost, the author feels that the concept is well explained using gray scale images.

(a) (b) (c) (d)

Figure 5.4: Frontal view generation during the calibration stage and reconstruction of the frontal image from the side view using the grid: (a) left image captured during the calibration stage. (b) virtual image constructed using the transformation tables and the right image during the calibration stage. (c) right image captured during the operational stage. (d) result of the reconstructed frontal view from the transformation tables and the right image during the operational stage

5.2.2 Virtual View Synthesis

Figure 5.6 shows the output image of the frontal view that is generated by our algorithm. This output is obtained by applying our warping and mosaicking algorithms to the left and right views shown in Figure 5.5.

Figure 5.5: Face images captured during the operational phase using the ESFCS

Figure 5.7 shows the side views of the human face captured using the HMFCS. The main problems of capturing the faces using HMFCS are:

1. Lighting variations

2. Distortion caused by the mirrors

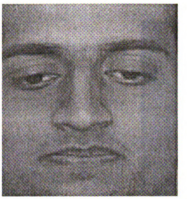
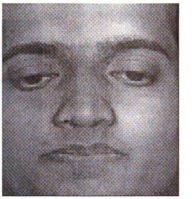



Figure 5.6: (a) Frontal view that is obtained from the camcorder and (b) virtual frontal view generated from our algorithm

3. Vibrations of the cameras and the mirrors


Figure 5.7: Face images captured using the HMFCS (a) left image and (b) right image

Figure 5.8 shows the output of the virtual view generated from the images captured using HMFCS.

5.2.3 Video Synchronization

Synchronization in the two videos is crucial in our application. Since, two views of a face with lip movements are merged together, any small changes in the synchronization will have high impact on the misalignment of the lips. This synchronization can

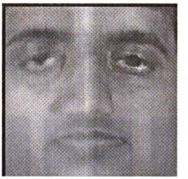


Figure 5.8: Virtual frontal view generated from the side views captured through HMFCS

Figure 5.9: (a) Top row: images captured from the left camera. (b) Second row: images captured using the right camera. (c) Third row: images captured using camcorder that is placed in front of the face. (d) Final row: virtual frontal views generated from the images in the first two rows

be evaluated based on sensitive movements such as eyeball movements (see Figure 5.10) and blinking eyelids (see Figure 5.11). Similarly, mouth movements can be analyzed from the virtual videos.

Figure 5.10: Synchronization of the eyeball movements: real video is in the top row and the virtual video is in the bottom row.

Figure 5.11: Synchronization of the eyelids during blinking: real video is in the top row and the virtual video is in the bottom row.

5.3 3D Face Model Construction

A 3D individualized head model can be constructed by adjusting a generic mesh model to fit the 3D feature points of an individual face. Depth information at the eye corners, mouth corners and nose tip is estimated using the stereo method. Section 5.3.1 deals with the results of the stereovision computations.

5.3.1 Stereovision Computations

Table 5.3 shows the error obtained during the estimation of 3D world coordinates on the calibration sphere. On average, the error in coordinates was less than 0.1 inches. The maximum total error was not more than 0.2 inches. For estimating the depth information of the facial feature points this accuracy might be sufficient.

Table 5.3: Depth estimation from left camera and projector for points on the calibration sphere. 3D coordinate dimensions are in inches.

	INPUT DATA							OUTPUT DATA			
Point	3D Coordinates			Left Camera		Projector		Computed points			Error
	x	y	z	u	v	x	y	x^1	y^1	z^1	
A	0.00	0.00	4.65	198	116	511	432	0.00	0.02	4.69	0.06
B	0.00	2.33	4.03	190	36	512	348	-0.02	2.30	4.12	0.14
C	-1.64	1.64	4.03	135	56	449	373	-1.65	1.64	4.07	0.05
D	-2.33	0.00	4.03	109	115	420	435	-2.38	-0.01	4.09	0.12
E	-1.64	-1.64	4.03	135	176	448	496	-1.62	-1.62	4.14	0.15
F	0.00	-2.33	4.03	187	202	509	524	-0.02	-2.34	4.04	0.04
G	1.64	-1.64	4.03	240	177	573	498	1.65	-1.64	4.04	0.02
H	2.33	0.00	4.03	262	121	600	437	2.32	-0.03	4.11	0.12
I	1.64	1.64	4.03	241	62	573	374	1.59	1.64	4.17	0.19
K	-2.85	2.85	2.32	74	16	409	336	-2.83	2.86	2.30	0.05
L	-4.03	0.00	2.33	30	116	363	441	-4.03	0.00	2.31	0.02
M	-2.85	-2.85	2.33	69	220	405	547	-2.84	-2.86	2.26	0.09

5.3.2 Customization of the 3D Face Model

The generic model used to generate the customized texture mapped face model was described in Figure 4.8. This model is made up of 395 vertices and a total of 2454 edges forming 818 triangles. Each triangle is defined by three vertices. This model is a complete head model that contains the ears and the back head portion. It also includes the eye balls, which are not common in some other face models. The customization of the generic head model is done by distorting the generic model as described in [IY96]. As discussed earlier, the 3D information of some of the prominent facial feature points are estimated and the 3D model is distorted to obtain the customized head model. The frontal view texture used for texture mapping the 3D model is shown in Figure 5.12. This texture map was obtained by extending the grid lines during the calibration stage. Figure 5.13 shows different rendered views of the 3D face model that was constructed from the two side views.

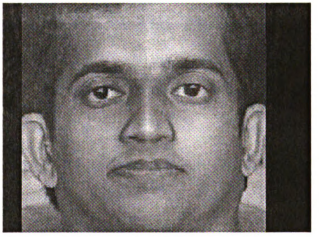


Figure 5.12: Frontal texture used for 3D face model construction

Figure 5.13: Different views rendered from the texture mapped 3D face model

Chapter 6

Assessment of the Results

Section 6.1 describes some evaluation procedures used to assess the quality of the generated virtual frontal views. Some more discussion of the results and issues related to the stability and portability of the FCS are given in Section 6.2.

6.1 Evaluation Schemes

Some researchers have worked on the evaluation of facial expressions [POM99, SCW+01]. Evaluating novel views is not studied much in the literature. In our case, we need to evaluate the synthesized videos in comparison with the real videos. This evaluation must give the accuracy of facial alignment, lip and eye movements and the perceptual quality of the synthesized videos. One can broadly classify the evaluation procedures into two kinds:

- Objective evaluation
- Subjective evaluation or quality assessment.

Section 6.1.1 describes the methods used for objective evaluation. The quality assessment of the videos is discussed in Section 6.1.2.

6.1.1 Objective Evaluation

One approach to assess the video is to theoretically evaluate the system. This approach doesn't require any human intervention or feedback. An error is obtained by comparing the virtual video frames with the real video frames of the frontal face. For effective comparison, the real video frames that are captured using a camcorder and the virtual video frames are normalized to a size of 200 X 200. The five images that were considered for evaluation are shown in Figure 6.1.

Figure 6.1: Images considered for objective evaluation (a)Top row: real video frames (b) Bottom row: virtual video frames

This evaluation can give some information regarding the facial feature alignment and facial movements which form the basis for facial interpretation and recognition. This can be done in two ways:

- Normalized cross-correlation of the 2D intensity arrays

- Euclidean distance measure between facial feature points

Figure 6.2 shows the bounding boxes of regions that are considered for evaluation using normalized cross-correlation method. The entire window was also used for evaluation.

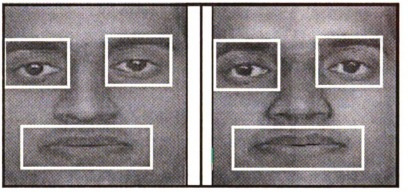


Figure 6.2: (a) Facial regions compared using normalized cross-correlation (Left: real view and Right: virtual view.)

Normalized Cross-correlation

Let,

h be the height of the image and

w be the width of the image.

The cross correlation between virtual image (V) and real image (R) of width w and height h is given by Equation 6.1.

$$CC_{VR} = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} V[i,j]R[i,j] \tag{6.1}$$

$$\| V \| = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} V[i,j]V[i,j] \tag{6.2}$$

$$\| R \| = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} R[i,j]R[i,j] \tag{6.3}$$

Equations 6.2 and 6.3 define the magnitudes of image V and R respectively. Equation 6.4 gives the normalized cross correlation between two images V and R.

$$NCC_{VR} = \frac{CC_{VR}}{\| V \| \quad \| R \|} \tag{6.4}$$

Table 6.1: Results of normalized cross-correlation between the real and the virtual frontal views. This normalized cross-correlation is applied in various regions of the face concentrating more at the eye and mouth regions

video	left eye	right eye	mouth	eyes + mouth	complete face
Frame1	0.988	0.987	0.993	0.989	0.989
Frame2	0.969	0.972	0.985	0.978	0.985
Frame3	0.969	0.967	0.992	0.978	0.986
Frame4	0.991	0.989	0.993	0.990	0.990
Frame5	0.985	0.986	0.992	0.988	0.989

The value of the normalized cross-correlation ranges between -1 and 1 with values of low absolute value indicating low similarity and absolute values near 1 indicating high similarity. In general, there was a high correlation between the real and the virtual images. Frames 2 and 3 shown in Figure 6.1 contain facial expressions (eye and lip movements) that were quite different from the expression used during the calibration stage and hence the generated view gave a lower correlation value when compared to the other frames. Also, the facial expressions in the frames 1 and 4 were similar to that of the expression in the calibration frame. Hence, these frames have a high correlation value compared to the rest. The eye and lip regions were considered

for evaluating the system because during any facial movement, these regions change significantly.

Euclidean distance measure

Using the Euclidean distance measure, the error is estimated to be the difference in the normalized Euclidean distances between some of the most prominent feature points. The feature points are chosen in such a way that one of them is relatively static with respect to the other. This will help us to evaluate the facial movements more accurately since the difference is calculated using a pseudo-reference(static) point. For example, if we consider some prominent feature points of the face (such as corners of the eyes, nose tip, corners of the mouth), the corners of the eyes are relatively static when compared to the corners of the mouth.

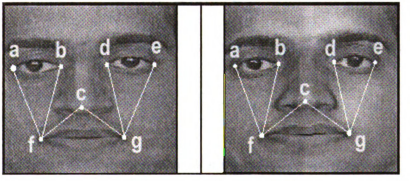


Figure 6.3: Facial feature points and the distances that are considered for evaluation using Euclidean distance measure (Left: real view. Right: virtual view.)

Figure 6.3 shows the most prominent facial feature points and the distances between those points that are considered for evaluation using the Euclidean distance measure. The basic assumption here is that the corners of the eyes and the nose tip are static with respect to the cameras. However, there will be a lot of mouth

movements and hence the position of the lip corners is not static. When distances between two feature points are measured, one point is chosen to be a static and the other one to be dynamic.

The Euclidean distance between two points whose coordinates are (x_i, y_i) and (x_j, y_j) is given by Equation 6.5

$$ED = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{6.5}$$

Let R_{ij} represent the Euclidean distance between two feature points i and j in the real frontal image and V_{ij} represent the Euclidean distance between two feature points in the virtual frontal image. The difference in the Euclidean distance is given by Equation 6.6.

$$D_{ij} = |R_{ij} - V_{ij}| \tag{6.6}$$

The total error ϵ for comparing the face images is defined by Equation 6.7

$$Total\ Error\ (\epsilon) = \frac{1}{6}[D_{af} + D_{bf} + D_{cf} + D_{cg} + D_{dg} + D_{eg}] \tag{6.7}$$

Table 6.2: Euclidean distance measurement of the prominent facial distances in the real image and virtual image and the defined average error. All dimensions are in pixels.

Frames	D_{af}	D_{bf}	D_{cf}	D_{cg}	D_{dg}	D_{eg}	Total Error(ϵ)
Frame1	2.00	0.80	4.15	3.49	2.95	3.46	2.80
Frame2	0.59	3.00	0.79	4.91	0.63	0.80	1.79
Frame3	1.88	3.84	4.29	4.34	2.68	1.83	3.14
Frame4	1.09	2.97	2.10	6.33	3.01	4.08	3.36
Frame5	1.62	2.21	5.57	4.99	1.24	1.90	2.92

The results in Table 6.2 indicate small error in the Euclidean distance measurements. An error of 3 pixels is not a significant quantity in an image of size 200 X

200. The facial feature points in the five frames were selected manually and hence the errors might have also been caused due to the unstability of manual selection. Also, one can note that the error values of D_{cf} and D_{cg} are larger than the others. This is probably because the nose tip is not a robust point when compared to eye corners. Hence, the errors in the distances involving the nose tip are more.

Some of the errors obtained in both of the above mentioned methods might also be due to the difference in the resolution of the images. The virtual frontal view is of resolution 162 X 192. The real frontal view is of higher resolution ($\approx$ 320 X 240).

6.1.2 Subjective Evaluation

Subjective evaluation involves some kind of a human intervention. The response of human in evaluating the system can vary from a simple "yes/no" to a numerical rating of the quality (1-6). Perhaps, this is a test that evaluates the quality of the virtual videos for supporting perception. The main factors that might affect quality of the virtual frontal face videos include

- Eye and lip movements
- Facial expressions
- Synchronization of the two halves of the face
- Color and texture of the face
- Quality of the audio
- Synchronization of audio

A preliminary subjective study has been made on some of the virtual videos by the author and CSE advisor. In general, the quality of the videos was assessed as adequate to support the applications for which the teleportal system is intended. The two halves of all the videos are well synchronized and color balanced. The quality of the audio is good and it has been synchronized well with the lip movements. Some observed quality factors were distortion in the eyes and teeth and in some cases a cross-eyed appearance. The face appears a little bulged when compared to the real videos. The main reason is that the cameras used to capture the side view have smaller focal length than that of the camcorder and hence the distortion in the images is more when compared to the real images captured using the camcorder.

Further analysis of these videos is a future task of the teleportal project. Tests can be conducted where the virtual videos and real videos are displayed in random order and human judges are asked to identify or evaluate the real and the virtual videos. This test can ask the subjects to evaluate the expressiveness of the video. Judgements can be rated on a scale of 1-6 where 6 represents highest confidence in evaluation of the expressions. Some standard expressions (e.g. joy, angry, surprise, sadness) could be judged.

6.2 Discussion of the Results

6.2.1 Time Taken

The time taken for processing the video is one important aspect of our system. Our goal is to make the system work in real-time. The total time taken can be mainly split into three parts. Pre-buffering estimates the time taken to transfer the images into the corresponding buffers. The next part is the time taken for doing the actual warping. This is the time taken for interpolating each of the grid blocks in the frontal image. The final part is the time taken for post processing. The post-processing consumes little time (less than 5% of the total time) because a linear filter is applied to smooth the image. The time taken for the overall procedure is directly proportional to the density of the grid. In our case, the average time taken per frame for processing the videos is around 60 ms using a computer with 746 MHz. Any processing that consumes less than 30 ms is considered to be real-time processing. However, the time taken in our case can be optimized and can be made to work in real-time in a high speed computer (with 2.6 GHz processing speed).

6.2.2 Positioning of Cameras and Mirrors

Placing the cameras and mirrors in the appropriate positions is crucial to the quality of the results. Most importantly, the cameras must focus the image of the face inside the mirror. The mirrors are to be adjusted in such a way that the face is well captured. Also, the vibrations of the cameras and the mirrors are to be minimized, especially when the person is speaking.

The angle at which the two side views are captured will have significant impact on the generated frontal view. Even though, it is possible to create a frontal view from widely separated side views using our algorithm, the facial expressions will be more distorted in the case of widely separated views. Hence, it is important that care should be taken to capture images that are not widely separated.

On the other hand, the accuracy in the depth measurements will be improved significantly if the cameras are widely separated because the depth information is extracted from the stereo between the cameras and the projector. As the cameras move farther apart, the rays from the camera and the projector tend to intersect at wider angles and thus the intersection points of the rays will be more accurate. If the views are widely separated then there will not be many common feature points between the two images captured from the cameras. In such a case, there cannot be stereo between the images captured from the cameras alone.

6.2.3 Depth of Field Issues

When the face capture is done through the mirror, lot of importance has to be given to the issue of depth of field. Distant facial features like the nose tip and the ear corners are focussed well. To make the system portable, one has to make sure that it can be used among various individuals without making any changes. This system was successful in porting to various individuals (with different head sizes) without changing the position and the focus of the cameras and the mirrors.

Chapter 7

Conclusion

This chapter gives a discussion of the completed work on the facial capture system and the future directions. Figure 7.1 summarizes both accomplishments and future work.

7.1 Summary

The proposed facial capture system will be able to capture the human face in a mobile environment in real-time. A real time video stream of the frontal view of the face is obtained by merging the two side views captured by two side cameras. We have developed customized mirrors based on the calculations made from the optical layout of the system. The algorithm being used can be made to work in real-time because of its computational simplicity and has been demonstrated to be near real-time using a PIII computer. This working prototype has been tested on a diverse set of 10 individuals. For comparisons of the virtual videos with real videos, we expect that

important facial expressions will be represented adequately and that feature locations will not be distorted by more than 2% . To further demonstrate the promise of this approach, a 3D head model has been created from the two captured side views.

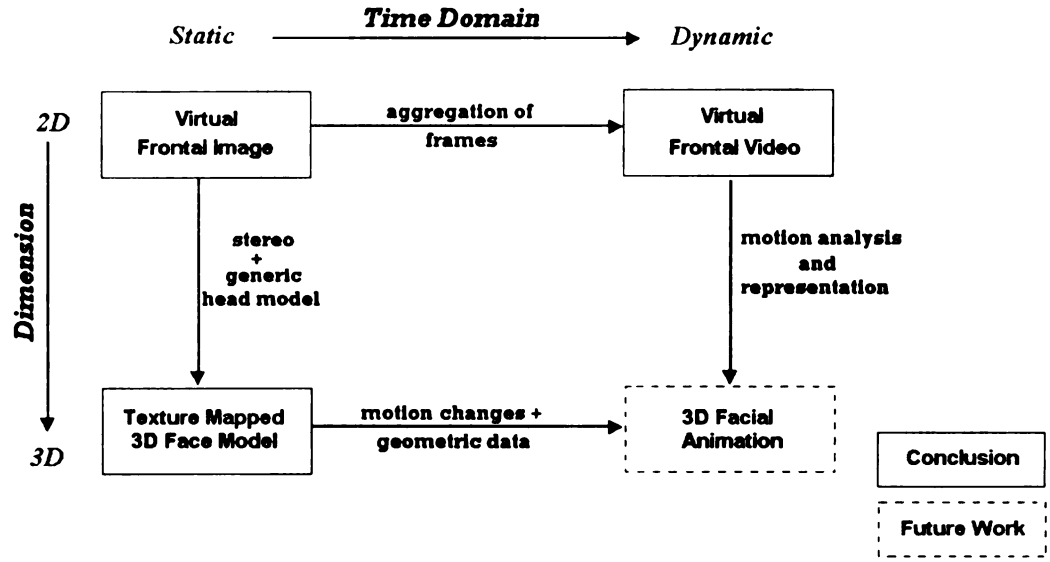


Figure 7.1: Conclusion and future work. Solid blocks indicate implemented subsystems. Dashed block indicates future subsystem.

7.2 Future Work

We need to simplify the calibration procedure by automating some of the steps during the calibration stage. This can be done by applying some of the thresholding algorithms on the two side views captured during the calibration. Online processing of the video in real-time and real-time remote collaboration should be demonstrated in the near future. Customizing the camera lenses might improve system quality. Algorithms for color-balancing should model the lighting conditions more effectively. Theoretically, it should be possible to reconstruct novel views from a range of view

points without reconstructing the 3D model. If this can happen in real-time, a floating window can show human faces from arbitrary views in space.

One can make the system work in real-time by using a high speed computer such as P4 with a processor speed of 2.6 GHz. If the calibration process is automated, the data can be processed online and transmitted via an Internet2 channel. Using the existing setup, two videos are grabbed simultaneously on the same computer. Instead of saving these videos into the hard disk, one can process them online in real-time. This concept of processing is called "quadruple buffering" which allows the user to process previous frames while current frames are being grabbed from both cameras.

Facial deformations that make significant alterations to the face surface will not be handled by the static warp discussed earlier. There might be need for a dynamic set of warps that can handle these alterations. The grid that was used might not handle various facial expressions effectively. A hierarchial grid that can project more dense grid patterns onto the eye and mouth regions may improve the quality of the output. The bilinear interpolation technique tends to be a block-wise operation. Cubic interpolation might give better results because cubic functions are used for modeling curved surfaces. The equipment has to be stabilized by fixing the capture system onto a more robust headset. An extra calibration step might be required to handle the distortion produced by the convex mirrors. This face capture system has to be integrated with the Head Mounted Projection Display [HGBR01].

In the years to come, one can enable a stereo field-of-view to be transmitted in time slices alternating between the human face view and the field of view of the user. This can be achieved by using an electro-optical glass in place of the mirrors [KS02].

Perhaps, one can achieve this by flipping the mirrors mechanically, allowing the same cameras to transmit the scene viewed by the mobile user. Thus, the FCS will be able to perform dual duty (1)to capture the face and (2)to capture the user's field of view.

The vibrations of the cameras and the mirrors are to be minimized especially when the person is speaking. These vibrations might be more while the person is in motion. If there are significant vibrations, then there will be a necessity for a video stabilization algorithm that has to be implemented while the image sequences are being captured.

Compression of the video streams will help in effective transmission of the data. Some parameters for an Internet2 transmission channel can be optimized for effective data communication for specific applications.

Appendix A

Conversion of Spherical to Cartesian Coordinates

The points on the calibration sphere are represented in spherical coordinates. These spherical coordinates are converted into cartesian coordinates with origin as center of the calibration sphere. Figure A.1 shows how a point P is defined in a spherical coordinate system. Let R be the radial distance of P from the origin. θ is the azimuthal angle in the xy-plane from the x-axis. ϕ is the polar angle from the z-axis. This is also called the "colatitude" of point P. The ranges of these angles are as follows

$0 \leq \theta \leq 2\Pi$ and $0 \leq \phi \leq \Pi$

Using basic trigonometry, the Cartesian coordinates (P_x, P_y, P_z) for the point P are defined from the spherical coordinates. (R, θ, ϕ)

$$P_x = RSin(\theta)Cos(\phi)$$

$$P_y = RSin(\theta)Sin(\phi)$$

$$P_z = RCos(\phi)$$

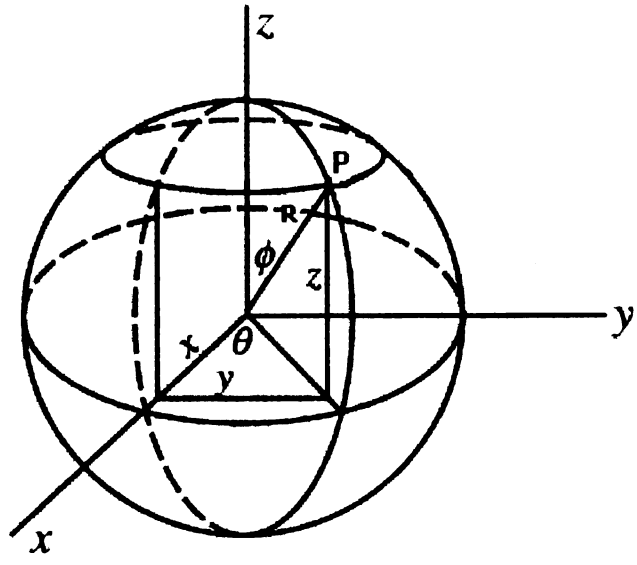


Figure A.1: Spherical coordinate system

Bibliography

[BC94] G. Burdea and P. Coiffet. *Virtual Reality Technology.* Wiley-Interscience, 1994.

[BF82] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14(4):553–572, 1982.

[BF98] D. Buxton and G. W. Fitzmaurice. HMDs, caves and chameleon: a human-centric analysis of interaction in virtual space. *Computer Graphics*, 32(4):69–74, 1998.

[BK87] K. L. Boyer and A. C. Kak. Color-encoded structured light for rapid active ranging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):14–28, 1987.

[BKP01] M. Billinghurst, H. Kato, and I. Poupyrev. The magicbook: Moving seamlessly between reality and virtuality. *IEEE Computer Graphics and Applications*, 21(3):6–8, 2001.

[Bor01] M. Bordenaro. *ASPs help make 'virtual meetings' successful.* Chicago Tribune, Nov. 19, 2001.

[BP93] R. Brunelli and T. Poggio. Face recognition: Features vs. templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, April 1993.

[BR00] F. Biocca and J. P. Rolland. *Teleportal face-to-face system.* Patent Filed, August, 2000.

[CNSD+92] C. Cruz-Neira, D. J. Sandin, T.A. DeFanti, R.V. Kenyon, and J. C. Cart. The cave: Audio visual experience automatic virtual environments. *Communications of the ACM*, 35(6):65–72, 1992.

[CW93] S. E. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of SIGGRPAH93*, pages 279–288, 1993.

[DMS98] D. DeCarlo, D. Metaxas, and M. Stone. An anthropometric face model using variational techniques. In *Proceedings of SIGGRAPH*, pages 67–74, 1998.

[DRHL+03] L. Davis, J. Rolland, F. Hamza-Lup, Y. Ha, J. Norfleet, and C. Imielinska. Alices adventures in wonderland: A unique technology enabling a continuum of virtual environment experiences. *Computer Graphics and Applications*, 23(2):10–12, 2003.

[DT96] F. W. DePiero and M. M. Trivedi. 3-d computer vision using structured light: Design, calibration, and implementation issues. *IEEE Computer*, 43:243–278, 1996.

[Fai90] G. Faigin. *The Artist's Complete Guide to Facial Animation.* Watson-Guptill Publications, 1990.

[FBA+94] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual space teleconferencing using a sea of cameras. In *Proceedings of the First International Symposium on Medical Robotics and Computer Assisted Surgery,*, 1994.

[Fei02] S. Feiner. Augmented reality: a new way of seeing. *Scientific American*, 286:48–55, 2002.

[Fis96] R. Fisher. *Head-mounted projection display system featuring beam splitter and method of making same.* US Patent 5,572,229, November 5, 1996.

[FSW97] K. E. Finn, A. J. Sellen, and S. B. Wilbur. *Video-mediated communication.* Lawrence Erlbaum Associates, 1997.

[GGSC96] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *Proceedings of SIGGRAPH*, pages 43–54, 1996.

[GGW+98] B. Guenter, C. Grimm, D. Wood, H. Malvar, , and F. Pighin. Making faces. In *Proceedings of SIGGRAPH*, pages 55–66, 1998.

[Goo95] D. Goodman. *Handbook of Optics 2nd Ed*, chapter General Principles of Geometric Optics. New York: McGraw-Hill, 1995.

[Ham01] M. Hamblem. *Avoiding travel, users turn to communications technology: Videoconferencing, Web collaboration use increasing in aftermath of attacks.* Computer World, Sept. 24, 2001.

[HB86] B. K. P. Horn and M. J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986.

[HGBR01] H. Hua, C. Gao, F. Biocca, and J. Rolland. An ultralight and compact design and implementation of head-mounted projective displays. In *Proceedings of IEEE Virtual Reality 2001*, pages 175–182, 2001.

[HGGR00] H. Hua, A. Girardot, C. Gao, and J. P. Rolland. Engineering of head-mounted projective displays. *Applied Optics*, 39(22):3814–3824, 2000.

[HS89] G. Hu and G. Stockman. 3-d surface solution using structured light and constraint propagation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(4):390–402, April 1989.

[IY96] H.H.S. Ip and L. Yin. Constructing a 3d individualized head model from two orthogonal views. *Visual Computer*, 12:254–266, 1996.

[Jar83] R. A. Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):122–139, 1983.

[KGH85] M. Krueger, T. Gionfriddo, and K. Hinrichsen. Videoplace - an artificial reality. In *Proceedings of ACM CHI'85 Conference on Human Factors in Computing Systems*, pages 35–40, 1985.

[KO97] R. Kijima and T. Ojika. Transition between virtual environment and workstation environment with projective head-mounted display. In *Proceedings of IEEE 1997 Virtual Reality Annual International Symposium*, pages 130–137, 1997.

[KS02] A. M. Kunz and C. P. Spagno. Simultaneous projection and picture acquisition for a distributed collaborative environment. In *Proceedings of IEEE Virtual Reality 2002*, pages 279–280, 2002.

[LC01] S.H. Lai and C.M. Cheng. Three-dimensional face model creation from video. In *Proceedings of SPIE Conference on Three-dimensional Image Capture and Applications IV*, 2001.

[Len98] J. Lengyel. Telepresence by real-time view-dependent image generation from omnidirectional video streams. *IEEE Computer*, 31(7):46–53, 1998.

[Lin01] C. Lindquist. *Analysis: 8 hot technologies for 2002.* CNN Online, Dec. 31, 2001.

[MD99] R.A. Manning and C.R. Dyer. Interpolating view and scene motion by dynamic view morphing. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 388–394, 1999.

[MK94] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems*, 77(12), 1994.

[MM97] P. Mouroulis and J. Macdonald. *Geometrical Optics and Optical Design.* Oxford Univ. Press, 1997.

[OK93] M. Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(4):353–363, April 1993.

[Ost01] M. Osterman. *Messaging subs for travel, snail mail since attacks.* Network World Messaging Newsletter, Dec. 03, 2001.

[OYTY98] Y. Onoe, K. Yamazawa, H. Takemura, and N. Yokoya. Telepresence by real-time view-dependent image generation from omnidirectional video streams. *Computer Vision and Image Understanding*, 71(2):154–165, 1998.

[PGD98] M. Proesmans, L. Van Gool, and F. Defoort. Reading between the lines — a method for extracting dynamic 3d with texture. In *Proceedings of International Conference on Computer Vision*, pages 1081–1086, 1998.

[PHL+98] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, 1998.

[POM99] I. Pandzic, J. Ostermann, and D. Millen. User evaluation: Synthetic talking faces for interactive servicess. *Visual Computer*, 15(7/8):330–340, April 1999.

[PR98] J. Parson and J. P. Rolland. A non-intrusive display technique for providing real-time data within a surgeon's critical area of interest. In *Proceedings of Medicine Meets Virtual Reality 98*, pages 246–251, 1998.

[PW96] F. I. Parke and K. Waters. *Appendix 1: Three-dimensional muscle model facial animation.* A. K. Peters, 1996.

[RF00] J. P. Rolland and H. Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators and Virtual Environments*, 9(3):287–309, 2000.

[RK75] F. Rocker and A. Kiessling. Methods for analyzing three dimensional scenes. In *Proceedings of 4th. International Joint Conference on Artificial Intelligence*, pages 669–673, 1975.

[Sat94] K. Sato. Silicon range finder :a realtime range finding VLSI sensor. In *Proceedings of IEEE Custom Integrated Circuits Conference*, pages 339–342, 1994.

[SCW+01] M.A. Sayette, J. Cohn, J.M. Wertz, M.A. Perrott, and D.J. Parrott. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, 25:167 - 186, 2001.

[SD96] S. M. Seitz and C. R. Dyer. View morphing. In *Proceedings of SIGGRAPH96*, pages 21–30, 1996.

[Sei01] S.M. Seitz. The space of all stereo images. In *Proceedings of International Conference on Computer Vision*, pages 26–33, 2001.

[SN97] H. Saji and H. Nakatani. Measuring three-dimensional shapes of a moving human face using photometric stereo method with two light sources and slit patterns. In *Proceedings of IEICE Transactions on Information and Systems*, pages 795-801, 1997.

[SS89] N. Shrikhande and G. Stockman. Surface orientation from a projected grid. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):650–655, April 1989.

[SS01] L.G. Shapiro and G.C. Stockman. *Computer Vision.* Prentice-Hall, 2001.

[Sut65] I. Sutherland. The ultimate display. In *Proceedings of IFIP65*, pages 506–508, 1965.

[TK92] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography—a factorization method. *International Journal on Computer Vision*, 9(2):137–154, 1992.

[VBK02] S. Vedula, S. Baker, and T. Kanade. Spatio-temporal view interpolation. In *Proceedings of the 13th ACM Eurographics Workshop on Rendering*, June 2002.

[WBL+96] R. Welch, T. T. Blackmon, A. Liu, B. A. Mellers, and L. W. Stark. The effects of pictorial realism, delay of visual feedback, and observer interactivity on the subjective sense of presence. *Presence: Teleoperators and Virtual Environments*, 5(3):263–273, 1996.

[web] The Internet2 website. http://www.internet2.edu.

[Woo80] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[XLH02] L. Q. Xu, B. Lei, and E. Hendriks. Computer vision for a 3-d visualization and telepresence collaborative working environment. *BT Technology Journal*, 20(1):64–74, 2002.

[YLW98] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. In *Acian Conference on Computer Vision*, pages 687–694, 1998.

[ZT92] J.Y. Zheng and S. Tsuji. Panoramic representation for route recognition by a mobile robot. *International Jounal of Computer Vision,*, 9:55–76, 1992.