Visually Dependent Nonverbal Cues and Video Communication

David Tong Nguyen

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-48
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-48.html

May 12, 2008

Acknowledgement

mom and dad

Visually Dependent Nonverbal Cues and Video Communication

by

David Tong Nguyen

B.S. (Northwestern University) 2002
M.S. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor John Canny, Chair
Professor Avideh Zakhor
Professor Coye Cheshire

Spring 2008

The dissertation of David Tong Nguyen is approved:

Chair Date

Date

Date

University of California, Berkeley

Spring 2008

Visually Dependent Nonverbal Cues and Video Communication

Abstract

Visually Dependent Nonverbal Cues and Video Communication

by

David Tong Nguyen

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor John Canny, Chair

Humans have a rich verbal and nonverbal language that allows us to communicate in powerful ways. Technology continues to play a significant role in enabling and enhancing the ways we communicate with each other when we are geographically separated. Machines capture our expressions and transmit them anywhere in the world for distant partners. But how expression is captured and how it is presented can have surprising effects on the way people communicate.

In this dissertation, I present background, designs, and evaluations of a new video conferencing system called MultiView which aims to improve the group-to-group video conferencing experience. I will (1) show the dependence of spatial information for the effective communication of nonverbal information such – as eye contact – and formalize the spatial shortcomings of modern video conferencing system design, (2) introduce a design based on a new multiple-perspective display to address these shortcomings, and (3) show that the MultiView design can effectively capture and present nonverbal cues in a natural way and that this new ability dramatically improves trust formation between remotely meeting groups.

Professor John Canny
Dissertation Committee Chair

to my mom and dad
Loc and Phuong Nguyen
your guidance continually brings me to wonderful places.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my committee members. To my advisor, John Canny, thank you for your tremendous energy, brilliant insight, continuous support, and unfailing drive as it has guided me through my tenure as a doctoral student. To my committee members, Avideh Zakhor, Coye Cheshire, and Marty Banks, thank you for pushing me beyond my comfort zone and introducing me to new ways of thinking about our mutually shared interests.

My research would not have been possible without the exceptional support from the XLab staff. Thank you Brenda Naputi, Miho Tanaka, Lawrence Sweet, and Bob Barde for those times I said, "I need 200 participants next week," and you said, "OK!".

Thank you Harlyn Baker and Bruce Culbertson, my collaborators at Hewlett-Packard Laboratories, for sharing the excitement of designing the future of communication technologies with me. Your vision, energy, dedication, and technical expertise are inspiring.

Thank you to my BiD lab mates who have created a social, intellectual, and technical base for simply awesome research. In particular, I would like to acknowledge Jeremy Risner for your kid-like excitement in building anything physical and expert skills in applying large, optically clear laminates. Nothing will beat trips to Ace Hardware, late nights in the tool shop, and putting together 80/20.

To the friends I have made through the years, you were there for me when life was down, there when life was up, just always there for me - Alice Chau, Joanna Lai, Sarah Kim, Eric Tang, Carri Chan, Ginger Perng, Preethi Kumaresan, Dan Ceperley, Erik Steltz, Eric Jensen, Steve Beitzel, and Minh Evans. In addition, toward the end when I spent many late nights in lab, thank you, David Sun, for working there as well and keeping me sane.

And, of course, my family. Life would not be the same without my brother and sister to go through it with. You two are crazy. Thanks for keeping me entertained through the years. To my parents, to whom this dissertation is dedicated. I increasingly appreciate the difficult decisions you have made, the life you have created, the role you have played in defining my life, and the guidance you continually provide. I admire your patience and wisdom in raising such a bratty child. I promise to grow out of it soon.

Part I

Background

Chapter 1

Introduction

> If, as it is said to be not unlikely in the near future, the principle of sight is applied to the telephone as well as that of sound, earth will be in truth a paradise, and distance will lose its enchantment by being abolished altogether.
> – Arthur Mee, p345

Everything you do says something whether it is verbal or non-verbal, conscious or unconscious. Our bodies are capable of powerful expression through words that are said, a smile that is shared, or the shake of a hand. At the same time, technology has come together quite nicely with people, capturing their expressions and delivering it in real-time to distant partners. Technology designers have brought us a myriad of communication tools that mitigates barriers of distance in real-time communication. We remain connected with our friends, families, and colleagues by technologies such as the cellular telephone and instant messaging. However, as useful as textual- and audio-only technologies are, we know that our bodies do a significant amount of communication to supplement, enhance, or replace the spoken or written word. It would seem to make sense, then, that a communication channel that also provides visual information would prove to be extremely valuable.

However, the terrain to creating a successful visual communication system that captures and communicates our bodily language is rough and difficult to navigate. The multimedia technologies of video conferencing systems are complex, only to be challenged by understanding the ways people communicate. Both challenges, however, must be met to appropriately design effective communication technology.

These obstacles have not prevented us from trying. As early as 1964, AT&T showed off a demonstration of its picture phone at the World's Fair and introduced a

product a few years later in 1970. AT&T's push was strong, investing significant amounts in the technology and business infrastructure. At it's peak, it only had 500 users and faded away in 1974. They tried again in 1992 with the VideoPhone 2500, but that failed again as that product only lasted until 1995. Other major players who have tried in the video conferencing space include IBM, Mitsubishi, Ericsson, Philips, Sony, and PolyCom.

Today, a handful of major video conferencing players fill certain needs where specialized contexts make video conferencing useful. However, we have yet to see video conferencing making changes on the scale of those produced by telephony or instant messaging technologies. This indicates that current video conferencing designs do not quite meet the current needs of the users.

In this dissertation, I present a radically new design in video conferencing systems for group-to-group communication based on an understanding of interpersonal communication and how people perceive images in a video conferencing setting. My system, called MultiView, is based on a new display. It more faithfully reproduces many of the nonverbal cues we use for communication; nonverbal cues which would otherwise be distorted by even the most advanced video conferencing systems available. I will show that the MultiView design dramatically improves the way groups can communicate. The hope is that some day, we can more effectively capture the way we communicate visually and allow geographically separated partners to fully take advantage of this potential.

Chapter 2

Video Conferencing System Design

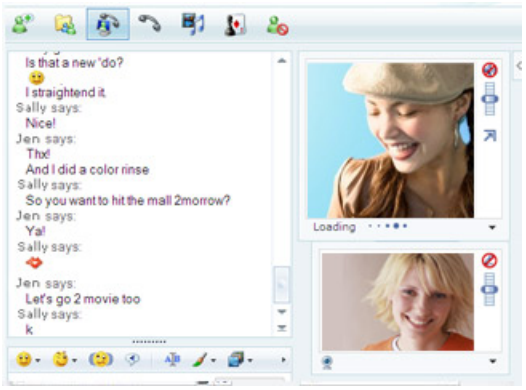
2.1 Product Space

There is a spectrum of video conferencing products available on the market that varies along many design dimensions. Additionally, people have become very creative in the ways they use video conferencing systems. This section will cover representative examples in the product spectrum to help situate the overall research and how these systems are typically used.

At one end of the spectrum are *desktop* video conferencing products. These products include Apple's iChat (Figure 2.1(a)), Microsoft's Windows Live Messenger (Figure 2.1(b)), and PolyCom's HDX4000 (Figure 2.1(c)). Generally, these products – like iChat and Messenger – are available as free software downloads and can run on a personal computer. They use the resources of the computers – camera, microphone, display, speakers, networking, and processing power – to create a complete video conferencing system. With these products, the video conferencing software is built into the instant messenger user interface. To start a connection is much like starting a chat with someone on your buddy list. Standalone systems – like the HDX4000 – include the necessary hardware for video conferencing in addition to all the necessary software. These systems usually require that you dial up a remote system's IP address, much like dialing a telephone number.

These products are designed to allow meetings between two and four sites with one participant at each site. When there are just two participants, these systems can provide convincing support for many nonverbal cues. By placing the camera right above the image

(a) Apple's iChat is an instant messenger tool freely available from Apple and bundled with their operating system. A feature of the iChat program is this video conferencing package which allows up to four remote participants to meet.



(b) Microsoft's Windows Live Messenger is an instant messenger tool available from Microsoft and is freely available. Like iChat, it also includes a video conferencing package.

(c) PolyCom's HDX4000 is a standalone unit including everything needed for a high-definition video conferencing system. Unlike the others, it does not run on a personal computer. The entire package currently retails for $4,999.

Figure 2.1: Three desktop video conferencing systems: (a) Apple's iChat, (b) Microsoft's Windows Messenger Live, and (c) PolyCom's HDX 4000.

of the remote person, convincing eye contact can be achieved [8]. By using the entirety of a desktop display, each person sees an adequate sized head-and-shoulders viewpoint of the remote participant. As more participants join the meeting, nonverbal cues become increasingly distorted and the image size of each remote person decreases with the number of participants in the meeting. Though primarily marketed as a system for one participant at each site, it is possible to use these systems with multiple people at each site.

Moving further along the spectrum are *meeting room* video conferencing products. These products include PolyCom's HDX Media Center (Figure 2.2(a)), Tandberg's Profile Series (Figure 2.2(b)), and LifeSize's Team MP systems (Figure 2.2(c)). These systems typically cost tens of thousands of dollars and include a large display, a camera system, a microphone system, an audio system, and a codec. The codec is responsible for capturing the camera and microphone signals, encoding audio and video, transmitting it to the remote location, decoding the remote signals, driving the displays, and driving the audio systems. Sometimes an additional display is available to share content. These systems connect to each other over an IP network. To connect to each other, they must dial each other's IP address.

These products are generally designed to allow between two and four sites of small groups to work with each other. The physical space that needs coverage is usually quite a bit larger than that of desktop systems: multiple people in the space of an entire room instead of just head and shoulders view of each person. To accommodate this, these video conferencing systems often utilize larger displays, when compared to their desktop counterparts, such as projection system or large flat panel displays. These systems are also typically characterized by pan, tilt, zoom (PTZ) cameras that allow both context – by letting the remote participants zoom out to see the entire room – and focus – by letting the remote participants zoom in and aim on a particular person or artifact. The usual configuration is to have this system at the end of a long conference table.

Recently, a new class of video conferencing systems have been introduced known as *telepresence* systems. These products include PolyCom's RealPresence (Figure 2.3(a)), Cisco's Telepresence (Figure 2.3(b)), and Hewlett-Packard's Halo (Figure 2.3(c)). These systems typically cost several hundred thousand dollars and require a monthly subscription fee that can cost tens of thousands of dollars. As of the time of this writing, HP's Halo system currently sells for $349,000 for the initial installation of the system and $18,000/month for service. These systems are purchased as rooms and include, in addition

(a) PolyCom's HDX Media Center

(b) Tandberg's Profile Series

(c) LifeSize's Team MP

Figure 2.2: Three meeting room video conferencing systems: (a) PolyCom's HDX Media Center, (b) Tandberg's Profile Series, and (c) LifeSize's Team MP.

to the video conferencing technology hardware, the tables, chairs, lighting, and decoration. These systems include between two and four large projection or flat-panel screens to display life-sized images of the remote participants. For each remote display, there is one local fixed camera to capture a specific portion of the room. The participants are unable to control these cameras like in the meeting room style video conferencing systems. An additional display is used for content sharing.

These products are designed for the round-table meeting structure with the goal of simulating face-to-face meetings as much as possible. There is usually a long table where the local participants sit on one edge, and a bank of displays sit on the other edge. As part of the package, these systems usually include a concierge service, a live representative who can control the entire setup and make sure the system is ready for the meeting.

Let us consider two factors: image size and support for gaze awareness. In desktop systems with a meeting dyad, the image of the remote participant is pretty close to life-size if the head and shoulders are captured. Additionally, with careful camera placement [8], convincing eye contact can be achieved. In meeting room systems that support group-to-group meeting structures, the physical space covered by the video conferencing systems grows much faster than the screen real estate. If we were to view the entire space, the image size of the remote participants would necessarily be smaller. Additionally, for reasons that will be explained in following sections, eye contact information is severely distorted. Telepresence systems introduce much more screen real estate, allowing all remote participants to be represented as life-sized images. But much like the meeting room systems, because of the group-to-group structure of the meeting, these systems still exhibit severe distortions in gaze information. In my work, I aim to provide life-size imaging and support for gaze awareness in group-to-group meetings.

2.2 Design Space

Creating a video conferencing system provides a rich and complex design space for several reasons. First, video conferencing systems incorporate many complex multimedia technologies. Second, if video conferencing systems are to be a versatile communication device, it must support a wide variety of human activities. Below, I discuss several general dimensions of designing a video conferencing system, though this list is far from comprehensive.

(a) PolyCom's Real Presence

(b) Cisco's Telepresence

(c) Hewlett-Packard's Halo

Figure 2.3: Three telepresence video conferencing systems: (a) PolyCom's Real Presence, (b) Cisco's Telepresence, and (c) Hewlett-Packard's Halo.

Audio/Video Quality Primarily, video conferencing systems need to capture, compress, transport, decompress, and display both audio and video information under tight latency and quality constraints. Visual factors include resolution, color quality and frame rate. Audio factors include sampling rate and echo cancellation. Both audio and video streams must be delivered quickly to the remote sites and presented to the remote participants such that they appear synchronized.

Data Collaboration Video conference systems often build in features which allow participants to collaborate over shared visual artifacts such as powerpoint slides or a physical item. Designs that support this include screen sharing or having extra cameras to focus on some object of interest.

User Interface Video conferencing systems involve a lot of complex technologies, and it is important to design effective ways for users to interact with the system. The user interface is responsible for functions such as allowing users to control the audio/video system, connect to each other, and share data.

Meeting Structure People meet in several different structures. Sometimes it is one-on-one, other times it is in small groups, or perhaps there is one person presenting to a large audience. Video conferencing systems can be designed to best support a specific meeting structure.

Engineered Environments Video conferencing systems are often times embedded into an environment. Sometimes the environment already exists and the system must be designed to integrate into these environments. Among high-end video conferencing systems, the designer may have control over the environment as well, being able to create the entire room so that each site is identical to the others.

Nonverbal Cues/Gaze Awareness One of the major goals of video conferencing is to capture and present the rich language of nonverbal communication. However, as I will demonstrate, video conferencing systems severely distort critical nonverbal cues. Of particular interest in prior research is the difficultly of conveying gaze information to remote participants. In this work, I will consider nonverbal cues more generally.

There is much research into improving all dimensions of video conferencing. The work presented here focuses on how video conferencing systems distort nonverbal cues

and affect communication. I will introduce a new design to improve the support of these nonverbal cues. I analyze and experimentally measure how distortions of nonverbal cues affects communication and introduce a new design which features life-sized images and preserves many nonverbal cues, like eye contact, in group-to-group meetings.

Chapter 3

Spatial Faithfulness

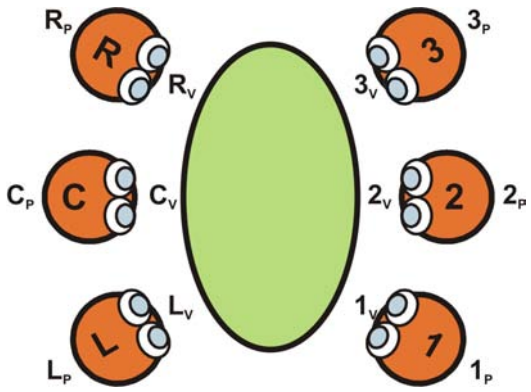
Video conferencing systems add a visual channel to remotely communicating people in an effort to capture the language of nonverbal communication. Because of the realistic imagery presented to the participants, the expectation is that this visual channel will provide an experience similar to that of a face-to-face meeting between the geographically separated participants [13]. However, at odds with expectations, regular users of video conferencing systems know that these meetings, though quite effective for certain tasks, rarely compare. While many aspects may affect the sense of face-to-face meetings, I explore the effects of differences in the visual channel between those meeting face-to-face and those meeting through video. In this section, I will outline the visual experiences of a video conference participant, introduce a model that will help explicate these experiences, and formalize the problems I am trying to solve in the design of a video conferencing system.

3.1 Observational Experiences

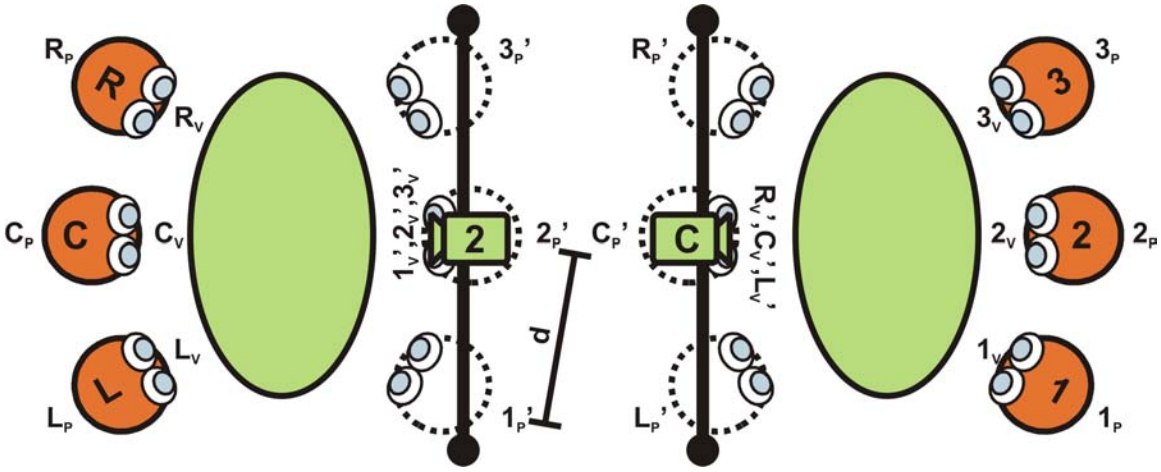
Nonverbal cues provide a rich language for communication. In this section, I will consider an example of what a participant may experience in face-to-face meetings and compare it to meetings in a video conferencing system. Let us first consider a group of six participants meeting face-to-face as shown in Figure 3.1(a). Suppose Participant C asks,

"How are you?"

Here, the sentence itself – the verbal communication – is not enough for interpretation. For instance, there is ambiguity as to who the intended recipient is. To



(a) An example structure of a face-to-face meeting between two groups of three.



(b) An example structure of a video conferencing meeting between the same two groups of three.

Figure 3.1: Two different meeting structures. (a) A face to face meeting and (b) and meeting using a video conferencing system.

resolve this ambiguity, each of the five remaining participants can easily turn to visual information, determine the gaze direction of the speaker, and quickly arrive at a shared understanding of the intended recipient of the question.

Absent of a visual channel – for instance, if all six participants were on a conference call – the speaker must be more explicit and encode the intended recipient in the question by including the name:

"How are you, 1?"

Now let us consider what happens to the same interaction among the same participants, but suppose they are meeting through a video conferencing system in two groups of three instead of face-to-face, as shown in Figure 3.1(b). For the purposes of this example, let us assume that three participants sit on one side facing a large display showing life-sized images of the remote participants. The video is captured by one fixed camera placed directly above the center of the remote display.

Given the realism of the image he sees of the remote participants, Participant C may – naturally and without much cognitive effort – form his sentence assuming he has the visual communication channel at his disposal. As a result, he may look directly at Participant 1 and ask, as he would if they were meeting face-to-face,

"How are you?"

Unfortunately, Participant 1 – as well as Participants 2 and 3 – see an image that suggests a message that is incongruent with the message being sent by Participant C. Specifically, because of the design of the video conferencing system, each Participant sees an image that suggests that Participant C is speaking to the person directly to the left: Participant 3 will believe Participant 2 is being addressed, Participant 2 will believe Participant 1 is being addressed, and Participant 1 will believe that he is looking off to the left somewhere. Since no one actually perceives himself as the recipient of Participant C's gaze, no one is prompted to respond. The image each local participant sees in this scenario is shown in Figure 3.2.

To compensate for this, the seasoned video conference user may find themselves reverting back to statements that disregard the use of the visual channel:

"How are you, 1?"

As shown by the example of face-to-face meetings in comparison with the audio-only meeting, the visual channel can be used to send and receive information needed to interpret a message. This example illustrates how the visual channel can be used to inform the group of the intended recipient of a message. As also shown, haphazardly adding a video channel that provides some semblance of a visual channel does not result in effective transmission of nonverbal cues. In this example, a distortion of gaze was introduced by the system, making that particular cue irrelevant and even misleading. New users of video conferencing systems may find these distorted nonverbal cues unnatural and disturbing. Seasoned video conferencing users may find themselves neglecting the visual information for certain tasks to the chagrin of those who designed or purchased the systems. In the following section, I present a structural analysis of what distortions to expect from video conferencing systems and why they exist.

Figure 3.2: These three figures show what each of the local participant in the meeting sees when viewing a standard display from each of their respective positions when all remote participants are looking at their rightmost remote partner. As can be seen, each local participant perceives that all remote partners are gazing one position to their left regardless of their actual seating position.

3.2 An Attention-Based User Model

I begin the analysis by introducing an attention-based model. Up until this point, the fixation of most video conferencing research has been on gaze behavior and eye contact in a setting where one remote person meets one other remote person. In other words, a dyadic meeting structure. By providing an attention-based model, I hope to accomplish two things. First, expand the scope of interest to include, not only gaze, but also any nonverbal cue that is generated by a participant. This can include a number of cues including pointing and body posture. At the same time, the scope expands beyond dyadic meetings to group meetings. Second, provide a vocabulary for discussion.

In this model there is an *attention source*, *attention target*, and *observer*.

Attention Source – a person who provides attention to the attention target. The method of attention can manifest itself in many different ways including, but not limited to, visual, gestural, positional, and directional.

Attention Target – a person or thing that receives attention from the source.

Observer – the person charged with understanding the presented information about attention – its source, its target, and any attached meaning.

Two common terms used in the gaze research community are *observer* and *looker*. Observer is used in the same way it is used here, but looker is a special case of an attention source where the type of attention is limited specifically to gaze information. Similarly, a *pointer* can be defined which would be an attention source who uses the gestural cue of pointing.

In analyzing video conferencing systems, it has been helpful to characterize the different types of *gaze information* that such systems can support. The literature uses the following definitions widely. Following Monk and Gale [22]:

Mutual Gaze Awareness knowing whether someone is looking at you. Also known as "eye contact"

Partial Gaze Awareness knowing the general direction someone is looking: up, down, left, or right.

Full Gaze Awareness ability to gauge the current object of someone else's visual attention.

However, these definitions leave some behavior around video conferencing systems undefined. For instance, consider a standard desktop setup where the camera is placed above the display. Because the camera is placed above the image of the remote participant, when the local participant looks in the eyes of the image, the remote participant will perceive downward directed gaze. Dourish et al. observed that with the initial use of this type of setup, users first obliged the remote user by looking into the camera to simulate a sense of direct eye contact for the remote user, but then re-adapted to looking at their partner's face as their understanding of the visual cues evolved. Once this level of understanding is achieved, the partners can accurately judge whether their partner is engaging in eye contact or not [14]. According to Monk's definition, these two partners know when their partner is looking at them but only through interpretation versus an automatic sensation. Can this system now be classified as supporting mutual gaze awareness or is the actual sensation of eye contact required?

The above issue demonstrates that a better understanding of the effects of the sensation of eye contact versus the knowledge of eye contact is required. I take the stance that the sensation of eye contact is more important than interpreted understanding. Non-verbal communication can function beyond any knowledge of it actually occurring – much non-verbal communication is neither consciously regulated nor consciously received, though its effects are certainly observable. Ekman demonstrates that facial movements and body gestures often occur without conscious thought, but can be reliable predictors of lying [16]. It has even been suggested that there are specialized brain functions for gaze detection [25].

In defining spatial faithfulness, I seek to both generalize from gaze awareness to attention cues as well as resolve the ambiguity between sensation versus interpreted knowledge.

Mutual Spatial Faithfulness A system is said to be mutually spatially faithful if, when the observer or some part of the observer is the object of interest, (a) it appears to the observer that, when that object is the attention target, it actually is the attention target, (b) it appears to the observer that, when that object is not the attention target,

the object actually is not the attention target, and (c) that this is simultaneously true for each participant involved in the meeting.

Partial Spatial Faithfulness A system is said to be partially spatially faithful if it provides a one-to-one mapping between the apparent direction – up, down, left, or right – of the attention target as seen by the observer and the actual direction of the attention target.

Full Spatial Faithfulness A system is said to be fully spatially faithful if it provides a one-to-one mapping between the apparent attention target and the actual attention target.

In characterizing the support of spatial faithfulness for video conferencing systems, let us consider the notion of *simultaneity*. Consider a dyadic meeting of two people, X and Y. As previously discussed, it is possible for one member to simulate for her partner the sensation of eye contact by looking directly into the camera, but this mechanism is asynchronous since she cannot see whether her partner is engaging in eye contact with her. Simultaneity is supported if a system allows both partners in the dyad to synchronously engage in eye contact. Simultaneity can apply to meetings of more than two members. Additionally, a system can be spatially faithful with respect to a certain type of attention. Most common are systems that explicitly support some level of spatial faithfulness for only gaze and not gesture. This is true for GAZE-2 [31] which I will describe in Section 4.

3.3 Perceptual Changes in Video Conferencing

In this section, I show that distortions in video conferencing are a result of the structure of a video conferencing system as well as the way people perceive images presented to them. I will present a phenomenon known as the *perspective invariance* and use it to describe how it introduces distortions in a video conferencing system.

3.3.1 Perspective Invariance

Perspective invariance is the cognitive phenomenon that allows viewers to view an image from a wide range of viewing angles and still form, automatically, an acceptable impression of the scene from the perspective it was captured. Our ability to view and

correctly form an impression is quite robust, making items like photographs, paintings, computer displays, and televisions powerful everyday tools. Unfortunately, for video conferencing, this exact ability causes spatial distortions for those hoping for an experience that rivals face-to-face communication.

To illustrate the phenomenon, let us consider what Leonardo da Vinci saw when painting Mona Lisa, replicated in Figure 3.3. His retinal image, and the perspective he consequently captured, was a function of his geometric relationship to his model. If he was to move 50° to his left, the painting would take on a completely different perspective with Mona Lisa's gaze directed 50° to his right.

Now, let us consider what a viewer of the painting sees. Suppose the viewer maintains the same geometric relationship with the painting that Leonardo da Vinci had with his model. This particular position is known as the *center of projection* or CoP. The retinal image of the painting would be quite similar to the retinal image of the painter and the viewer, not surprisingly, would perceive a scene quite close to that perceived by the painter. However, if the viewer of the painting were to move 50° to her left from the center of projection, her perception of the scene, unlike Leonardo da Vinci's, remains unchanged. This occurs even though her position, and, consequently, her retinal image, has changed. She experiences the painting as if she remained at the center of projection. This is the phenomenon known as perspective invariance. Though perspective invariance is universal to all images, no example has been as prominent as Mona Lisa's gaze. As a result, perspective invariance is sometimes known as the *Mona Lisa Effect* [17].

There are many hypotheses as to the exact cognitive mechanics of perspective invariance. The *local slant hypotheses* suggests that viewers take a local slant measurement of the image surface for each point of interest and make an appropriate adjustment of the retinal image [33]. The *indiscriminability hypothesis* suggests that the changes in the retinal image as we view images from varying angles are below some noticeable threshold [12]. Other theories suggest that familiar shapes, such as faces and certain fixed geometries, allow the viewer to appropriately interpret slanted images [6]. *Pictorial-compensation* theories suggest that information in the image allow us to determine the center of projection and reinterpret the retinal image accordingly [1].

Whatever the reason, the phenomenon of perspective invariance has been well documented and results in the viewer taking on the perspective of the scene as if they were viewing from the center of projection – defined by the geometric relationship of the painter

or the camera with respect to the scene. The interpretation of the image from the center of projection happens regardless of the actual viewing angle. As I will show in the next section, this is a major source of spatial distortions in video conferencing.

3.3.2 Distortions in Video Conferencing

Perspective invariance has a profound effect on how we perceive remote participants when we consider the structure of video conferencing systems. It introduces several distortions that participants must overcome through additional effort. In this section, I consider a system like the one shown in Figure 3.1(b) where a group of three meets another group of three. I will provide a model – the *View-Presence Model* – to explain the distortions in video conferencing and outline three specific problems to be solved by design.

The example system shown in Figure 3.1(b) features a large display on which each of the three remote participants is shown. To introduce the View-Presence Model, I introduce the term *virtual presence position* that is the spatial position where each remote participant is represented. It is at the virtual presence position of a particular remote participant where local participants will look when they wish to engage in eye contact or point at when they are gesturing. The virtual presence position becomes the attention target for many interpersonal interactions between the partners. The virtual presence position of a remote participant can also provide a point of reference for the local participant when he refers to other objects or people. For instance, the local participant may point to the left or right of a virtual presence position. In Figure 3.1(b), the virtual presence positions of each of the participants are the corresponding dashed lines on the screens and labeled with subscript V.

To continue with the model, I will introduce the term *virtual viewing positions* which is the spatial position corresponding to the impression formed by the participant. In the example, a centered camera perches on top of the display. Because of perspective invariance, the local position of the camera defines the center of projection for the remote viewers of the display that, in turn, will define the virtual viewing position. Because of the structure of the video conferencing system, multiple participants are viewing the same video from this single camera. The direct result is that all remote participants share the same virtual viewing position defined by that camera and thus perceive the local scene as if they are viewing from the same position. Since the camera position is directly above the images

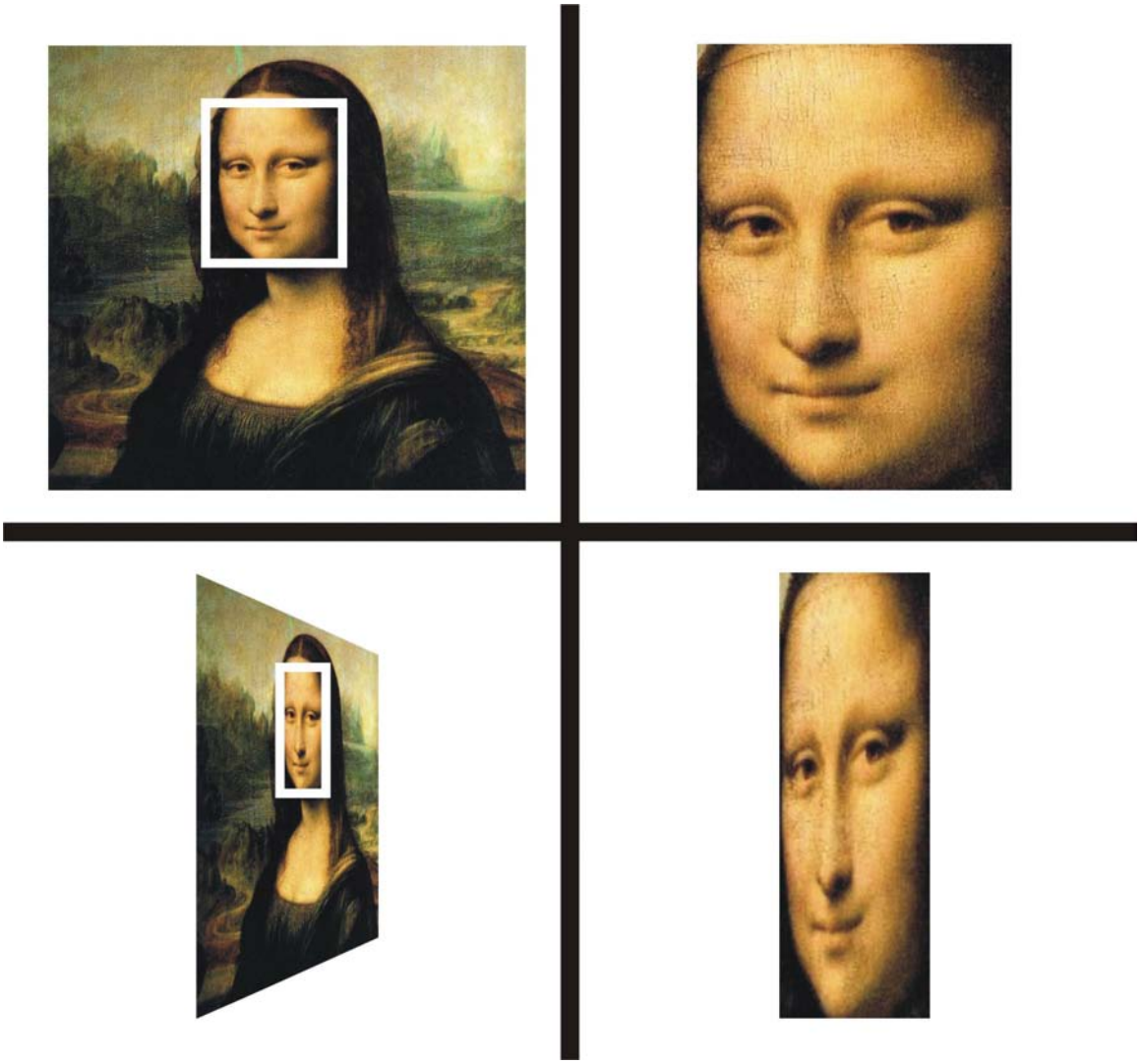


Figure 3.3: The Mona Lisa Effect: On the top left image is a frontal view of Mona Lisa. On the top right is a closeup of the face from this perspective. Notice how Mona Lisa seems to be looking directly at you. On the bottom left corner is an image of the Mona Lisa rotated as if it were viewed from 50° to your left. On the bottom right is a closeup of her face from this perspective. From the closeup, there is still a strong perception that Mona Lisa is still looking directly at you. The same sensation would be had if you turned this page 50° to your left and viewed the top left image. Humans have automatic cognitive functions that automatically adjust for rotations of 2-dimensional images allowing us to interpret this photo as if we were viewing it from the center of projection [33]. This perspective invariance allows the sensation Mona Lisa's gaze to dominate no matter where she is viewed from.

View-Presence Model

Term	*Definition*
Real Viewing Position	the spatial positions where each participant is viewing from.
Real Presence Position	the spatial position of the participants body.
Virtual Viewing Position	the spatial position corresponding to the impression formed by the remote participant.
Virtual Presence Position	the spatial position where each remote participant is represented.

Table 3.1: The View-Presence Model.

of Participant 2 and Participant C at their respective sites, the virtual viewing position for all participants will correspond to where Participant 2 and Participant C are seated.

For both virtual viewing and presence positions, there are also real viewing and presence positions. The real viewing position corresponds to the spatial positions where each participant is viewing. Real presence position is the spatial position of the participant's body. Real viewing and presence positions are usually the same position given the fact that our eyes are usually attached to our bodies.

In video conferencing system, both real and virtual positions exist for each of the participants. To define the challenges, I look at the relative geometric structure of these positions to see how they compare to face-to-face meeting. I then define three distortions to overcome in order to achieve full spatial faithfulness.

In face-to-face meetings, where only real viewing and presence positions exist, the real viewing position and the real presence position coincide by virtue of the fact that the eyes are attached to the body for each participant. Figure 3.1(a) illustrates a face-to-face meeting where real viewing – R_V, C_V, L_V, 1_V, 2_V, and 3_V – and real presence – R_P, C_P, L_P, 1_P, 2_P, and 3_P – positions are marked. For each participant, the coinciding viewing and presence positions form a round-table meeting structure.

In a video conferencing meeting, virtual positions are introduced and the structural relationships begin to look different from face-to-face meetings. Because of the nature of displays, it is possible to maintain the geometric relationships between real and virtual *presence positions*. In the example shown in Figure 3.1(b), there are two sites and the

system attempts to preserve the same meeting structure. Each site has a set of real presence positions and a set of virtual presence positions. On one side are the real presence positions of the local participants – R_P, C_P, and L_P – with the virtual presence positions of the remote participants – $1'_P$, $2'_P$, and $3'_P$. As can be seen, the relative structures between the real and virtual presence positions are congruent to the relative structure of real presence positions in the face-to-face meeting of Figure 3.1(a). The same can be said about the other site in the video conferencing meeting.

However, the geometric relationships between real and virtual *viewing positions* show incongruence. By virtue of perspective invariance, all remote viewers of a local scene will take on a perspective as if they were viewing from the center of projection defined by the location of the camera. Figure 3.1(b) shows that on one side, the local participants maintain their real viewing positions – R_V, C_V, and L_V. However, the remote participants all share the same virtual viewing position – $1'_V$, $2'_V$, and $3'_V$ – leading to a different structure than in the face-to-face condition. In the example introduced at the beginning of this chapter, where Participant C was looking toward Participant 1, it is now clear why all the remote participants would share the same perception that Participant C was gazing one person to his left. Another distinguishing event occurs when Participant C looks toward Participant 2. In this example, the position of the camera happens to correspond with Participant 2's image. Because of the shared virtual viewing position, all remote participants will perceive direct eye contact simultaneously which is an event that is impossible in face-to-face meetings. I call this first distortion the *Collapsed Viewer Effect*.

Related, but not identical, is a second problem because of perspective invariance. In face-to-face meetings, the real viewing positions and the real presence positions are tightly coupled for each participant. However, in a group-to-group video conferencing, there will always be a horizontal displacement between the virtual presence position and the virtual viewing position of at least one participant. In any group structure, two people will always occupy two positions while a single camera can only occupy one. This results in a decoupling of the virtual viewing and virtual presence positions. Because of this displacement, when the attention source produces a cue toward the virtual presence position of the attention target, a transformation will take place based on the horizontal displacement. For instance, continuing with this example, when Participant C looks toward Participant 1, Participant C looks directly at the virtual presence position of Participant 1. However, because the virtual viewing position for Participant 1 is displaced from the virtual presence position

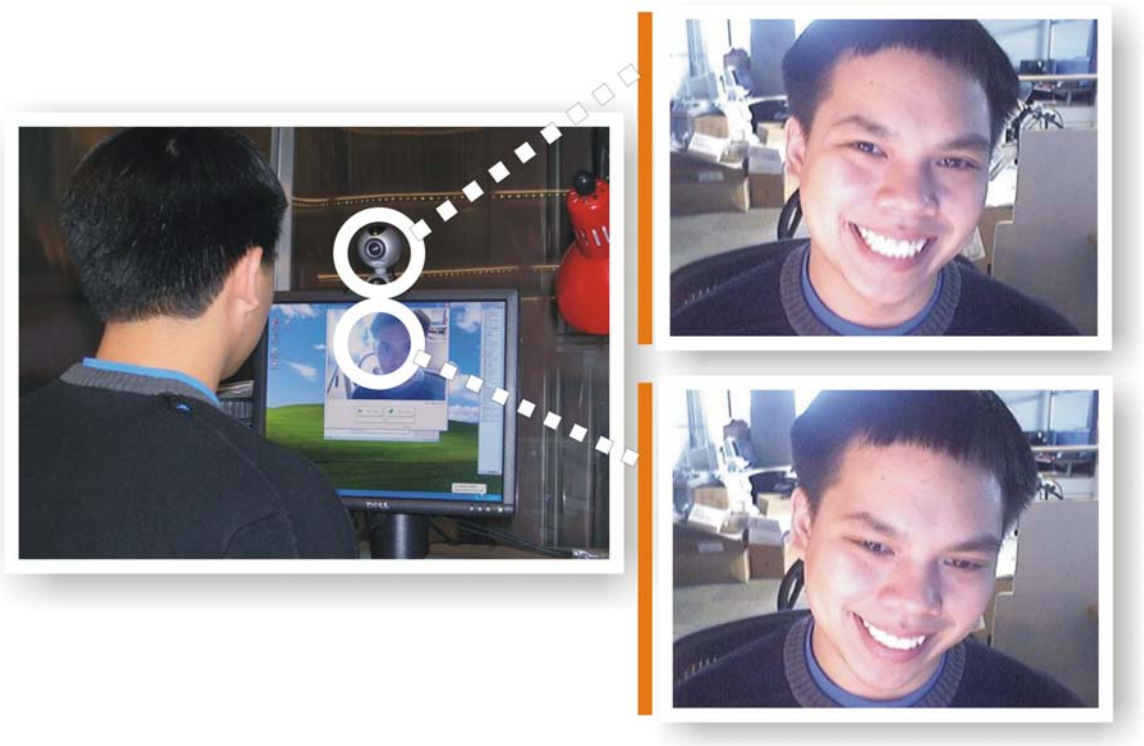


Figure 3.4: A demonstration of vertical parallax with gaze. When the local participant looks at the image of the remote participant in the eyes, the remote participant sees an image which suggests they are being looked down upon because of the displacement between the camera and the image. The local participant can simulate direct eye contact for the remote participant by looking directly into the camera, but now the local participant is forced to look at the camera. This, ironically, forces him to miss out on the visual information that was the purposes of video conferencing in the first place.

by one position, marked as d in Figure 3.1(b), Participant 1 would see gaze directed one position away instead of the intended direct eye contact. I call this second distortion the *Horizontal Parallax Effect*.

So far, only horizontal distortions in video conferencing systems have been considered. Additionally, there are vertical effects that must be address. For practical reasons, cameras are often times perched on top of displays producing a vertical displacement in addition to the horizontal displacement described above. Due to perspective invariance, this produces a virtual viewing position that is above the virtual presence position of each of the participants leading to the perception that the remote participants may be looking downward at them. This downward looking gaze is often times loaded

with various meaning depending on the context. Figure 3.4 illustrates this effect with a desktop video conferencing system. The camera is perched on top of the display. When the participant looks into the image of their remote partner, their partner sees an image of the participant looking downward. A sense of eye contact can be simulated for the benefit of the remote participant if the local participant looks directly into the camera, but then the local participant no longer has access to the visual. Ironically, this is the purpose of video conferencing in the first place. I call this last distortion the *Vertical Parallax Effect*. The keen reader may notice that horizontal and vertical parallax effects are structurally the same problem, but they are separated in my analysis because a different solution is used for each.

In this section, I presented *perspective invariance* and the *View-Presence Model* allowing us to analyze and articulate the structural differences between face-to-face meetings and video conferencing meetings. I used this analysis to outline three major distortions to overcome by design: (1) the Collapsed Viewer Effect, (2) the Horizontal Parallax Effect, and (3) the Vertical Parallax Effect. I will revisit these distortions in Part II and introduce a design that mitigates these distortions. Next, I will present an overview of how spatial information is used in communication and how video conferencing can affect it.

3.4 Effects of Video In Communication

Gaze has a critical role in group communication. According to Kendon [20], its functions include turn-taking, eliciting, and suppressing communication, monitoring, conveying cognitive activity, and expressing involvement. By removing or distorting gaze perception, we risk adversely affecting the processes of communication that depend on these functions. For instance, Vertegaal et al. [31] found that participants took 25% fewer turns when eye contact was not conveyed in a three-person meeting.

However, an arbitrarily added video channel will not necessarily result in better communication. Connell et al. [9] found that audio alone might be, in fact, preferable in routine business communication. Bos et al. [5] measured the effects of four different mediated channels – face-to-face, text, audio only, and video and audio – on trust building. They found that adding video did not significantly contribute to trust building when compared to audio-only channels in people who have not met face-to-face.

Furthermore, Short et al. [30] notes that a video channel may actually disrupt some communication processes when compared to audio only channels. For instance, the lack of mutual eye contact can lead one participant to feel like she is making eye contact with a remote participant when the other does not, leading to an asymmetry in the understanding of the shared context. Argyle et al. [2] found that such asymmetries lead to noticeable increases in pause length and interruptions.

Chapter 4

Prior Work

The problem with video conferencing systems and nonverbal cues is not new and has a long history of research which tries to mitigate the effects of these distortions. In this section, I review several systems which try to restore nonverbal cues, mostly gaze, using a variety of novel techniques.

Hydra [29], shown in Figure 4.1, supports multi-party conferencing by providing a camera/display surrogate for each remote participant in the meeting. This surrogate occupies the space that would otherwise be occupied by the corresponding participant. Because of the scale and setup of a Hydra site, there is still a noticeable discrepancy between the camera and the image of the eyes, resulting in the same lack of support for mutual gaze awareness that standard desktop setups have. Hydra does add an element of mutual spatial faithfulness in that it appears to an observer that she is being looked at when she is indeed the attention target and not being looked at when she is not the attention target in group meetings.

GAZE-2, shown in Figure 4.2, is another system developed to support gaze awareness in group video conferencing. GAZE-2 uses an eye tracking system that selects from an array of cameras the one the participant is looking directly at to capture a frontal facial view [32]. This view is presented to the remote user that the participant is looking at, so that these two experience realistic eye contact. The other participants in the group meeting see this *frontal planar* image in a 3D space rotated toward the image of the person being looked at. Even with significant rotations of frontal views, the images will still be perceived as frontal ones, while a side view of those participants is what is desired. To mitigate this, GAZE-2 blurs the image, attaches them to a 3D box, and rotates the image

Figure 4.1: Hydra

by 70° or more and to create a spatial perception that overwhelms the perception of the face itself. This distortion is not spatially faithful, and there is no attempt to preserve gesture or relations with objects in the space.

MAJIC, shown in Figure 4.3, produces a parallax-free image by placing cameras behind the image of the eyes using a semi-transparent screen [23]. MAJIC supports mutual, partial, and full spatial faithfulness since the images are free of parallax, so long as there is only one participant at each site since they employ single view displays.

An extreme approach to preserving spatiality is to use a mobile robotic avatar such as Personal Roving Presence, or PRoP, as a proxy for a single remote user [24]. Shown in Figure 4.4, PRoPs suffer from the Mona-Lisa effect at both ends, but are not intended for group-to-group interaction. At the robot end, they mitigate the effect by using the robot's body and camera as a gaze cue like GAZE-2's virtual monitors. When multiple users operate PRoPs in a shared physical space, full spatial faithfulness is preserved.

All the above systems claim to support multi-site meetings. A striking limitation on all these systems, however, is that they only work correctly and provide their claimed affordances when used with *one participant per site*. This will be a problem with any system based on viewer-independent displays. In real physical space, different users *do not share*

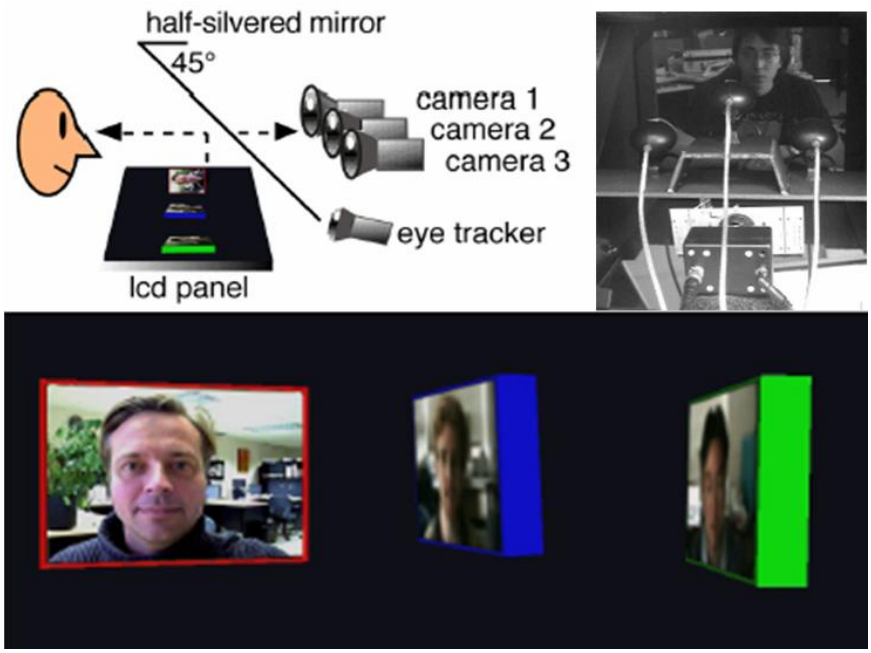


Figure 4.2: GAZE-2

the same view with others. MultiView provides a practical solution to this problem, using a custom view-dependent display.

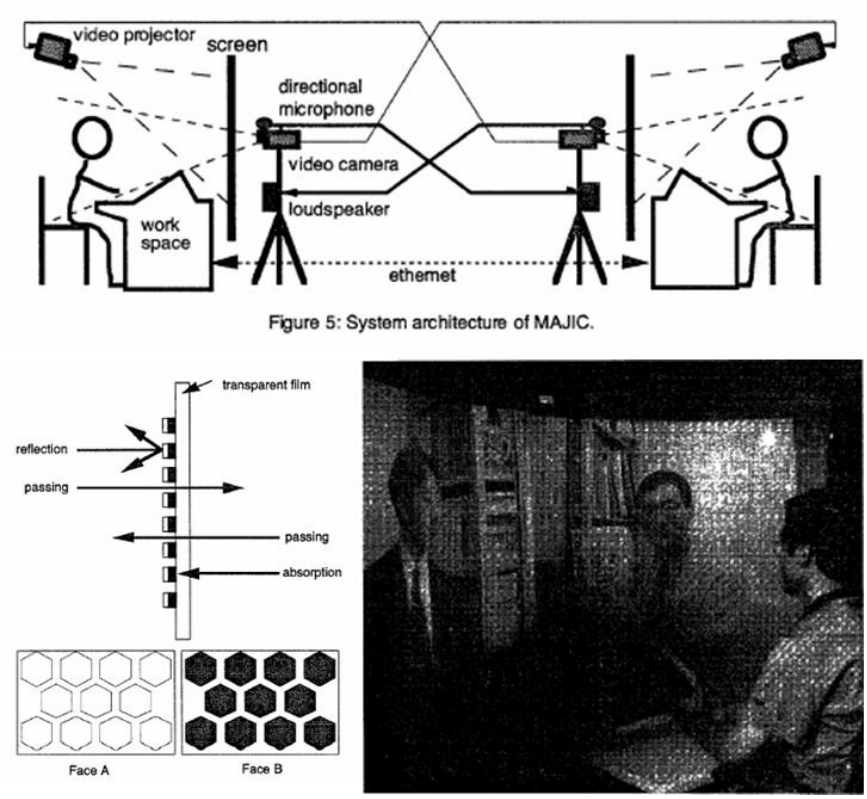


Figure 4.3: MAJIC

Figure 4.4: PRoP

Part II

Design

Chapter 5

Many-to-Many Video Conferencing

The design space of video conferencing systems is very rich. Industry and researchers alike have taken advantage of the multitude of ways we can incorporate video into communication to produce a wide range of innovative products allowing people to communicate more effectively. In this chapter, I explore yet another section of this rich space and present a new design that supports meetings between groups of people. The MultiView design understands the way people perceive images, and how this presents multiple challenges to designing video conferencing systems. It is sensitive to the fact that when people see faces and bodies, our perceptual system is primed for certain expectations [13]. MultiView is designed to support a fully spatially faithful meeting.

The MultiView design, shown in Figure 5.1 for a meeting between two groups of three, deviates from the design of standard video conferencing systems in two major ways. The first major distinction is the introduction of several carefully arranged cameras. There is one remote camera for each local participant. Each camera is placed on the screen directly above each of the images of the local participants and captures a unique perspective of the entire remote scene for each of them. Because each camera captures the entire remote scene, each will be seeing the same things, but from slightly different perspectives. In this example – where each site has three local participants – there are three remote cameras and each camera captures all three remote participants, just from slightly different perspectives.

The second major distinction is the introduction of a multiple viewpoint display. This is a special class of displays that is capable of supporting multiple viewers, allowing

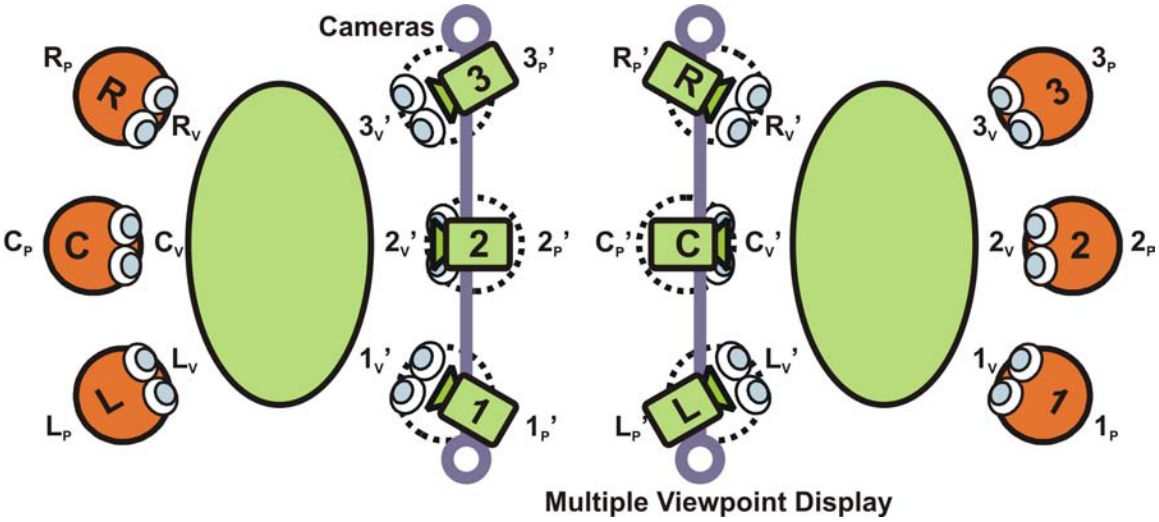


Figure 5.1: A diagram of the MultiView Video Conferencing System.

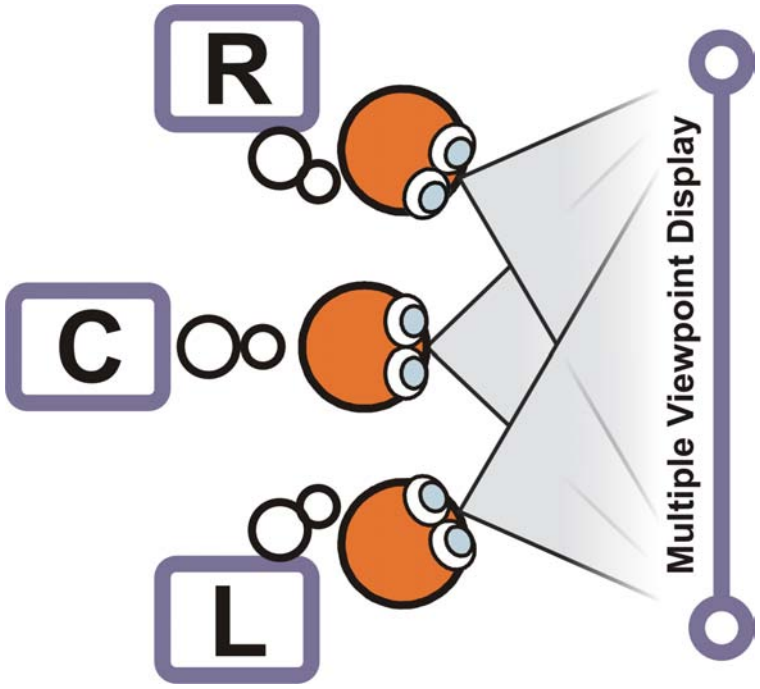


Figure 5.2: A diagram of a multiple-viewpoint display. Using these displays, several viewers can be looking at the same display, but each see a completely different images without any knowledge of what the other viewers see. In this example, one participant sees 'R', another 'C', and the last 'L'.

Challenge	*Definition*
Collapsed Viewer Effect	The effect where all remote participants share the same virtual viewing position of the local scene.
Horizontal Parallax Effect	The effect where there is a horizontal displacement between the perceived attention target and the actual attention target because of a horizontal displacement between the virtual viewing position and the virtual presence position of a remote participant.
Vertical Parallax Effect	The effect where there is a vertical displacement between the perceived attention target and the actual attention target because of a vertical displacement between the virtual viewing position and the virtual presence position.

Table 5.1: Spatial Distortions in Video Conferencing

each viewer to look at the same display, but see a completely different image[1]. The high-level behavior of multiple-viewpoint displays is illustrated in Figure 5.2. Though multiple-viewpoint displays exist as products, no displays fit the requirements needed for effective video conferencing. The displays were too small, the resolution was too low, the image brightness was too low, it required wearing special glasses, or some combination of these reasons. As part of the design process, I produced a display that would meet the needs of the design. The details of this design will be saved for Chapter 7. Using this display, each participant views their own video stream from the camera that is above their image at the remote site. This setup that addresses the three distortions I defined in Section 3.3.2 and summarized in Table 5.1.

In standard video conferencing systems, the *Collapsed Viewer Effect* is a result of perspective invariance paired with the sharing of a standard display driven by a single camera by all the participants. MultiView overcomes this distortion by providing each participant with his or her own camera. This allows each participant to have their own

[1]Multiple-viewpoint displays are also sometimes known as "his and her televisions" because they allow him to watch his show while allowing her to what something else with neither of the pair aware of what the other is watching.

unique virtual viewing position of the local scene so that they no longer need to share this position with all remote participants.

The second distortion – the *Horizontal Parallax Effect* – is caused by a decoupling of the virtual viewing and virtual presence positions and an introduction of a horizontal displacement between the two. This is necessary in standard video conferencing systems because only one virtual viewing position could be provided regardless of the number of remote participants. With MultiView, each remote participant can now have his or her own unique virtual viewing position. By carefully placing the cameras directly above the image of each remote participant, virtual viewing positions and virtual presence positions can be recoupled, reproducing the structure seen in face-to-face meetings. Let us consider one video conferencing site of the example shown in Figure 5.2. In standard video conferencing, and illustrated in Figure 3.1(b), there is incongruence between the structures of the virtual viewing positions when compared to the real viewing positions of a face-to-face meeting. In MultiView, each remote participant is provided with a unique virtual viewing position that is recoupled with the respective virtual presence position, faithfully reproducing the structure seen in face-to-face meetings. Figure 5.1 marks each of the real and virtual viewing and presence positions.

The third distortion – the *Vertical Parallax Effect* – is caused by a vertical displacement of the cameras and the image of the eyes. This is necessary in standard video conferencing systems because if the cameras were to be at the same position as the eyes, either the screen or the camera would be occluded. To resolve this issue, I take a close look at how sensitive we are to distortions in gaze. As it turns out, remote partners will reliably perceive direct eye contact over video even when their partner is looking up to 5° in the downward direction [8]. This means that as long as the camera is placed above the screen and the vertical displacement between the camera and the image of the remote participant's eyes is less than 5° with respect to the local participant's eyes, the remote participant will still reliably perceive direct eye contact. Figure 5.3 illustrates the necessary relationship between the image of the remote participant, the local participant, and the camera. Because of the scale of MultiView, the displacement between the cameras and the image of the eyes rarely exceeds 5° so breaks in eye contact because of vertical parallax should not be observed. This is not necessarily true for desktop systems. The reason for this is, even though the displacement between the camera and the image may be less than in MultiView, the viewer is closer resulting in an angle that is greater than 5°.

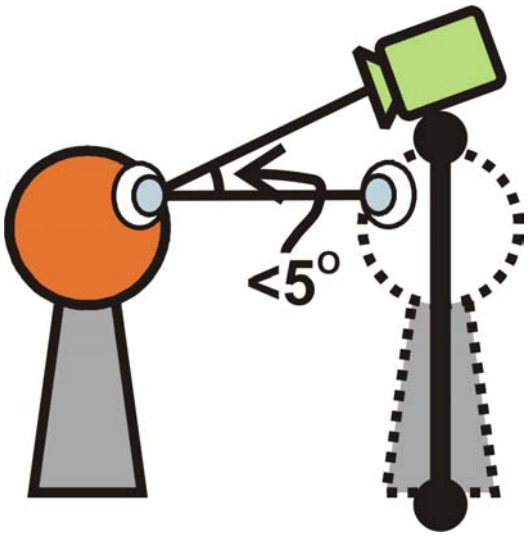


Figure 5.3: As long as the camera is placed above the image, and the angle defined by the image of the eyes of the remote partners, eyes of the local partner, and the camera is less than 5°, than the remote partner will still reliably perceive direct eye contact.

The basic MultiView design deviates from standard video conferencing systems in two major ways. First, I introduce a camera for each remote participant taking part in the meeting. Second, I introduce a multiple viewpoint display. With the appropriate configuration, these two pieces come together to solve the three spatial distortions I outlined in Section 3.3.2. In Chapter 7, I will cover the details of implementing a multiple viewpoint display, but first I will cover some the design iterations before reaching the current implementation.

Chapter 6

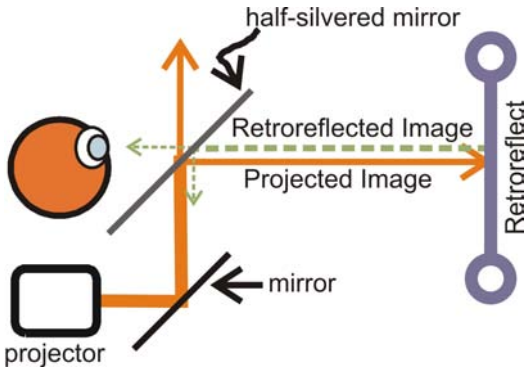
Design Implementations

As part of the design process, I focused on iterative design. I implemented several systems based on the basic design ideas presented in Chapter 5. Collecting and learning from each design through personal experience and user experimentation, new iterations were informed by the previous ones. In this section, I present three of the major design iterations.

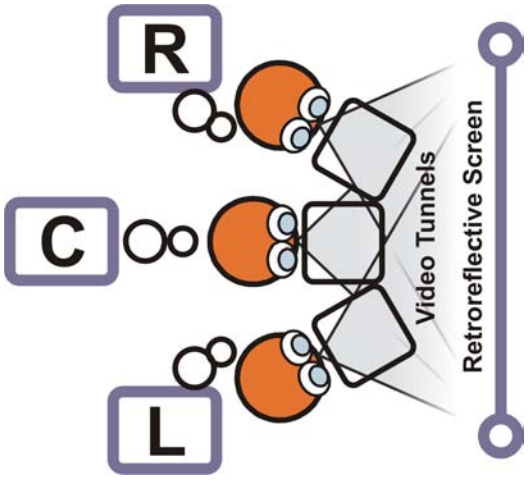
6.1 Video Tunnel Technique

In the first approach to designing MultiView, I adopted a "video tunnel" technique. In this approach, a projector projects onto a retroreflective surface. Retoreflectors have the property that, ideally, all the light that hits it is reflected back in the direction it came from. The problem is that in order to see the image from a projector, one would have to be in the same physical position as the projector which is impossible. To get around this, I built a video tunnel as diagrammed in Figure 6.1(a).

The video tunnel works using a system of mirrors and half-silvered mirrors to put the eyes in line with the retro-reflection of the image. The projector projects onto the bottom mirror which reflects it upward toward the half-silvered mirror as shown in Figure 6.1(c). Some of the light continues upward while the rest of the light reflects toward the retroreflective screen as shown in Figure 6.1(d). The retroreflective screen then retroreflects the light directly back at the half-silvered mirror. Some of the retroreflected light is reflected downward while the rest of it continues to the viewer's eyes.



(a) A diagram of the video tunnel in operation.



(b) A diagram of how the video tunnel fits into a video conferencing implementation.

(c) The video tunnel with projector, mirror, and half silvered mirror. The entire apparatus was held together with clear acrylic.

(d) The retroreflective screen onto which the image was projected.

Figure 6.1: The video tunnel based implementation of MultiView.

I used the video tunnel video conferencing system as shown in Figure 6.1(b). Though the video tunnel proved useful for implementing a multiple-viewpoint display, it was quickly apparent that it would not be feasible for video conferencing. The main problem is that the video tunnels, placed in front of each participant, blocks the cameras' line of sight. Additionally, properties of the half-silvered glass severely affected the color quality of the image.

From this implementation, I learned that a multiple viewpoint display that does not rely on devices that block the cameras from capturing its image must be developed. This not only precludes video-tunnel approaches, but any approach that requires wearing glasses such an anaglyph or polarized approaches.

6.2 Diffused Retroreflection Technique (v1)

From experience with video tunnels, it is clear that a display was needed that allows the viewers see their own images while allowing the cameras to capture all of the viewers for the remote participants. I therefore developed a display that removes all dependencies on artifacts that may occlude the camera, which I call *diffused retroreflection*.

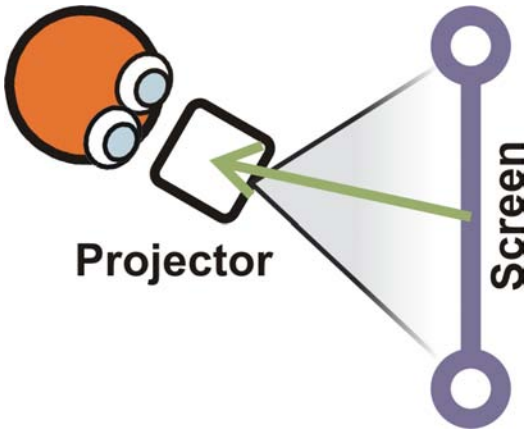
In diffused retroreflection, projectors project onto a screen which retroreflects in the horizontal direction but diffuses in the vertical direction. Desired diffusion behavior is illustrated for the horizontal and vertical directions in Figures Figure 6.2(a) and 6.2(b), respectively. The horizontal retroreflection allows a participant lined up with a projector to see the image coming from that projector while remaining unaware of the images produced by neighboring projectors. The vertical diffusion allows the viewer to be above or below the projector and still see the image. This is different from a simple retroreflector where the image is only viewable from the exact position of the light source creating a dependence on something like the video tunnel. The details of the implementation of this display will be presented in Section 7.1. A system based on this display is diagrammed in Figure 6.2(c). The implementation of this type of system is shown in Figure 6.2(d).

With this implementation of MultiView, participants sits at a desk with a projector in front of them and all view the same screen. However, each viewing participant will see a slightly different image from the other participants allowing each participant to arrive at a shared understanding of the meeting. Figure 6.3 demonstrates what each local participant would see if all remote participants were looking at the center participant.

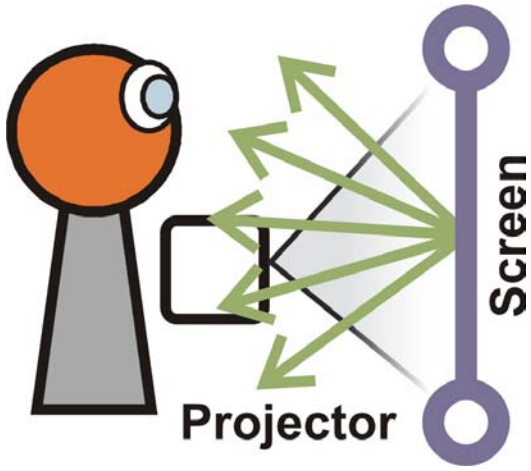
Because of available materials, the screen was 36"x48" large. However, this was not large enough to present life-sized images of the remote participants so the image was scaled down by 2/3. This scaling put the virtual participants a distance *behind* the plane of the screen. If combined with the 12' of physical distance between the screen and the participants, the total effective distance to the remote participants becomes 18'.

On this system, I ran a user study to test the functionality of the system as well as get feedback for design improvements. The details of this study are presented in Chapter 8, but the design lessons learned are summarized here.

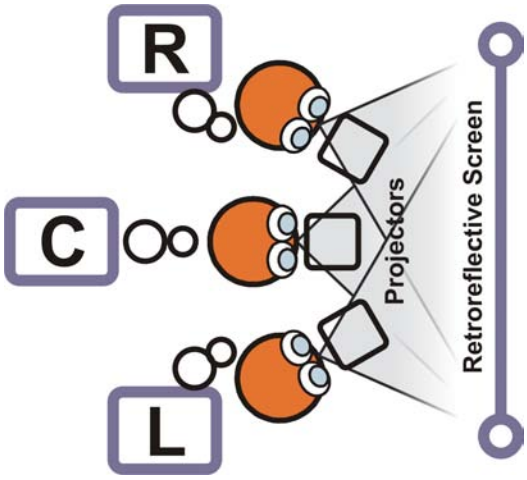
- Due to the throw distance of the projectors, the system separated the participants from the screen by 12'. This is much further apart than participants would sit if they were meeting face-to-face.
- The screen was too small to fit all three remote participants without scaling the image down. If the image scaling was taken into affect, the participants were virtually separated by 18'.
- The quality of the image was not yet good enough to allow participants to reliably resolve eye contact with each other.
- The setup of the system placed the projectors on the desk in front of each participant. The projectors produced an uncomfortable amount of fan noise as well as heat.



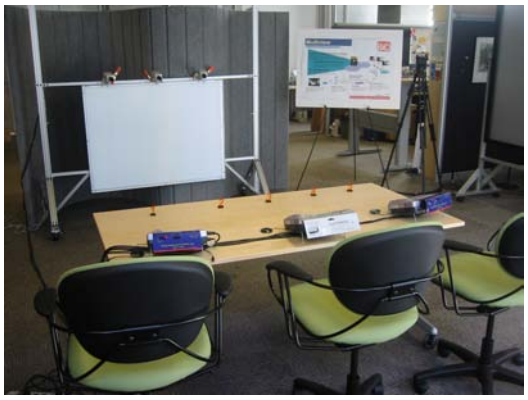
(a) A diagram of the horizontal retroreflection of the display.



(b) A diagram of the vertical diffusion of the display.



(c) A diagram of the MultiView system based on the diffused retroreflector display.



(d) A photograph of the MultiView system based based on this display

Figure 6.2: The MultiView system based on the diffused retroreflector display.

Figure 6.3: These three figures show what each of the local participant in the meeting sees when using MultiView (v1) from each of their respective positions when all remote participants are looking at the center remote partner. As can be seen, each participant sees a slightly different perspective, allowing each to arrive at the shared understanding that the center participant is being addressed.

6.3 Diffused Retroreflection Technique (v2)

Version 2 of the diffused retroreflection system is much like the previous iteration, but I took into account the lessons learned from the user study and implemented many changes. The current implementation is shown in Figure 6.4.

From the above findings, I developed the latest iteration of MultiView. Participants now sit in front of a conference table about 8' from the screen. Each viewing position is separated by 27" or 16° with respect to the screen.

The display has been improved in several respects. MultiView now features a larger, wider screen (72"W x 32"H, 9:4 aspect ratio) so images can be life-sized. Though the basic optical functionality is the same as in the previous iteration, higher precision optics were used in this iteration that greatly enhances the image quality. To complement the new screen, new short-throw XGA (1024x768 pixel) projectors allowed us to reduce the viewing distance from 18' to 8'. The new projectors are mounted above the participants, clearing the work surface, and directing heat and noise away from the participants.

To capture the images, new high-resolution (1024x768 pixel) firewire cameras replace QVGA CCTV (320x240) cameras of the previous iteration. Due to the mismatch between the screens 9:4 aspect ratios and the 4:3 aspect ratio of the projectors and cameras, the image is vertically higher than necessary. As a result, the lower 40% of the pixels was discarded in both the cameras and projector. Cameras are placed to minimize the vertical disparity between the cameras and the images of the eyes. The image of the eyes was generally 6" below the position of the cameras. Given that the participants are viewing the screen from 8', there will be about a 3.6° disparity between the actual gaze direction and the perceived gaze direction in the downward direction. However, even with this disparity, people should still register correct eye contact given that it is below the angular threshold beyond which people perceive a break in eye contact [8]. Figure 6.5 demonstrates what each local participant would see if each remote participant was looking at their rightmost remote partner comparing a view independent screen in the top row with a MultiView screen in the bottom row.

Sound is recorded using a single echo-canceling desktop conferencing microphone (ClearOne AccuMic PC). Speakers are mounted on the top of the screen.

All audio and video are encoded and decoded using MPEG-2 codecs. Each video and audio stream was encoded at 6Mbps constant bit rate and has been tested over both

Figure 6.4: The current MultiView implementation

local gigabit ethernet and with an Internet2 connection between Berkeley and Palo Alto sites.

The current implementation of the system shows dramatic improvement over the previous iteration in terms of user experience and video quality. It has provided an effective platform for user studies probing the effects of spatial faithfulness on communication. For instance, as I will cover in detail in Chapter 9, it has allowed us to test the effect of spatial faithfulness on trust formation in group-to-group meetings. However, much engineering work still needs to be done to improve image quality of the presented images. In the next chapter, I present the engineering design work for the current implementation of the screen and a possible avenue for future designs.

Figure 6.5: Three remote participants are gazing at viewing position 1 (Figure 5.1). Column 1 is the view from position 1, column 2, position 2, and column 3, position 3. The bottom row is what is seen using MultiView and shows appropriately changing perspectives. The top row, for comparison, is what is seen from the respective positions with non-directional video conferencing and demonstrates perspective invariance.

Chapter 7

Multiple Viewpoint Displays

Though multiple-viewpoint displays exist as a product, there was no particular product that met all of the design needs for video conferencing. Reasons included image size, image resolution, image brightness, or requiring the viewers to wear special goggles. Because of these reasons, I began the design of a specialized display more appropriate to the design needs of MultiView. In the previous chapter, I presented the design of MultiView assuming the existence of a multiple-viewpoint display. In this chapter, I present two different implementations of a display itself. The first one, diffused retroreflection, is the implementation used in the system at the time of this writing. The second one, a lenticular method, is a technique I am just beginning to explore and, therefore, have not yet been implemented into the MultiView system.

A multiple viewpoint display's main function is to display a different image to each viewer depending on the viewing position. A conventional screen will display the same image regardless of their viewing position. The design of this type of display is very similar to designing three-dimensional displays and thus has a long and rich history of engineering. In this section, I present details of a novel front-projection system I developed for the initial prototype based on a retroreflective element and present a second design based on lenticular imaging for future consideration.

7.1 Diffused Retroreflection Displays

In the current implementation of the MultiView multiple-viewpoint display, I design a custom front projection screen that carefully controls the direction of reflected

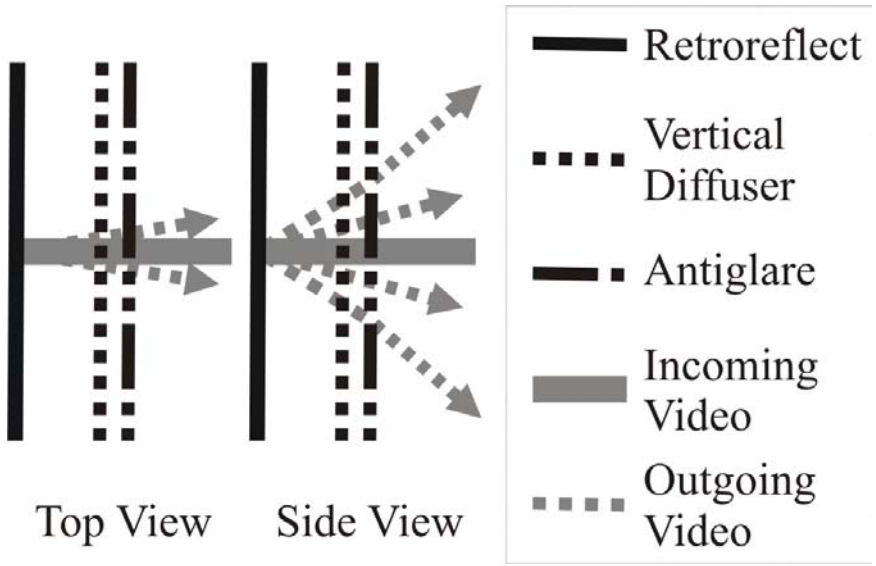
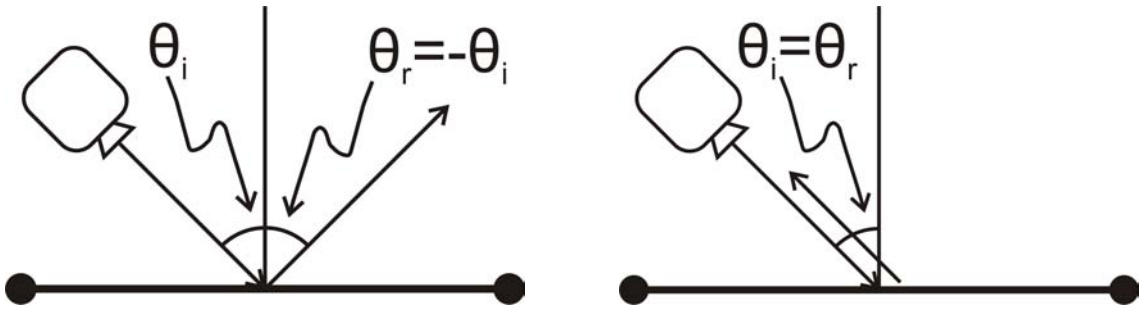


Figure 7.1: Slice views showing the multiple layers of the MultiView screen. The backmost layer is is a retroreflective sheet. The center layer is a vertical diffuser. The frontmost layer is an antiglare layer. The "Top View" shows a small amount of diffusion in the left and right directions. The "Side View" shows a large amount of diffusion in the up and down direction.

light. The system currently employs several projectors, each projecting a unique image on the entirety of the screen. The screen's optics carefully retroreflects the light in the direction of the projector but diffuses it vertically, allowing viewers to see the image from any position above or below the projector. This allows us to create viewing zones defined by the position of the projectors.

7.1.1 Implementation

The MultiView screen uses multiple layers to create its viewing zones. A diagram of layers is shown in Figure 7.1. The back-most layer is a retroreflective material. An ideal retroreflective material bounces all of the light back to its source: $\theta_r = \theta_i$. This differs from an ideal mirror where the light bounces along the reflective path: $\theta_r = -\theta_i$. Both reflection and retroreflection may exhibit a certain amount of diffusion based on the properties of the material. Additionally, materials can exhibit properties of a Lambertian surface that diffuses light in all directions equally. A practical retroreflective material exhibits all properties – given a source of light, some of the light bounces back to the source with some diffusion,



(a) Reflection with a perfect mirror.

(b) Retroreflection with a perfect retroreflector.

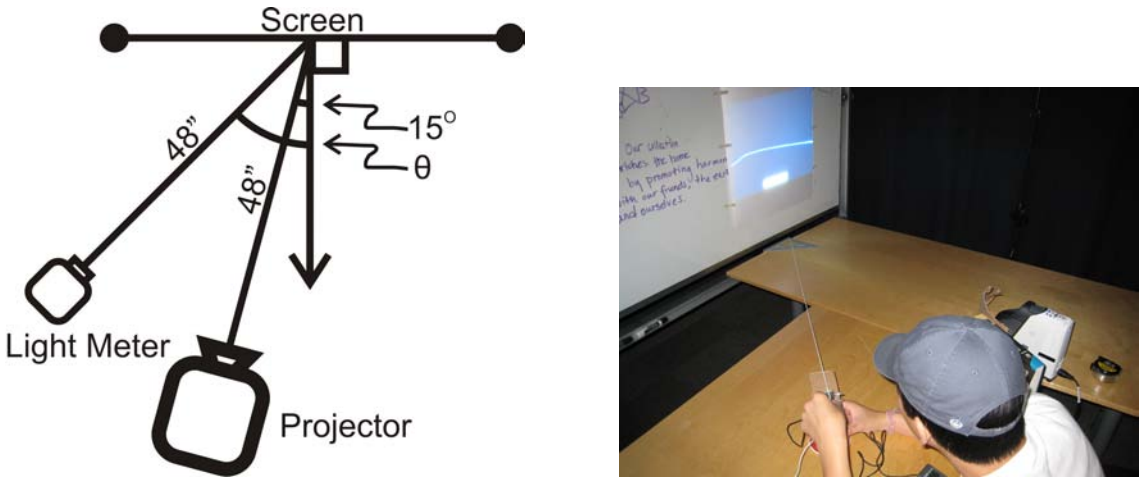
Figure 7.2: A comparison of reflection versus retroreflection. In reflection (a), the angle of reflection is equal the opposite of the angle of incidence. In retroreflection (b), the angle of reflection is equal the angle of incidence and light is sent directly back to the source.

some of the light gets reflected along the reflective path with some diffusion, and there will be constant diffusion across all angles. There are many retroreflective materials available in the market, but the Reflexite AC1000 was chosen because of its strong retroreflective specification.

The next layer is a one-dimensional diffuser that extends the viewing zone vertically. Without it, the image would only be visible directly on the projection axis. This is problematic because if a person were in front of the projector, she would block the projected image, and if she were behind it, the projector would block her view. In this implementation, a lenticular sheet was used as the diffuser[1]. The vertical diffuser actually amplifies some of the diffusion present in the retroreflected image. As a result, space is required between the retroreflective layer and the diffusive layer. A spacing of 1/4" or more between retroreflect and lenticular sheet is recommended, otherwise the diffusion effects of the lenticular will be undone by the retroreflect. It is possible to reduce this spacing if needed by using a lenticular sheet with finer pitch.

The last layer is an antiglare layer. The high gloss finish of the lenticular sheets produced a very distracting glare along the path of reflection. As a result, an antiglare film produced by DuPont (HEA2000 Gloss 110) was applied. The film has a pressure sensitive

[1]Note: Lenticular sheets are often used in directional displays for multiple image separation and have been used in this way in previous spatial displays. This often confuses readers trying to understand MultiView. The lenticular sheet was *not* used as a lenticular imager, but simply as a directional diffuser. Any other diffuser could be used, but others are currently much more expensive.



(a) A diagram of the experimental setup.

(b) A photograph of the experimental setup.

Figure 7.3: The experimental setup for measuring the screen diffusion profile.

adhesive (PSA) so a method similar to that used in applying window tints was used on the smooth side of the lenticular sheet.

7.1.2 Measurements

To measure the properties of the screen, an image was projected on the display and the brightness was measured from different viewing angles. The image was projected using a single Hewlett Packard mp3130w projector $-15°$ from the normal at a distance of 48”. A light meter was used to measure the illuminance of the resulting image from several different angles for each of the conditions. The distance between the light meter and the center of the screen was also kept at a constant 48”. The experimental setup is illustrated in Figure 7.3. This measurement provides a diffusion profile to characterize the properties of the screen.

7.1.3 Results and Discussion

The above measurement provides a diffusion profile shown in Figure 7.4. At the optimal viewing angle of 15°, the image brightness was 127 lux. As I move away from the optimal viewing angle, the brightness of the image quickly drops off. Sampled at $-40°$, the brightness was 1.6 lux.

Since the maximum brightness of the image is known, how bright a competing image needs to be before it is noticeable can be calculated[2]: 10.2 lux. Using this information, the size of the viewing zone is can be calculated. In the experiment, there is about 5.5° from the optimal viewing zone to the angle where the brightness of the image equals the calculation above. Thus, a competing image would need to be less than 5.5° away from the primary image in order to be noticed by the viewer. The image was a fixed blue screen and does not take into account the full range of color video.

[2]This value was calculated using Weber's Law which states that the ratio between the just noticeable difference, ΔI, and the current stimulus, I, is constant, k. Or in other words: $\frac{\Delta I}{I} = k$. For light instensity, $k = 0.08$

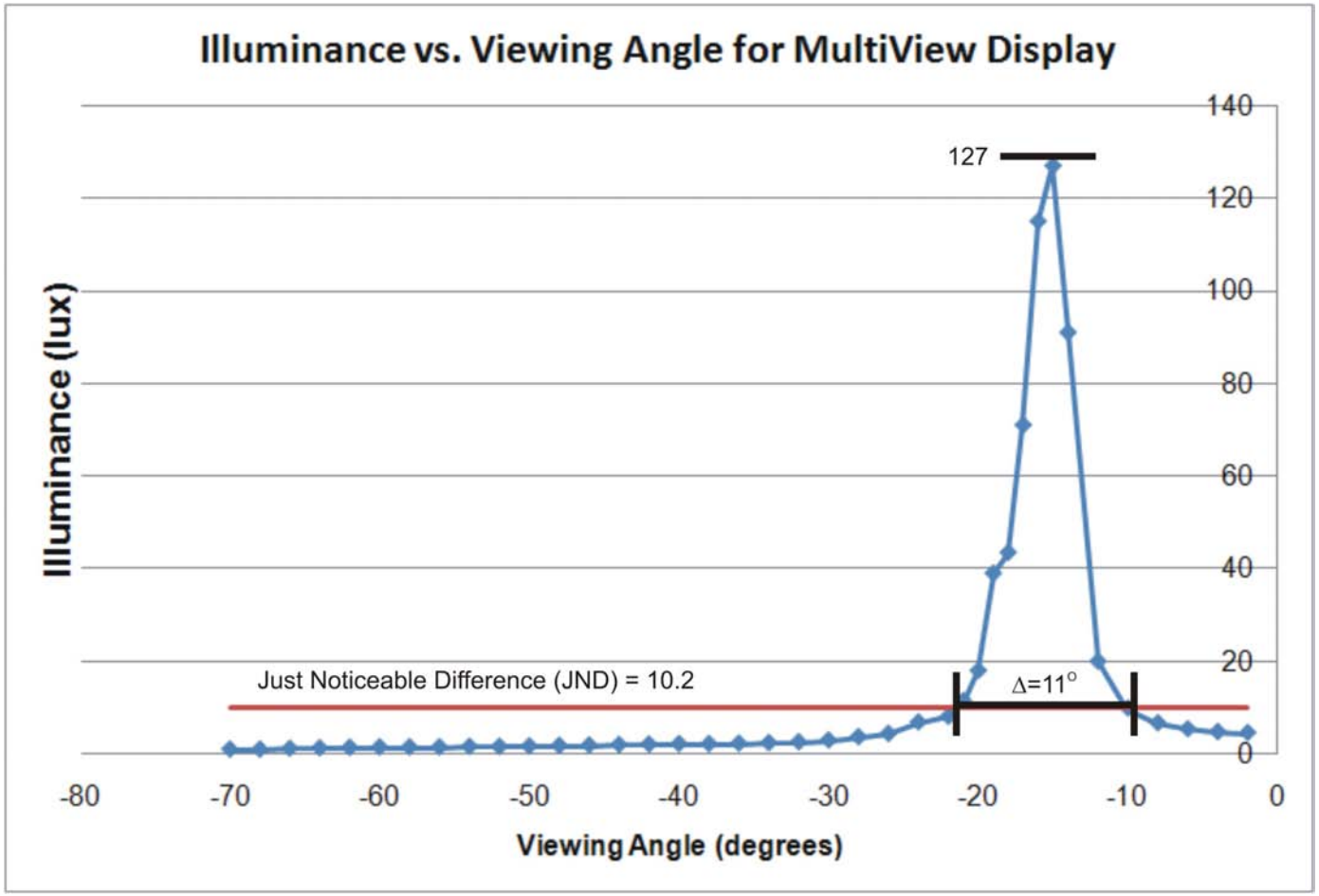


Figure 7.4: Illuminance versus viewing angle measurements for the retroreflective screen. The Just Noticeable Difference (JND) is also plotted to show the intensity of light needed from a competing image before it is perceived by a viewer at the optimal viewing position.

A major drawback with this configuration of optical elements is that it introduces a couple artifacts, reported qualitatively, because of the materials used and the space between the retroreflector and the diffuser. First, the diffusive layer actually amplifies the slight diffusion angles of the retroreflective layer. Because of this, the image is softened a bit making image slightly out of focus. Second, because an antiglare layer was applied to the lenticular sheet to reduce specularity, the diffuser actually images what is projected onto it both with the light enters the screen as well as when the light exits the screen.

Though the diffused retroreflector method exhibits the diffusion profile needed to act as a multiple-viewpoint display for MultiView, it also introduces several artifacts which take away from the possible quality of the image. In order to overcome this, I will explore lenticular-based displays in the next section.

7.2 Lenticular Method Based Displays

Another approach to producing a multiple viewpoint display is based on the lenticular method. In this method, a *lenticular lens* – a sheet of half-rods as shown in Figure 7.6 – is placed on top of a carefully crafted *lenticular image* which actually contains the information for multiple images.

This method has been developed since the early 1900's for still images and has produced several successful novelty items. Figure 7.5 shows a notebook with a lenticular image on the cover. When viewed from one angle, the cover shows an image of Diane Prince. As the viewing angle changes, the image changes and Diane Prince transforms to Wonder Woman.

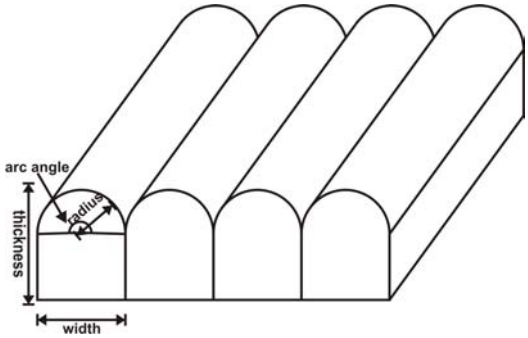
By replacing the lenticular image plane with a diffusive backing, it is possible to create a multiple-viewpoint projection screen [21]. Several engineering parameters can affect the quality of each viewpoint including features of the lenticular lens and properties of the diffusive backing.

I believe the lenticular method can be used to engineer new and better displays. In this section, I introduce the lenticular methods and evaluate the effect of varying the properties of the diffusive backing on display performance. I am specifically interested in two qualities of the display. First is the overall brightness of the intended image when at the appropriate viewing angle. Second, is the brightness of unintended images from competing viewing angles.

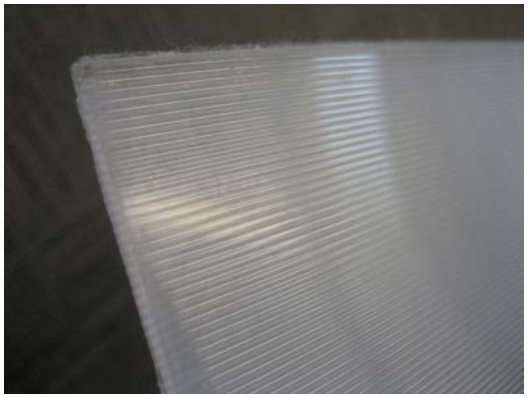
(a) When viewing this notebook from one angle, we begin to see the transformation from Diana Prince to Wonder Woman.

(b) As the notebook is tilted and the image is viewed from another angle, the transformation to Wonder Woman completes.

Figure 7.5: A novelty notebook using lenticular imaging to animate the transformation of Diana Prince to Wonder Woman as the notebook is tilted up and down.



(a) A diagram of a lenticular sheet. The design of these sheets have several parameters including arc angle, radius, thickness, and width.



(b) A photograph of a standard lenticular sheet used in MultiView's construction

Figure 7.6: Lenticular lens sheet

7.2.1 Implementation

The lenticular method works by directing different parts of the lenticular image toward different directions as shown in Figure 7.7. Typically, the lenticular lens is mounted so that the individual lenticule lenses run in the vertical direction and the curvature is in the horizontal direction. When viewing the lenticular display, the lenticules, focuses a unique portion of the lenticular image to the viewer. The focused portion of the lenticular image is unique to each viewing angle in the horizontal direction given the vertical mounting of the lens. Because there is no curvature in the vertical direction, the lens does not affect the image in that direction. This technique has been used successfully to create up to seven viewing zones in practice.

This technique has been quiet successful for multiplexing multiple still images with a static lenticular image on the back plane. It has also been adapted to support full motion video using one of several methods. The most common method is to apply a lenticular lens on top an LCD or plasma display. Making sure that an appropriate lenticular image is provided by the display, a multiple viewpoint display can be created. Many products are available on the market today that use this approach. The advantages of this method are that implementation is easy, low-cost, and the materials are readily available. The drawbacks of this method are that it effectively reduces the resolution of each image by

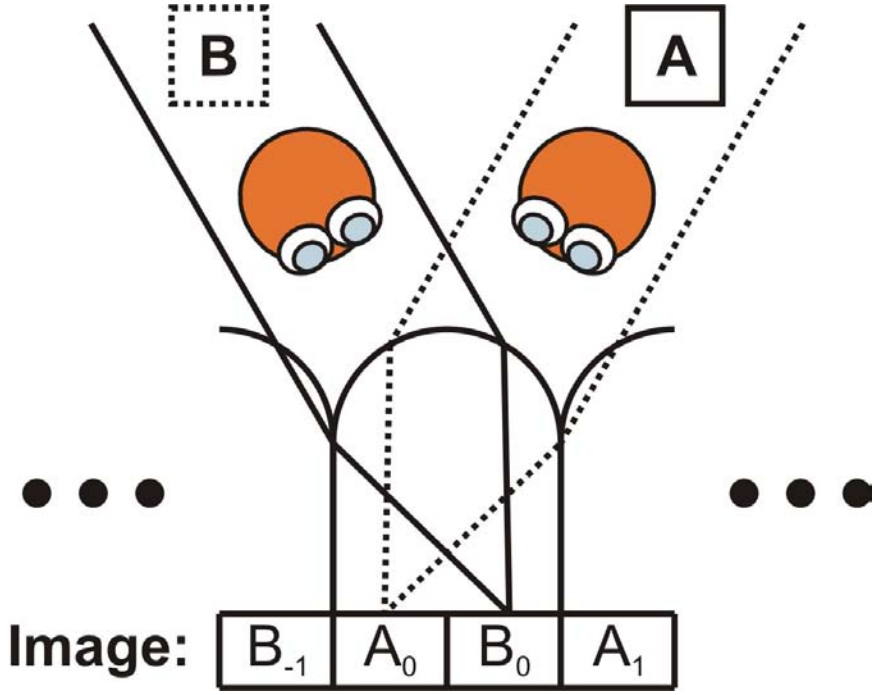


Figure 7.7: An example of the lenticular method. This example considers how a single lenticular sheet multiplexes two images on a lenticular image – 'A' and 'B' – to two viewers viewing the screen from different angles. The lenticular sheet is mounted so that the lenticules run vertically. A single lenticule is considered and the lenticular image underneath it: 'A_0' and 'B_0'. When viewer B views the lenticular screen, the lenticule focuses on a portion of B_0. When viewer A views the lenticular screen, the lenticule focuses on a portion of A_0. Each lenticule provides a vertical slice of the entire image in the same fashion and considering all lenticules provides a complete image. (Note: Not drawn to scale.)

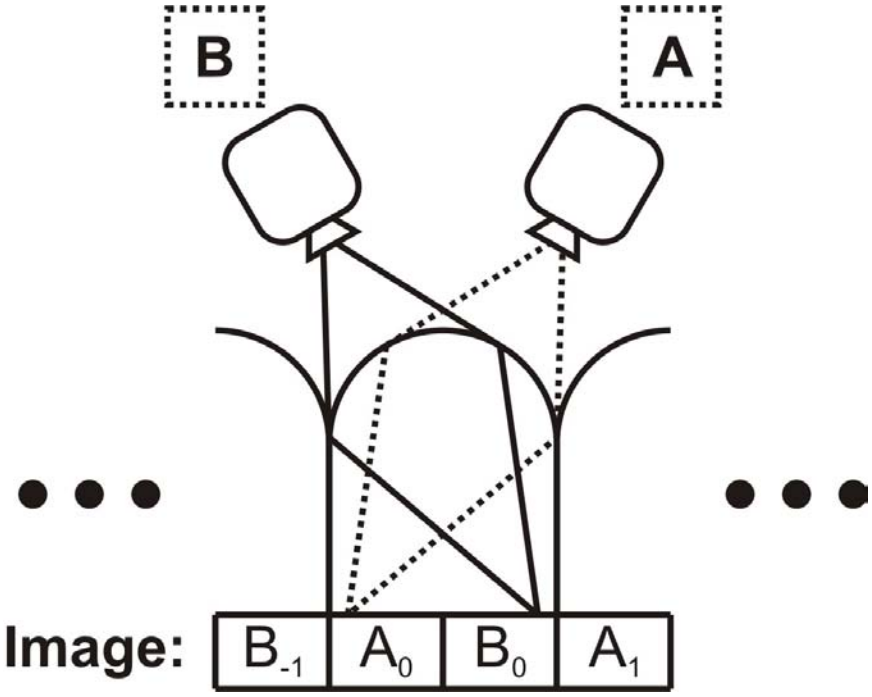


Figure 7.8: An example of using the lenticular method as a front-projection multiple-viewpoint display. The lenticular image backing is replaced with a diffuser. Two projectors are placed in front of the lenticular display, each projecting their own respective images: 'A' or 'B'. Each lenticule focuses a portion of each of the images to a unique portion of the diffuser creating a multiplexed lenticular image. As a result, any person viewing from the same position as one of the projectors will see only the image projected by the respective projector. Viewers can be vertically displaced from the projectors since the image is diffused in the vertical direction. (Note: Not drawn to scale.)

a factor equal to the number of viewpoints supported. For example, for two images, the resolution of each image would be half the native resolution of the underlying display. Image size is also limited to the underlying display.

Though several other approaches exist, a promising approach is described by Matusik et al. [21] and involves a front projection method. If a lenticular lens is backed with a diffuser and multiple projectors are placed in the same horizontal positions as the desired viewing positions, a multiple viewpoint display can be created where the viewing zones are defined by the positions of the projectors.

At the macro level, the behavior of this screen is much like the behavior of the diffused retroreflector method described in Section 7.1. When images are projected onto

the lenticular screen, the lenticular lens produces the multiplexed lenticular image needed on the diffusion layer to show that image to a viewer in the same position as the projector. This results in retroreflective behaviors in the horizontal direction. Because there is no lens curvature in the vertical direction, the image will be diffused by the backing without any lens effects resulting in standard diffusion behavior in the vertical direction. This allows the viewer to be vertically displaced from the projector. Figure 7.8 illustrates a two-zone multiple-viewpoint display.

In this next section, I begin an engineering analysis of the lenticular method based display by measuring the effect of different diffusive backings on the image quality. I will vary the transparency of the diffusive backing and measure the illuminance profile of the screen from different viewing angles to evaluate the effectiveness of this configuration as a multiple-viewpoint display.

7.2.2 Measurement

The lenticular display I built consisted of a lenticular sheet as shown in Figure 7.6(b) (Supplier: MicroLens, Lenticules per Inch: 30, Thickness: 0.052", Height x Width: 32"x42") that had an optically clear pressure sensitive adhesive (PSA) on the back allowing us to mount several different diffusing backings. I created three different displays, each differing by the diffusive backings:

mylar a single sheet of mylar plastic

single paper a single sheet of ink jet plotter paper.

double paper two sheets of ink jet plotter paper. The first sheet was affixed to the lens using the adhesive on the lenticular lens. The second sheet was affixed to the first sheet of paper using a spray adhesive

These three diffusing backings vary along transparency. The mylar sheet is a translucent material allowing much light to both scatter within the material and pass right through it relative to the other two diffusing backings. Paper is more opaque, with little scatter of light within the paper and less light passing through it. Doubling up paper allows even less light to pass through.

To measure the diffusion profile of the screen, I used the same method as described in Section 7.1.2.

7.2.3 Results and Discussion

The results show several trends between the three different conditions shown in Figure 7.9.

I begin this analysis with the brightness of the intended image at the optimal viewing angle. The optimal viewing angle in the experiment would correspond to the projection angle of $\theta = 15°$. The double paper design showed the brightest image (96.0 lux), followed by the single paper (66.1 lux), with the mylar design showing the lowest (20.6 lux). These results are in line with expectations based the materials. The mylar backing allows for more light to transmit through it than the single paper design, resulting in less light returned to the viewer. Double paper allows even less light transmission than single paper, resulting in the brightest returned image.

The second measure is the brightness of the image when viewing from outside the viewing zone. The brightness of the image drops sharply as measurements are taken further away from the optical viewing position and settles at a constant brightness. Sampled at $-40°$, constant brightness was highest with the mylar design (6.00 lux), followed by double paper (4.41 lux) and single paper (3.80 lux). These results are also in line with expectations based on the materials. Since mylar allows light to scatter much more throughout the material, more light should bleed outside the viewing zones. The single paper and double paper designs exhibited very similar behavior, though the double paper design was slightly more reflective since less light was lost to transmission.

When approaching $-53°$, there is an increase in brightness again marked in Figure 7.9 as the lenticular repetition. Sampled at $-53°$, brightness was highest with the double paper design (12.0 lux), followed by single paper (8.60 lux), with the mylar design showing the lowest (6.40 lux). This increase in brightness is a result of a property of the lenticular lens that can be controlled by modifying the arc angle of each lenticule. The brightness difference between each of the three conditions can be attributed to the amount of light loss to transmission.

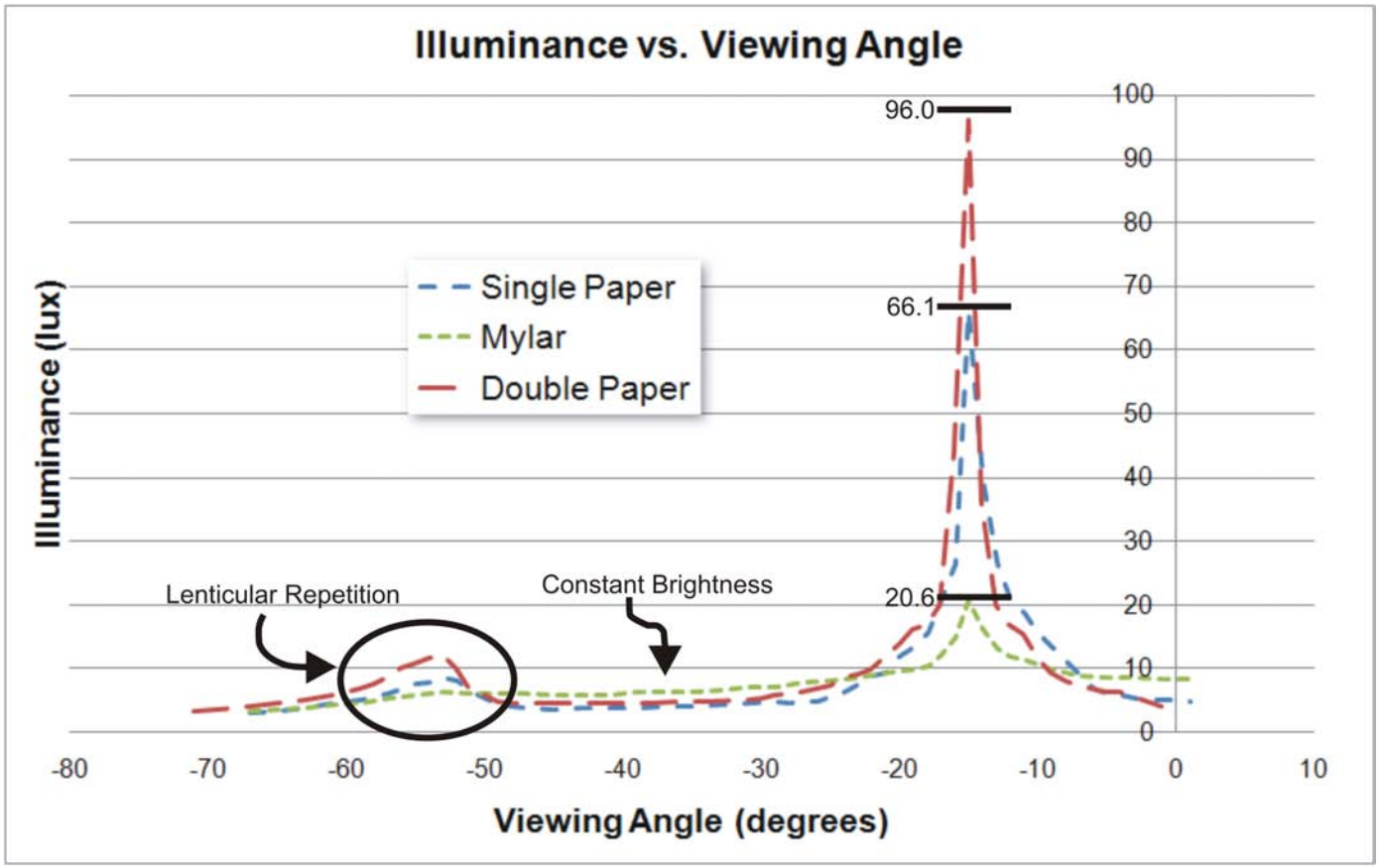


Figure 7.9: Illuminance versus viewing angle measurements for three variations of the lenticular screen. This chart compares (a) mylar, (b) single layer of paper, and (b) double layer of paper as the diffusive backing.

Evaluation of the diffusive materials shows that image quality can be greatly affected by the choice of diffusion material and that a carefully engineering diffusion backing may lead to desired display properties. These experiments show that controlling the amount of light transmission can directly affect the brightness of the return image. They also show that controlling the amount of internal scattering can control the brightness of an image outside the viewing zone.

7.3 Conclusion and Future Work

In this chapter, I evaluated two methods of producing multiple viewpoint displays. The first method, the diffused retroreflector method, utilized a retroreflector to return the light in the direction of the projector for horizontal retroreflection, and a diffusion layer for vertical diffusion. The second method, the lenticular method, utilized lenticular image for horizontal retroreflection and vertical diffusion.

Both designs exhibited the properties needed for effective multiple viewpoint displays and supported the necessary technical specifications needed for video conferencing; namely large image size, high resolution, full motion video, and not requiring viewers to wear special goggles. However, both approaches need work in improving the image brightness of the appropriate image and reducing the brightness of competing images. In others words, eliminating crosstalk.

The diffused retroreflection technique, the first one described in this capture, involves several layers with a spacing between them. The several optical layers introduce construction complications as well as several artifacts that must be controlled for. To help alleviate some of these issues, I am currently exploring lenticular displays. Its simple construction uses very few optical elements compared to the diffused retroreflective display, leaving little room for introducing unwanted artifacts.

Because of this, I began exploring lenticular-based methods, the second technique described in this chapter. Future work will involve exploring more diffusive materials to characterize the precise effect properties of the materials have on the image quality. Diffusers may include holographic diffusers, coatings, and surface treatments. In addition to the diffusive backing, further experimentation and engineering can be explored with the lenticular lens including tuning the lens dimensions, applying different surface treatments, or exploring new lenticular profiles that deviate from the standard circular lenticule.

Part III

Evaluation

Chapter 8

Perceptual

8.1 Introduction

MultiView is the first design of a video conferencing system that supports full spatial faithfulness in a group-to-group setting. In this evaluation, I set out to demonstrate that the system functions as a spatially faithful video conferencing as well as collect user feedback to help inform future design iterations of the video conferencing system.

8.2 Method

8.2.1 Participants

Seven groups of three and one group of two were used for testing. Overall, 23 participants took part in the user study. They were recruited from the undergraduate and graduate student population at University of California, Berkeley. Each participant was paid $10 upon completion of the experiment. In addition to the participants, a set of researchers were recruited from the Berkeley Institute of Design to provide the visual stimuli in the experiments. There was a pool of six researchers used in sets of three. The makeup of the researcher group for each session was determined by availability.

8.2.2 Apparatus

During this evaluation, the iteration of MultiView available was the one described in Section 6.2 and shown in Figure 8.1. The results of this evaluation were subsequently

Figure 8.1: The MultiView setup used in the evaluation. On the desk are acrylic cutouts of numbers 1 through 5.

used to inform the design of the latest described in Section 6.3. Two systems were setup back to back in the same room as diagrammed in Figure 8.2. At one site, several acrylic cutouts in the shapes of numbers '1', '2', '3', '4', and '5' were placed across the table and served as attention targets. Numbers 1, 3, and 5 were placed directly in front of each chair while numbers 2 and 3 were placed between them. There was about 8" of separation between two consecutive targets.

8.2.3 Measuring Gaze and Gesture Detection

To measure gaze, I performed a stimulus-response experiment and frame it in terms of the attention-based model presented in Section 3.2. There were two sites in the experiment, a site for three researchers who played the role of attention sources and a site for three observers consisting of recruited participants. Acrylic cutouts were placed at the observer site and provided attention targets for the attention source.

Using this setup, the attention sources individually present an attention cue toward one of the attention targets at the observer site. The observers then record what they

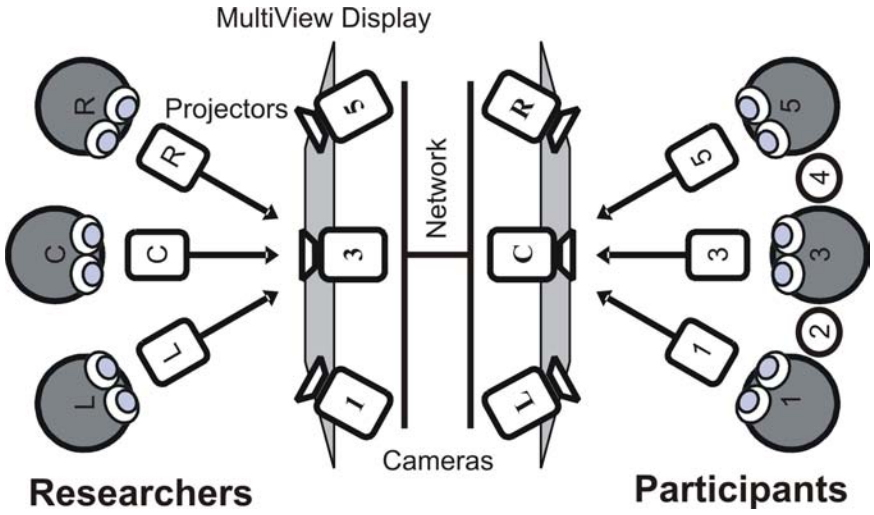


Figure 8.2: A diagram of the current MultiView Setup with two sites. Each site can support up to three participants. Researchers sat at positions left, center, and right which were designated by the letters L, C, and R, respectively. Experiment participants sat at position 1, 3, and 5. Positions 2 and 4 were targets between participants 1 and 3, and 3 and 5, respectively.

perceive to be the attention target. The intended attention target of the attention sources can then be compared to observed attention targets of the observers providing a performance metric of attention cue registration.

For example, suppose at the attention source site, one of the participants gazes toward target '2' in the observer site. The observer is then asked which target the gazing participant appears to be looking at and they can choose between targets 1, 2, 3, 4 or 5. The response from the observers may be correct or may include some error. Spatial faithfulness is operationalized as the amount of error that is introduced by the video conferencing system – a low error indicates high spatial faithfulness while a high error indicates low spatial faithfulness.

The measurement consists of repeated rounds where in each round, the three attention sources present three individual attention cues. The intended attention target is independent for each attention source. So one attention source may be focusing on Target 2 while another is focusing on Target 4. Each observer is then asked to determine the attention target of each of the attention sources. Because there are three observers and

each observer is asked to determine the the attention target for each of the three attention sources, 9 total responses are be provided for each round.

8.2.4 Procedure

The experiment took 60 minutes for each session. Upon arrival, participants were assigned to one of three seating positions located on the observer side of the video conferencing system. At the attention source site of the video conferencing system are three members of the research team. These three members were selected from a larger set of researchers based on availability.

Once all the participants arrived, they were shown to their seats and presented with consent materials. They were instructed that there would be three tasks to the experiment and that each section would be preceded by specific instructions.

Task 1 Each researcher was instructed to look at one of the five positions. The positions were randomly generated prior to each session of the experiment and provided to each researcher on a sheet of paper. If the position happened to have a participant in it – positions 1, 3, and 5 – they were instructed to look into the image of the participants eyes on the screen. If the position was in between two participants – positions 2 and 4 – they were asked to look toward that position at the average eye level of the participants. The participants were then asked to record which position each researcher appeared to be looking at on a multiple choice answer sheet. They were carefully instructed to avoid trying to determine which target they felt like the researcher actually was looking at, but to instead concentrate on which target the image of the researcher appeared to be looking at. This process was repeated ten times.

Task 2 This task is similar to task 1, except that instead of gazing at each of the positions, the researchers were asked to point in the direction of the position. This process was repeated ten times.

Task 3 In this task, participants and researchers were paired off. The researchers were asked to gaze at points on the screen relative to their participant partners eyes. They were asked to look at one of the following: above the camera, at the camera, at the participants eyes, below the eyes, slightly to the right of the eyes, or slightly to

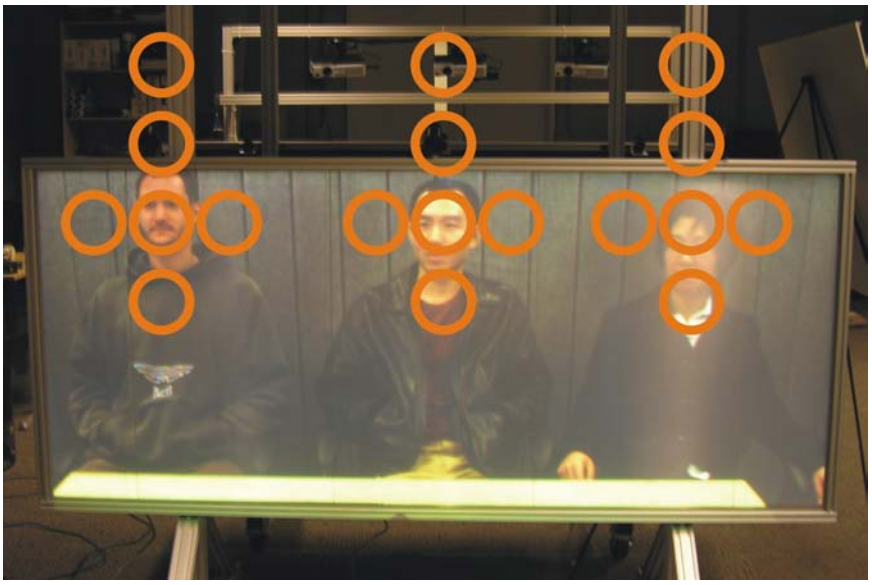


Figure 8.3: Gaze position targets for Task 3. The attention sources were asked to gaze either (a) at the eyes, (b) below the eyes, (c) to the left of the eyes, (d) to the right of the eyes, (e) at the camera (above the eyes), or (f) above the camera.

the left of the eyes. Targets are illustrated in Figure 8.3. The order of the targets was randomly generated before each session of the experiment. Each participant was asked, "Do you feel as though the researcher is looking directly into your eyes?" After 10 trials, participants and researchers switched partners. This process was repeated until all pairs were exhausted.

At the end of the experiments, the participants were asked the following question in order to help us interpret results, provide insight into the way MultiView was used, determine possible design improvements, and guide future work:

> "Please use the space below for any comments you have on our new system. This may include, but is not limited to, details about the system, reactions to how you felt about using the system, any perceived differences between using MultiView and face-to-face meetings, perceived differences between MultiView and other video conference systems you have used, etc."

Stimulus Target

Perceived Target	1	2	3	4	5
1	78.5	24.2	10.6	1.0	0.0
2	18.5	33.9	18.8	6.2	0.9
3	0.0	30.3	46.3	24.9	5.6
4	3.1	10.3	20.6	35.8	29.0
5	0.0	1.2	3.8	32.1	64.5

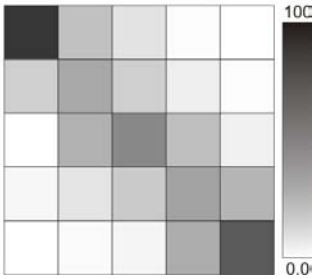


Figure 8.4: The confusion matrix for Task 1. Each column represents the actual target of the gaze stimulus and each row represents the target as perceived by the participants.

8.3 Results and Analysis

8.3.1 Task 1: Group Gaze

The results of task 1 are presented in different ways that are relevant to the discussion that follows. Figure 8.4 presents the results in the form of a confusion matrix. Each column represents the actual target of the gaze stimulus and each row represents the target as perceived by the participant given the gaze stimulus. For example, for all gaze stimuli directed at position 3 (column 3), 10.6% of the responses perceived that the gazer was looking at position 1, 18.8% at position 2, 46.3% at position 3, 20.6% at position 4, and 3.8% at position 5. For the condition of gaze, 91.7% of the responses were at most one target off.

Another measure takes a closer look at error in perceiving the attention target. Error of any given stimulus i (ϵ_i) is defined to be the difference between what the observer perceived to be the attention target of the image (t_{pi}) and the actual attention target of the researcher producing the gaze stimulus (t_{ai}) and is reported in units of *positions*:

$$\epsilon_i = |t_{pi} - t_{ai}|$$

Table 8.1 presents the mean error and standard deviation of error by the observer's viewing position. For instance, the mean error for observers sitting at position 1 was 0.70*positions*. An analysis of variance showed that viewing position had no significant effect on mean error, $F(2, 687) = 1.48, p = 0.23$. This is to be expected, in fact, it is a validation of the Mona

Viewing Position	$\mu(positions)$	σ
1	0.70	0.65
3	0.63	0.67
5	0.60	0.70
Combined	0.64	0.68

Table 8.1: The mean error (μ) and standard deviation (σ) in gaze direction perception by viewing position.

Gaze Target	$\mu(positions)$	σ
1	0.28	0.63
2	0.79	0.67
3	0.68	0.71
4	0.73	0.62
5	0.43	0.65

Table 8.2: The mean error (μ) and standard deviation (σ) in perceived gaze direction for each set of stimuli directed at each target in Task 1.

Lisa Effect – the effect implies that perceived view is not affected by viewer angle relative to a screen.

Table 8.2 presents the mean and standard deviation of error by the target of the gaze stimuli. For instance, the mean error of responses to all stimuli targeted at position 2 was 0.79. The Tukey HSD procedure showed significant differences in any pairing between stimuli whose target was 2, 3, or 4 and stimuli whose target was 1 or 5. There was no significant difference for any other pairing.

8.3.2 Task 2: Gesture

The results found in Task 2 were very similar to those found in Task 1. They are summarized in Figure 8.5, Table 8.3 and Table 8.4 without further discussion. For the condition of gesture, 94% of the responses were at most one target off.

8.3.3 Task 3: Mutual Gaze

A summary of the results from task 3 are given in Table 8.5. The first column ("Gaze Direction") describes the direction of the gaze. The second column ("Total") is the total number of stimuli presented in that direction. The third column ("Yes") is the

Stimulus Target

Perceived Target	1	2	3	4	5
1	78.9	35.7	4.1	2.7	0.0
2	19.3	38.4	23.4	6.8	1.3
3	1.8	22.2	47.4	29.9	9.0
4	0.0	3.8	22.8	35.4	33.3
5	0.0	0.0	2.3	25.2	56.4

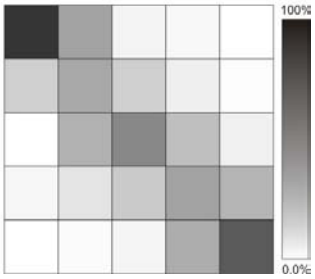


Figure 8.5: The confusion matrix for Task 2. Each column represents the actual target of the gesture stimulus and each row represents the target as perceived by the participants. The confusion matrix is represented textually on the left and graphically on the right.

Viewing Position	$\mu(positions)$	σ
1	0.55	0.61
3	0.53	0.60
5	0.65	0.67
Combined	0.58	0.63

Table 8.3: The mean error (μ) and standard deviation (σ) of perceived gesture direction perception by viewing position.

Gesture Target	$\mu(positions)$	σ
1	0.23	0.46
2	0.65	0.55
3	0.59	0.61
4	0.76	0.69
5	0.55	0.71

Table 8.4: The mean error (μ) and standard deviation (σ) in perceived gesture direction for each set of stimuli directed at each target.

Gaze Direction	Total	Yes	No	Rate
Above Cam	100	54	46	54.0%
At Cam	132	91	41	68.9%
At Eyes	127	81	46	63.8%
Below Eyes	136	76	60	55.9%
Left of Eyes	123	74	49	60.2%
Right of Eyes	72	37	35	51.4%

Table 8.5: The responses of the participants based on the direction of gaze in Task 3.

number of times a participant replied positively as to whether or not they felt the researcher was looking directly into their eyes. The fourth column ("No") is the number of times a participant replied negatively to that same question. The fifth column ("%Rate") is the rate at which the participants answered positively.

8.4 Discussion

Referring back to Figure 8.2, let us consider the seventh trial of the third session. Researcher L is instructed to look at target 1, Researcher C at target 1, and Researcher R at target 5. All the participants, mindful of being asked to record where they think the *image* of the researcher is looking, respond correctly for each researcher. If this trial were reproduced using a standard single view setup with the camera positioned at the center of the screen, then the observer sitting at position 1 would perceive Researcher R looking, incorrectly, at position 3 and Researchers L and C looking beyond the available targets to her left. An observer at position 5 would also have similar distortions. The only one with the correct perspective would be the observer at position 3 since the position of the remote camera correlates to that person's perspective.

The position of the observer had no significant effect on the mean error. Observers were often able to respond to a stimulus in a matter of a second. The mean error in determining the direction of a person's gaze was 0.64. The rather low accuracy is probably due to the large distance between the two sets of participants, discussed later.

In much of the established literature on gaze, acuity is often measured in degrees. Given the above geometry, the change in angle between any two adjacent attention targets can be calculated as:

$$\Delta\alpha = \arctan\left(\frac{20cm}{300cm}\right) = 3.82^\circ$$

If this value is multiplied by the mean error, an extremely rough estimate of sensitivity in degree measure can be produced: $0.64 \cdot 3.82^\circ = 2.45^\circ$. This value is roughly on par with previous empirical values for gaze direction acuity [8, 18]. This task does not have the precision required to accurately measure human acuity and was not intended to do so.

The two end positions, 1 and 5, enjoyed a significantly lower mean error than the interior positions, 2-4. From the comments gathered during the experiment, it seems that this is due to a self-calibration phenomenon resulting from the setup of the experiment. The participants were aware that the target set consisted of only five positions, and quickly learned what the images looked like when looking at the end positions. Comments like "I thought the last one was a 5, but it was not because this time she is looking even more to the right," were common.

Task 3 was designed to provide more precise characterization of MultiView's support for mutual gaze awareness. The expectation was that participants would answer "yes" near 100% of the time when gaze was directed at the camera. However, the rate for this case was actually at 68.9%. In addition, there is little difference between the rates of perceived eye contact between each gaze direction. When asked for comments at the end of the experiment, it was repeatedly mentioned that it was difficult to make out the exact position of the pupil because of the distance and image quality.

However, the participants also mentioned that they had a strong sensation of eye contact during impromptu conversations with researchers between experiments. They felt like the entire *context* of the conversation, combined with the visual information, provided a strong sensation of eye contact even with the limited ability to determine pupil position.

This highlights a separation between the ability to determine the position of a pupil and the sensation of eye contact. In [25], Perrett describes the existence of

a *direction-of-attention detector* (DAD), which is a specialized brain function used to determine the attention target. His theory suggests that, though the eyes are the primary source of information, the DAD can come to depend more on other cues such as head orientation and body position when the eyes are viewed from a distance or otherwise imperceptible, as is the case with MultiView. The task presented to the participants required them to judge pupil direction, but the differences between the images of two different gaze points were apparently imperceptible.

8.5 Conclusion

To test the early design of MultiView and to collect feedback for future designs, several users were brought in to interact through the video conferencing system in a controlled way. In this particular study, we introduced a method of testing group-to-group video conferencing systems for spatial faithfulness. This method was used to demonstrate the effectiveness of the MultiView video conferencing system in preserving spatially dependent nonverbal cues.

User feedback was also elicited as part of the experiment to help inform the design of the third iteration of MultiView described in Section 6.3. Several themes emerged. The first involved improving the image quality. The second involved the comfort of users, especially considering the noise and heat produced by each of the projectors.

I have presented the first video conferencing system that is able to faithfully present spatial information of nonverbal cues when there are multiple participants at any given site.

Chapter 9

Trust Formation

9.1 Introduction

> We're never so vulnerable than when we trust someone – but paradoxically, if we cannot trust, neither can we find love or joy – Walter Anderson, 1998, p44

It has always been suggested that a system that supports eye contact and gesture awareness – a system that I call spatially faithful – can dramatically improve communication between two meeting groups. However, up until the design of MultiView, there has never been a way to experimentally test this hypothesis in a group-to-group meeting structure. With only spatially unfaithful systems available, researchers generally make the comparison to face-to-face meetings when measuring the effect of spatial distortions in video conferencing on communication. The drawback of such an experiment is that the precision is very limited. Many factors change between meeting through a spatially unfaithful video conferencing system and meeting face-to-face, making it difficult to attribute any measurable differences to spatial distortions alone. Other factors – such as image quality, network latency, or the ability to shake hands – may, just as well, have affected the results. The MultiView design provides the first platform to test the effects of spatial distortions on communication for group-to-group meetings. Up until this point, I have shown prior work which can only suggest that spatial faithfulness may improve the way two parties communicate with each other. With MultiView as a platform, I will now experimentally conclude that spatial faithfulness significantly improves an important aspect of communication: trust.

There is a growing body of literature on how computer-mediated communication systems affect trust formation. For instance, Drolet and Morris show that dyads playing a conflict game tend to show more cooperative behaviors when communicating face-to-face than when communicating over the telephone [15]. Rocco showed that 6-person groups playing an investment game tend to show more stable and cooperative investing when communicating face-to-face than when communicating over non-anonymous mailing lists [28]. Bos et al. had 3-person groups play an investment game across four different communication channels: face-to-face, video-conferencing, audio-conferencing, and instant messenger. They found that participants communicating face-to-face showed higher and more consistent levels of cooperative investing than those using computer-mediated communication systems [5]. All the above measures, however, limit their studies to one participant per video conferencing site. This leaves out the common meeting structure with multiple participants at each site.

9.2 Trust Measurement and Validity

Trust, unfortunately, is one of those concepts whose understanding has been the subject of volumes of work and is still highly debated. Despite the complexity in trying to define trust, there are many experimental methods that purportedly measure trust [3, 4, 5, 11, 15, 19, 27, 28]. Of course, with any measure, we must take into account the internal validity that the experiment is actually measuring the phenomenon of interest. It is not the intention of this work to contribute to the working understanding of trust or undermine the complexity of trust, but present broadly how trust is currently understood, how it is measured, and how it is that this current measure captures or fails to capture certain aspects of trust[1].

Though we need not subscribe to any one particular conception of trust, there are several aspects of trust that are generally regarded as necessary in any conception of it. First, trust is a three-part relation: a person trusts another person or group to do some action. Second, trust is an assessment of some truth beyond our control and that this assessment is to be separated from any action informed by trust. Third, there must be an element of risk for trust to be relevant. Fourth, the trustee's action is a result of a larger

[1]For a more detailed analysis of trust and limitation of trust measurement instruments, see [10, 11, 27]

	Cooperate	Defect
Cooperate	Reward(3)/Reward(3)	Sucker(10)/Temptation(0)
Defect	Temptation(0)/Sucker(10)	Punishment(7)/Punishment(7)

Table 9.1: Payout structure of Prisoner's Dilemma game.

context that just the trust relationship and other factors may trump in the decision making process [19].

Many measures of trust are built upon the basic structure of a game known as the Prisoner's Dilemma or PD for short. In this classic game, you and a partner are asked to pretend that you have just been arrested. The two of you are separated and you are given a choice, you can confess to the crime or you can remain silent. Your partner will have the same choice. If you confess but your partner remains silent, charges against you will be dropped (0 years) and your testimony will be used to make sure your partner does some serious time (10 years). Likewise, if you remain silent, but your partner confesses, her testimony will be used to make sure you do serious time (10 years) while the charges against her are dropped. If you both confess, you will both be convicted but with early parole (7 years). If you both remain silent, both of you will be convicted for a minor crime (3 years) [26].

The specifics of the prison context can be abstracted away revealing an underlying reward/punishment structure with payouts labeled as "reward", "sucker", "temptations", and "punishment". By definition, a prisoner's dilemma holds the following relationships between each of the outcomes in terms of desirability: $Temptation > Reward > Punishment > Sucker$. Of course, other relationships can exists leading to different types of games that measure different aspects of cooperation.

To analyze the outcome of PD experiments, the joint payoff is often taken as the measure of cooperation and the indicator of trust. In the Prisoner's Dilemma game set up above, the participants will serve 6 years if they both cooperate with each other, the minimum amount of total years indicating maximum trust. Should one defect while the other cooperates, than a total of 10 years must be served indicating a medium level of trust. Should they both defect, a total of 14 years must be served indicating a minimum level of trust. The outcomes are summarized in Table 9.2.

As a measure of trust, several shortcomings have been discovered with the basic PD game. In order to highlight these shortcomings, I will contrast the PD game to a game introduced by Cook et al. known as Prisoner's Dilemma with Risk or PD/R [11].

First, the fact that trust is an assessment of another person brings into question as to what the interesting questions to ask are. Often times, PD games are presented as one-shot, binary choice games. Though this might model an interesting class of trust relationships, many relationships in life are developed over time and involve varying levels of risk. PD/R addresses these issues by allowing each player to choose, along a continuous scale, how much to risk. Instead of a one-shot game, experiments are executed in multiple rounds. This allows researchers to measure varying levels of trust and to see how that develops over time.

Second, as mentioned above, trust is a three-part relation where one person trusts another to do some action. In developing an experiment that measures trust, it is important to consider precisely who is trusting, who is being trusted, and what action constitutes trust behavior. Because both participants make a single, simultaneous, binary decision in standard PD games, there is no clear separation between the action of the truster and the action of the trustee. In addition, there is no clear separation between a trusting action and a cooperative action, two separate phenomenons that can happily exist without one another [10]. For instance, a player may choose to defect either because they do not trust the other person or they are not willing to cooperate. With no way to separate the actual reason, the experiment, as a measure, becomes less precise. There may be some relation in that trust may lead to higher levels of cooperation, but trust, as noted earlier, is only a single factor in the decision making process toward cooperation.

PD/R addresses this issue by separating the game into two steps. In the first step, each player is given 10 coins. Each player then decides how many of those coins they wish to entrust to their partner. After receiving information on how many coins they received from their partner, each player decides how many coins to return knowing that their partner will receive double the number of coins that are returned. The first step, measures trust each person has in their partner, knowing that if their partner does not cooperative, they will lose their coins. The second step measures cooperation, each partner knowing that they have the option of keeping the coins for their own immediate benefit.

I have highlighted two different shortcomings with the standard Prisoner's Dilemma game: limitations of one-shot, binary choice games and confounding cooperation

and trust. Of course, there will be many more issues depending on the definition of trust one subscribes to in terms of both internal and ecological validity of the experiment. However, the goal here was to highlight some of the key issues of the standard Prisoner's Dilemma game by comparing to the Prisoner's Dilemma with Risk game. In the next section, I describe the actual game used in the experiment, another variant of a PD game. The goal of this section was just to situate the Daytrader game in the current understanding of trust.

9.3 DayTrader: Measuring Trust

I used another instantiation of a Prisoner's Dilemma game called Daytrader [5] with some modifications. In the previous section, I highlighted several shortcomings of the standard Prisoner's Dilemma game. Daytrader addresses some issues but still exhibits some of the known issues. Namely, participants can choose their level of cooperation and the game is repeated several times, but it still exhibits issues of confounding cooperation and trust. This specific measure was chosen despite this shortcoming because it still provides a valid behavioral measure of an important aspect in group-to-group meetings and it allows us to compare and build upon results of prior studies. I describe Daytrader in this section.

In this study, a modified version of Daytrader was used to measure levels of trust in group-to-group communication. The rules of the game are as follows:

- There are 2 groups, each group consisting of 2 or 3 participants.
- The groups will play an unknown number of rounds.
- In each round each group is given 60 credits. Each group must decide how many of their credits to cooperatively invest with the other group (cooperate) and how many they wish to save for themselves (defect).
- For each round, a new group leader should make the final decision as to what the investment is going to be.
- The cooperative investment is put into a fluctuating market which will average 50% return over the course of the entire game.
- The earnings from the cooperative investment are divided evenly among the two groups, regardless of each groups' contribution to the cooperative investment. Each

group is told how much they earned, but they are not told what the other group earned.

- After every 5 rounds a "Rich Get Richer" bonus is awarded to the two groups. 60 credits are placed into the fluctuating market. The earnings are divided between the two groups such that the proportion of the awarded bonuses is equal to the proportion of the groups' earning in the previous 5 rounds.

- Discussion is allowed at any point in time, either with groupmates or with the opposing group. However, groups have about one minute between each round. After the bonuses are awarded at the end of 5 rounds, the groups are given extra time and are encouraged to have a discussion. Groups are not allowed to share precise numerical investment and earning amounts with the other group.

This game differs from the one presented by Bos et al. [5] by using a fluctuating market which adds noise to the information available to the groups. The goal was to make it ambiguous as to whether returns were the result of the other group's action or the market performance. A fluctuating market provides a way to hide defection moves as well as sabotage cooperative moves. It was an attempt to induce more dependence on the communication channel. The participants are made aware that the market is guaranteed to earn 50% on top of the investment by the end of the game and encouraged not to invest based on what they think the market is going to do but on what they think the other group is going to do. The fluctuation was determined before the experiment and was the same across all sessions. The market was determined using a random number generator with an even distribution between -50% to 150% averaging 50%. By adding noise, the game structure becomes an instantiation of what is known as Iterated Prisoner's Dilemma with Imperfect Monitoring [4].

Additionally, the original formulation of this game called for each participant to make a decision about the investments. In the formulation used in this experiment, a group needs to decide together how much to invest. To enhance group behavior while reducing effects of dominant and freeloading behaviors, a new group leader, who was in charge of making the final decision, was required in each round.

Though a group can decide to invest any amount between 0 and 60 credits, I will illustrate the game with four possible scenarios assuming average market performance. If

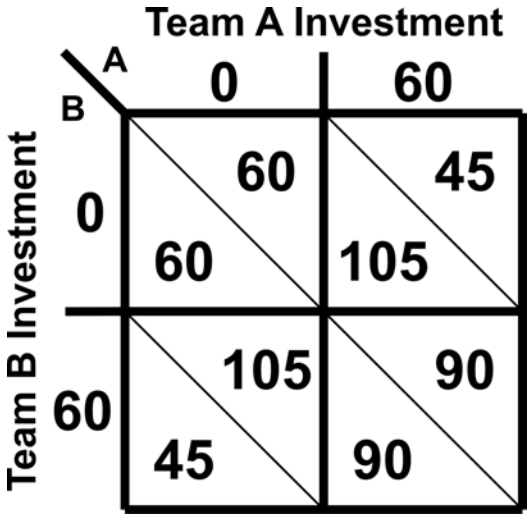


Figure 9.1: The payoff structure of Daytrader in four scenarios assuming average market earnings.

both groups invest 0 credits, each group will earn 60 credits for that round since they both just saved their credits. If both groups invest all 60 credits cooperatively, both groups will earn 90 credits. If Group A invests 60 credits while Group B makes a defection move and invests 0, then Group A earns only 45 credits while Group B earns 105 credits and vice versa. These payouts are illustrated in Figure 9.1.

As can be seen from these examples, by investing, a group puts itself at risk for defection by the other group resulting in less earnings than if they invested nothing at all. Additionally, by defecting, they also have the chance to earn more if the other group decides to invest cooperatively. The rational choice is to consistently defect. But once both groups settle on this strategy, both groups will earn less than if they invested irrationally – hence the dilemma.

In this experiment, the measure of trust is operationalized as the sum total of cooperative investments between the two groups. In each round, the measure of trust can be from 0 – where both groups invest nothing – to 120 – where both groups invest all their credits.

9.4 Hypotheses

In this experiment, I specifically make the following hypotheses based on previous findings:

Hypothesis 1 (H1): Groups meeting face-to-face will demonstrate higher levels of trust than groups meeting through non-directional video conferencing systems.

There is limited precedence in measuring trust formation in video conferencing conditions for the group-to-group structure. Finding support showing a difference between face-to-face and non-directional video conferencing conditions in group-to-group meetings adds credence to the problem to be solve. It also provides a basis for comparison. Specifically, I hypothesize the following:

Hypothesis 1a (H1a): Groups meeting face-to-face will show higher levels of overall *trust than groups meeting through non-directional video conferencing.*

That is, I expect that the total cooperative investment by groups meeting face-to-face will be significantly higher than the total cooperative investment by groups meeting through non-directional video conferencing.

Hypothesis 1b (H1b): Groups meeting face-to-face will show reduced delay *in trust formation when compared to groups meeting through non-directional video conferencing.*

Bos et al. found that trust generally increases over time. They called this phenomenon *delayed trust* [5]. They found that trust increased more slowly with participants meeting through video conferencing compared to groups meeting face-to-face. I expect to extend their results to the group-to-group setting and show that there will be a greater delay in trust formation for groups meeting through non-directional video conferencing when compared to groups meeting face-to-face.

Hypothesis 1c (H1c): Groups meeting face-to-face will show reduced fragility *in trust formation when compared to groups meeting through non-directional video conferencing.*

Bos et al. also found that there was a decrease in cooperative investment when bonuses are about to be offered. They called this phenomenon *fragile trust* [5]. They found that trust in participants meeting through video conferencing was less resilient to bonuses than for participants meeting face-to-face. I expect to extend their results to the group-to-group setting and show that groups meeting through non-directional video conferencing will exhibit more fragile trust than groups meeting face-to-face.

The second hypothesis compares the trust formation patterns of groups meeting through directional versus non-directional video conferencing systems. I expect that full spatial faithfulness provided by MultiView should improve trust by preserving many of the nonverbal cues which are distorted in non-directional video conferencing systems. Similar to Hypothesis 1, I make the following hypotheses:

Hypothesis 2 (H2): Groups meeting through directional video conferencing will show higher levels of trust than groups meeting through non-directional video conferencing.

Specifically, I hypothesize the following:

Hypothesis 2a (H2a): Groups meeting through directional video conferencing will show higher levels of overall *trust than groups meeting through non-directional video conferencing.*

Hypothesis 2b (H2b): Groups meeting through directional video conferencing will show reduced delay *in trust formation when compared to groups meeting through non-directional video conferencing.*

Hypothesis 2c (H2c): Groups meeting through directional video conferencing will show reduced fragility *in trust formation when compared to groups meeting through non-directional video conferencing.*

9.5 Method

9.5.1 Participants

Participants were recruited by the Experimental Social Science Laboratory (XLab) at University of California, Berkeley. The XLab maintains a database of university affiliated students and staff members who are interested in taking part in experiments. Participants are emailed about experiments and opt-in by signing up via an online calendar. There were 169 participants: 110 females (65%), 59 males (35%), 156 students (92%), and 13 staff members (8%). The average age of student participants was 20 years old, and the average age of staff member participants was 39. These participants formed 29 groups of 2 and 37 groups of 3. Groups played against each other in three different conditions in a between-group study.

The experiment occurred in two-hour sessions with between four to six participants. Because participants do not always shows up to their scheduled sessions,

up to ten participants were recruited for each session. If a participant could not be accommodated, they would be compensated with a $5 show-up fee. Participants taking part in the experiment were paid according to the outcome of the experiment, but were guaranteed at least $22.50.

9.5.2 Apparatus

This experiment used the diffused retroreflection based system (v2) as described in Section 6.3. The two systems were in separate but adjacent rooms and connected via local gigatbit ethernet.

9.5.3 Treatment Conditions

Face-to-Face In this condition, the two groups met in the same room. One group sat on one side of the conference table and the other group sat on the other side. The two groups were separated by 8'.

Directional Video Conferencing In this condition, the two groups met in separate rooms and communicated through the MultiView video conferencing system which takes advantage of the multiple viewpoint directional display and represents a spatially faithful video conferencing system. The groups sat 8' from the screen to mimic the distance of the face-to-face condition.

Non-Directional Video Conferencing This condition was identical to the directional video conferencing condition except the multiple viewpoint display was covered with a standard projection screen material and only the center camera and projector was used. Image quality remained the same. This condition mimicked the commonly found, spatially distorted video conferencing system.

9.5.4 Measurement Instruments

Task Performance Measure

The two groups played the variant of Daytrader as described above. The measure of trust is operationalized as the sum total of cooperative investments between the two groups. In each round, the measure of trust can be from 0, where both groups investing nothing, to 120, where both groups invest all their credits.

Post-Questionnaire

An adaptation of Butler's Conditions of Trust Inventory [7] was administered to the participants. The original inventory consisted of 110 Likert scale questions measuring 11 different conditions. Questions were selected and modified from this pool for appropriateness of the condition to be measured and brevity of the questionnaire. The conditions chosen were *trust in other group* (11 items), *trustworthiness* (5 items), and *consistency* (3 items). This inventory included questions like "I trusted the other group members in this game," "I could be trusted by the other group," and "During the game I behaved in a consistent manner." The participants responded on a scale of 1 (strongly disagree) to 7 (strongly agree).

Post Interview

Upon completion of the experiment, each group was interviewed separately. There were no predetermined questions, but the topics covered were general impressions of the other group, any specific incidents in the game that stood out, and discussion of any strategies they used. The post interview was to help explain some observed events during the game and to guide future research.

9.5.5 Procedure

The experiment took 120 minutes for each session. Upon arrival, each participant was immediately assigned to one of two groups. If participants were acquainted with another participant, they were placed in the same group. In the computer-mediated conditions, participants were escorted to their assigned rooms to minimize any face-to-face contact with opposing group members.

Once assigned to their groups, they were shown a set of videos that walked them through the consent materials and the rules of Daytrader. This process took about 30 minutes.

If applicable, the video conferencing systems were turned on and connected at this point. The participants were allowed to introduce each other to the other group and were given time for discussion before the game begun. Once they were ready, they would submit their investment amounts to a *fund manager*.

The fund manager is a program designed to prompt the groups for their investments. The groups interacted with the fund manager through America Online's (AOL) Instant Messenger (IM) program installed on a laptop on their conference table. Once the fund manager received the amounts, it would calculate each group's earnings and report them to the respective groups. Groups did not know the opposing group's earnings. The researcher could command the fund manager to send a "time out" warning, indicating to the participants that they are taking too long to make their decisions. This was necessary to get through enough rounds in the allotted experiment time.

This portion of the experiment lasted for 45 minutes. All groups played at least 30 rounds. The actual number of rounds played was variable between each session and groups were not made aware of how many rounds there would be.

Once the end of the game was reached, the two groups were allowed to say goodbye to each other. In the video conferencing conditions, the systems were shut down and connections were severed. In the face-to-face condition, the groups were separated into different rooms.

Each participant then filled out the questionnaire individually and an interview was conducted. This took about 30 minutes.

The participants were compensated for their participation in the study. The amount of their compensation was based on the number of credits their group earned and the number of rounds they played. Basing the compensation on the number of credits earned during the session provided motivation to do well in the game. The average compensation was $26.21, the maximum was $31.42, and the minimum was the guaranteed $22.50 even if the credits their group earned was worth less.

Each group left at separate times as to avoid meeting again.

9.6 Results and Analysis

9.6.1 Overall Cooperative Investment

I begin by looking at *overall trust* which is measured by the total cooperative investment across the entire game. All cooperative investments by both groups for the first 30 rounds are summed up for each session. The maximum cooperative investment is 3600 credits (60 credits/group * 2 groups/round * 30 rounds).

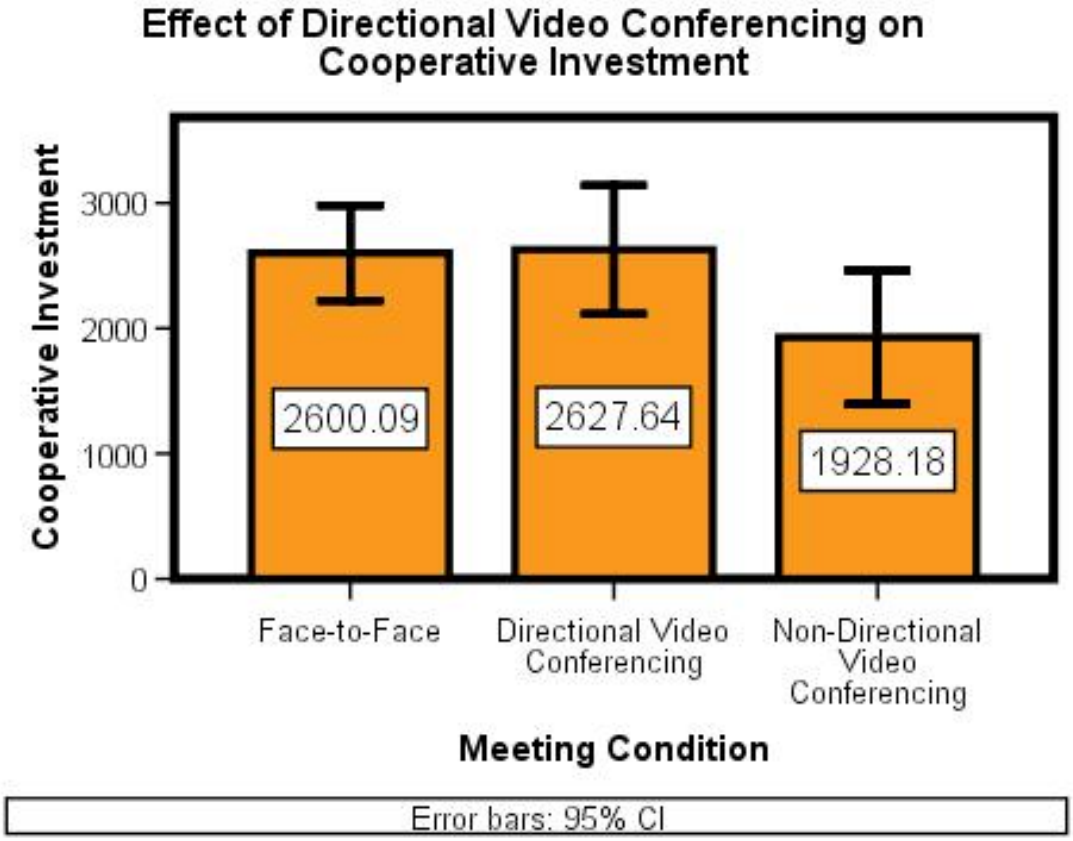


Figure 9.2: Overall cooperative investment by meeting condition.

Means for cooperative investment are shown in Figure 9.2 for each of the conditions. Three Planned Comparisons were performed using one-way analysis of variance. The analysis showed that *cooperative investment* by groups meeting face-to-face was significantly higher than by groups meeting through non-directional video conferencing, $F(1, 20) = 5.21, p < .05$. It also showed that cooperative investment by groups meeting through directional video conferencing was significantly higher than by groups meeting through non-directional video conferencing, $F(1, 20) = 4.42, p < .05$. No significant difference in cooperative investment was found between groups meeting face-to-face and groups meeting through directional video conferencing, $F(1, 20) = .01, p = .92$.

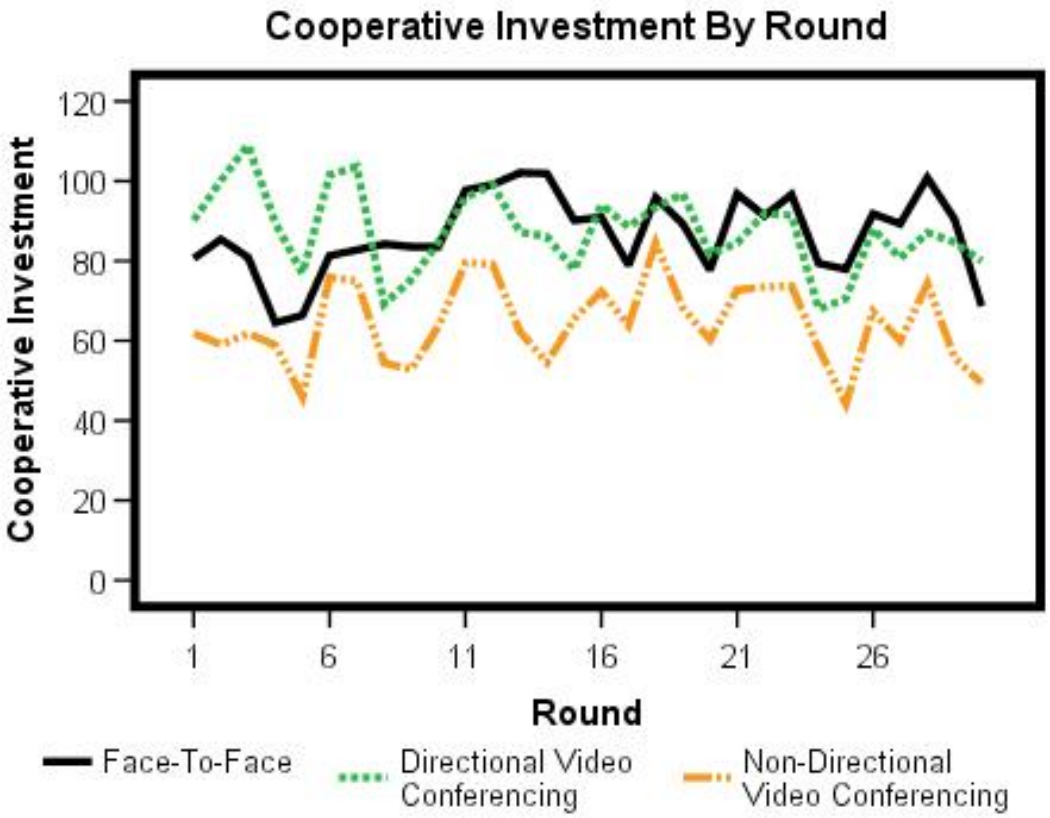


Figure 9.3: Cooperative investment amounts in each round.

9.6.2 Round-By-Round Cooperative Investment

Next I take a look at investments made round-by-round. For each round, both groups' cooperative investments were summed up. The maximum cooperative investment per round is 120 credits (60 credits/group * 2 groups). Figure 9.3 shows the average of all cooperative investment for rounds 1 through 30. Each line represents a different meeting condition.

Prior work[5] and the data presented in Figure 9.3 suggest Daytrader data exhibits two different phenomena: (1) *delayed trust*, which is a function of the number of rounds since the start of the game, and (2) *fragile trust*, which is a function of the number of rounds since the last discussion.

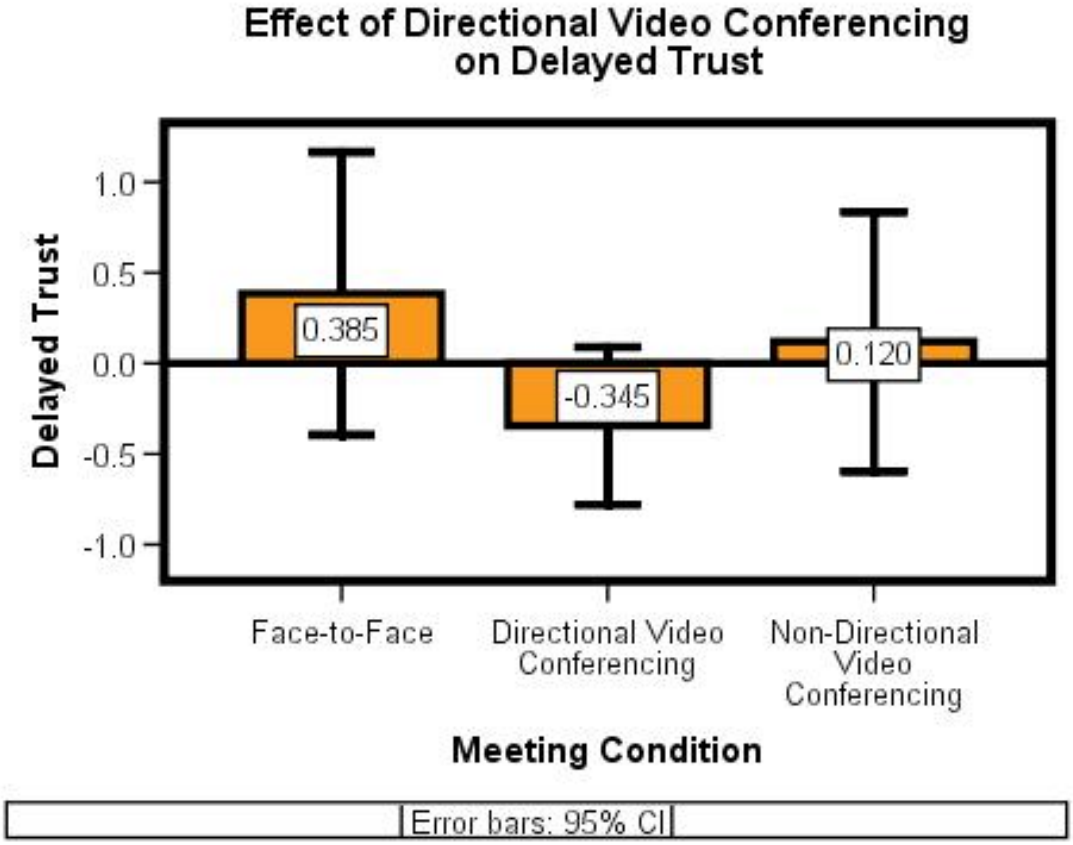


Figure 9.4: Delayed trust by meeting condition.

To use fragile trust in statistical analysis, Bos et al. [5] defined a new variable called *discussion distance*. It is the number of rounds since the last 5-round discussion. For example, round 6 and 11 both occur right after a discussion, so both rounds would have a discussion distance of 1.

The measure of *delayed trust* is the slope of a regression line between *cooperative investment* versus *round*. Delayed trust was calculated for each of the 33 sessions played. Discussion distance was added as a covariate to control for the effect of fragile trust. The means are presented in Figure 9.4. I performed three Planned Comparisons using one-way analysis of variance. The analysis showed no significant difference in delayed trust for any of the comparisons: (1) face-to-face versus non-directional video conferencing, $F(1, 20) = .31, p = .58$, (2) directional versus non-directional video conferencing, $F(1, 20) = 1.53, p =$

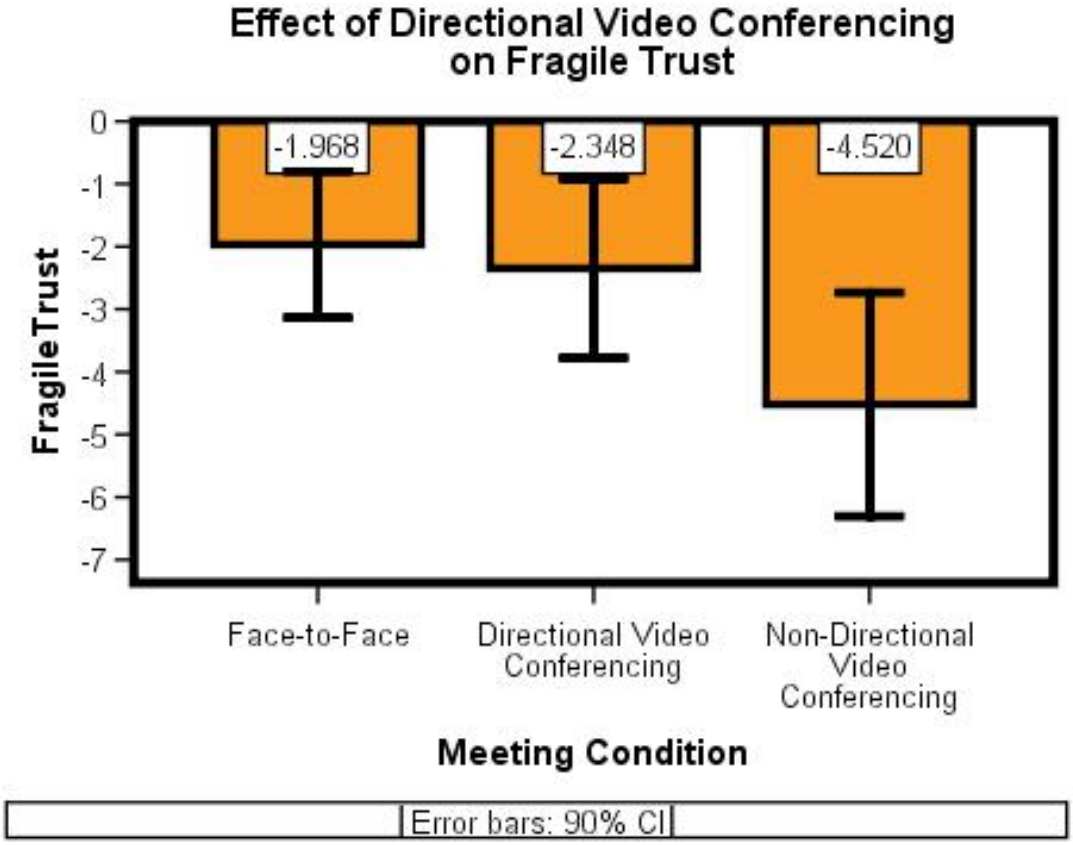


Figure 9.5: Fragile trust by meeting condition.

.23, and (3) face-to-face versus directional video conferencing, $F(1, 20) = 3.31, p = .08$. One-way t-tests show none of the delayed trust measurements were significantly different from 0: (1) face-to-face, $t(10) = 1.10, p = .30$, (2) directional video conferencing, $t(10) = -1.77, p = .11$, and (3) non-directional video conferencing, $t(10) = .37, p = .72$.

The measure of *fragile trust* is the slope of a regression line between *cooperative investment* versus *discussion distance*. Fragile trust was calculated for each of the 33 sessions played. Round number was added as a covariate to control for the effect of delayed trust. The means are presented in Figure 9.5. I performed three Planned Comparisons using one-way analysis of variance. The analysis showed trust in groups meeting face-to-face were significantly less fragile than in groups meeting though non-directional video conferencing, $F(1, 20) = 4.70, p < .05$. It also showed that trust in groups meeting through

directional video conferencing was significantly less fragile than in groups meeting through non-directional video conferencing at a more experimental level, $F(1, 20) = 2.96, p < .10$. No significant difference in fragile trust was found between groups meeting face-to-face and groups meeting through directional video conferencing, $F(1, 20) = .14, p = .71$.

9.6.3 Questionnaire

For each session, the responses to each questionnaire item given by all the participants from both groups were averaged to create an aggregate session response. One questionnaire was disregarded since that participant just circled the same number for all items. The questionnaire measured *trust in other group* (11 items, $\alpha = .96$), *trustworthiness* (5 items, $\alpha = .92$), and *consistency* (3 items, $\alpha = .65$). One item was removed from *consistency* to improve the internal consistency (2 items, $\alpha = .70$). Spearman rank correlation showed a significant and positive correlation between each of the conditions measured in the questionnaire with the total cooperative investment of the game: *trust in other group*, $\rho(31) = .57, p < .01$, *trustworthiness*, $\rho(31) = .50, p < .01$, and *consistency*, $\rho(31) = .48, p < .01$.

I then compared self-reported trust measures by meeting condition. The means are presented in Figure 9.6. I performed three Planned Comparisons using one-way analysis of variance. The analysis showed that groups meeting face-to-face self-reported significantly higher trust than groups meeting through non-directional video conferencing, $F(1, 20) = 12.61, p < .05$. It also showed that groups meeting through directional video conferencing self-reported significantly higher trust than groups meeting through non-directional video conferencing at a more experimental level, $F(1, 20) = 3.60, p < .10$. No significant difference in self-reported trust was found between groups meeting face-to-face and groups meeting through directional video conferencing, $F(1, 20) = .01, p = .94$. No significant differences were found between experimental conditions for any of the other questionnaire conditions measured. The results of these comparisons match the results of the comparisons of overall and fragile trust from the Daytrader measurements.

9.7 Discussion

On the basis of the above findings, I will now revisit the hypotheses set out earlier.

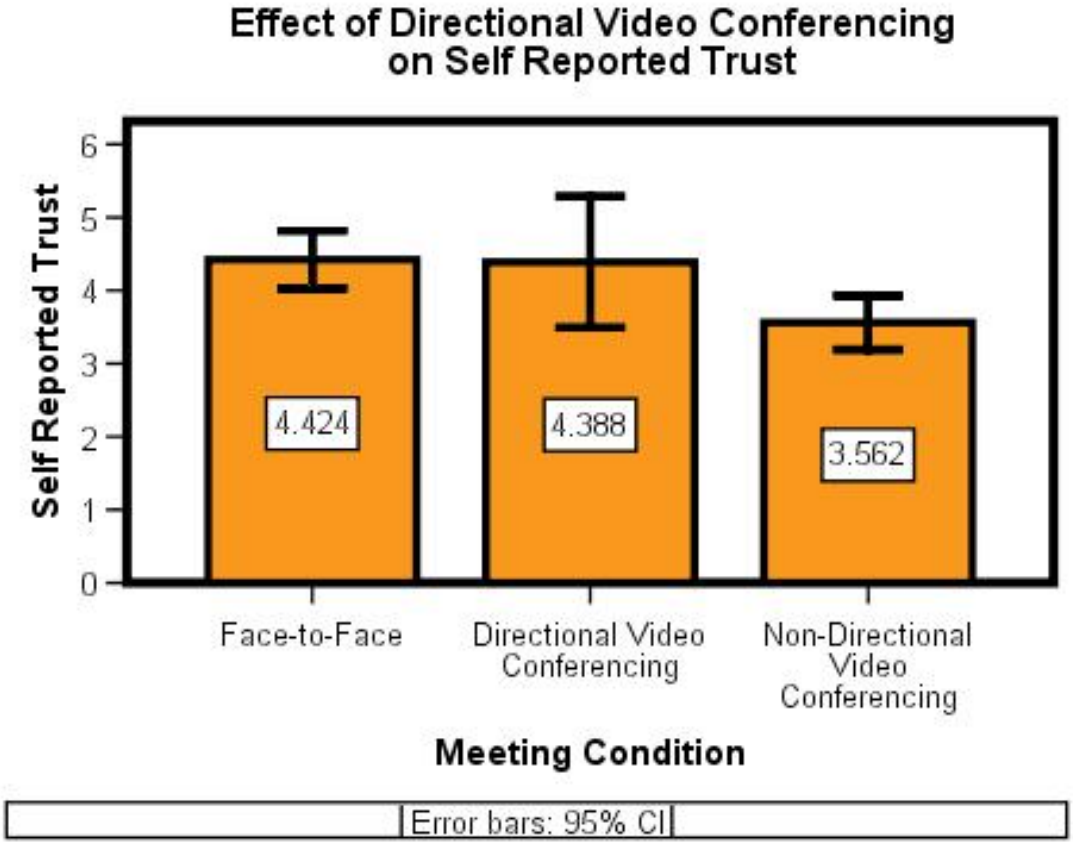


Figure 9.6: Self-reported trust by meeting condition.

For this experiment, I used a variant of Daytrader as a measure of trust. The results from the trust inventory show a positive and significant correlation between *investment amounts* and *trust scores* adding internal validity to Daytrader as a trust measurement device.

Hypothesis 1a (H1a): Groups meeting face-to-face will show higher levels of overall *trust than groups meeting through non-directional video conferencing.*

This hypothesis is supported on the basis of the descriptive statistics. In comparing the total cooperative investment amount by both groups in all 30 rounds, I found that the total cooperative investment by groups meeting face-to-face was significantly higher than those by groups using non-directional video conferencing.

Additionally, results from the questionnaire trust inventory show a statistically significant difference in the *trust in other group* condition between the face-to-face and non-directional video conferencing.

Hypothesis 1b (H1b): Groups meeting face-to-face will show reduced delay *in trust formation when compared to groups meeting through non-directional video conferencing.*

This hypothesis is not supported by the results and analysis. The results do not suggest a difference in delayed trust formation between face-to-face and non-directional meeting conditions. In fact, no change in trust was measured at all in either of these conditions. This may be the result of either a lack of the delayed trust phenomenon in group-to-group interactions or limitations in the power of the experiment for measuring delayed trust. Further studies would be needed to clarify this.

Hypothesis 1c (H1c): Groups meeting face-to-face will show reduced fragility *in trust formation when compared to groups meeting through non-directional video conferencing.*

This hypothesis is supported by the results and analysis. The results show that groups meeting face-to-face tended to be more resilient to breakdowns in trust when compared to groups that met through non-directional video conferencing.

Hypothesis 2a (H2a): Groups meeting through directional video conferencing will show higher levels of overall *trust than groups meeting through non-directional video conferencing.*

This hypothesis is supported on the basis of the descriptive statistics. In comparing the total cooperative investment amount by both groups in all 30 rounds across the three meeting conditions, I found that the total cooperative investments by groups using directional video conferencing was significantly higher than those by groups using non-directional video conferencing and that the investments in the directional video condition tended toward the investments made by those who met face-to-face when compared to non-directional video.

Additionally, results from the questionnaire trust inventory show a statistically significant difference in the *trust in other group* condition between the directional and non-directional video conferencing conditions. The self-reported trust level in the directional video conferencing condition tended toward the levels in the face-to-face condition which is in agreement with the hypothesis.

From the results, those that met face-to-face invested an average of 2600.09 credits. By using non-directional video conferencing, average cooperative investment reduced to 1928.18 credits, a reduction of 26%. Meeting through directional video conferencing system restored the average cooperative investment up to 2627.63 credits, similar to face-to-face levels. I are careful here not to claim that using a directional video conferencing system like MultiView will fully restore trust lost in using non-directional video conferencing systems, but I do present the lack of measurable difference as support for the dependence of trust on spatial faithfulness.

Hypothesis 2b (H2b): Groups meeting through directional video conferencing will show reduced delay *in trust formation when compared to groups meeting through non-directional video conferencing.*

Similar to the discussion for H1b, this hypothesis is not supported by the results and analysis. The results do not suggest a difference in delayed trust formation between directional and non-directional meeting conditions. No change in trust was measured at all in either of these conditions.

Hypothesis 2c (H2c): Groups meeting through directional video conferencing will show reduced fragility *in trust formation when compared to groups meeting through non-directional video conferencing.*

This hypothesis is supported by the results enough to warrant further exploratory work. The results show a statistically significant difference in fragile trust between the directional and non-directional video conferencing conditions at a reduced level of confidence ($p < .10$). Groups that met through directional video conferencing tended to be more resilient to breakdowns in trust when compared to groups that met through non-directional video conferencing. The measure of fragile trust tended toward that of face-to-face and there was no measurable difference between the face-to-face and directional video conferencing conditions.

9.8 Conclusion

For many different types of meetings, trust can play an important role. For these types of meetings, standard video conferencing systems will be less effective than other available alternatives. The use of standard video conferencing systems can significantly hinder the trust formation process in multiple-participant sites. The group-to-group

configurations in the study cooperatively invested less and trust was more fragile when meeting through non-directional video conferencing than when meeting face-to-face.

However, using a spatially faithful video conferencing, such as MultiView, helps improve the trust formation process. Groups meeting through directional video conferencing cooperated more than groups who met through standard video conferencing systems and were more resilient in their cooperation in the face of temptation. For all the measures of trust, there was no measurable difference in cooperative behavior between groups meeting face-to-face and groups meeting through MultiView.

What I have shown here is that spatial faithfulness is very important in the trust formation process and that a spatially faithful system should be used for high-stakes communication where trust is important.

Part IV

Conclusion

Video conferencing technology continues to evolve. However, the basic design of the system has remained relatively unchanged. Since its introduction at the World's Fair in 1964, engineers and researchers continue to push the boundaries on video conferencing design. We continue to see increasing quality in the network infrastructure, sound and display systems, and the corresponding audio and video processing techniques. Though progress continues, the design goals remain largely static. We continue to push for better audio and video performance without looking at the overall design of the video conferencing system and with lack of perspective on how people actually communicate and use it.

To make matters more complex, many goals may not even align with the goal of creating a better meeting experience. For instance, there is actually very little evidence that shows increasing the video resolution has any effect on the efficacy of the meeting in any measurable way yet we continue to invest resources on supporting higher resolutions with the move from QVGA to SD, and now from SD to high-definition (HD). Another example is the telephone that, in all its ubiquity, actually supports a very limited audio spectrum but remains a very useful everyday tool. I believe that a more radical design change is required.

In this dissertation, and with designing things in general, I have argued for the understanding of the related issues that are important to the overall goal, proposing designs that speak to those issues, and show that they actually do have an effect on communication. The goal of video conferencing is to allow two geographically separated parties to meet effectively. With the understanding that nonverbal cues can play a significant role in communication, this dissertation analyzes the mechanisms required for effective use of nonverbal cues with respect to how current video conferencing systems fail to support these mechanisms (Part I). I introduced a novel display (Part II) and follow up with studies that show that it works and makes a significant difference in factors that are important and in line with the overall goal of communication (Part III).

To outline the specific contributions I have made, I will overview each part. In Part I, I presented an analysis that allows us to define the spatial problems of group-to-group video conferencing that understands the way people perceive images. I presented the notion of *spatial faithfulness* in video conferencing systems that is a description of how well spatial information is supported using a video conferencing system. In defining spatial faithfulness, I introduced an*attention-based model* that allowed us to expand the scope of gaze analysis toward all nonverbal communication cues and include group-to-group structures instead of just dyadic pairs. I then introduced the *perceptual invariance* or the *Mona Lisa Effect* that

describes how users perceive flat images. I then presented the *view-presence model* that allowed us to look at and compare spatially relevant structures between video conferencing and face-to-face meetings. Using the view-presence model, I then defined a set of three spatial distortions to be solved by design: (1) the Collapsed Viewer Effect, (2) the Horizontal Parallax Effect, and (3) the Vertical Parallax Effect.

In Part II, I introduced a new design of a group-to-group video conferencing system that solves the three defined problems using a carefully crafted arrangement of cameras and the introduction of a multiple-viewpoint display. I then presented the iterations I went through as I improved the MultiView design and showed the details on how to engineer a multiple-viewpoint display based on a front projection system.

In Part III, I introduced several empirical methods and performed studies based on these methods to improve our understanding of the technological and the social implications of the MultiView design. I began by introducing a method for testing the spatial faithfulness of a video conferencing system and used it to demonstrate the spatial faithfulness of MultiView. During that same experiment, I elicited feedback from the users to figure out what aspects of the design needed to be improved in the third iteration. I then improved upon a method of measuring trust formation patterns, allowing us to experiment in the group-to-group setting. Using this method, I explored its support for trust formation and showed that spatial faithfulness is important for supporting the interaction needed in the formation of trust.

People communicate in powerful ways. Technology and people have come together to enhance communication in many ways. Nothing has been as powerful, though, as being able to immediately communicate with distant partners in real time. Because of the nature of people, there will always be ways to improve how I communicate. In this dissertation, I hope to have contributed to a theoretical understanding of human, nonverbal communication. I hope to have informed future design of communication technology and to have presented a new design of video conferencing system that helps harness the power of nonverbal communication.

Bibliography

[1] Kenneth R. Adams. Perspective and the viewpoint. *Leonardo*, 5(3):209–217, 1972.

[2] Michael Argyle, Mansur Lalljee, and Mark Cook. The effects of visibility on interaction in a dyad. *Human Relations*, 21(1):3–17, February 1968.

[3] Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, March 1981.

[4] Jonathan Bendor. Uncertainty and the evolution of cooperation. *The Journal of Conflict Resolution*, 37(4):709–734, December 1993.

[5] Nathan Bos, Judy Olson, Darren Gergle, Gary Olson, and Zach Wright. Effects of four computer-mediated communications channels on trust development. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 135–140, New York, NY, USA, 2002. ACM Press.

[6] Thomas A. Busey, Nuala P. Brady, and James E. Cutting. Compensation is unnecessary for the perception of faces in slanted pictures. *Perception & Psychophysics*, 48(1):1–11, 1990.

[7] John K. Butler. Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, 17(3):643–663, September 1991.

[8] Milton Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 49–56, New York, NY, USA, 2002. ACM Press.

[9] Joanie B. Connell, Gerald A. Mendelsohn, Richard W. Robins, and John Canny. Effects of communication medium on interpersonal perceptions: Don't hang up the phone yet! In *GROUP '01: Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, pages 117–124, New York, NY, USA, 2001. ACM Press.

[10] Karen S. Cook, Russell Hardin, and Margaret Levi. *Cooperation Without Trust?* Russell Sage Foundation, New York, 2005.

[11] Karen S. Cook, Toshio Yamagishi, Coye Cheshire, Robin Cooper, Masafumi Matsuda, and Rie Mashima. Trust building via risk taking: A cross-societal experiment. *Social Psychology Quarterly*, 68(2):121–142, 2005.

[12] James E. Cutting. Rigidity in cinema seen from the front row, side aisle. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):323–334, 1987.

[13] Judith Donath. Mediated faces. *Lecture Notes in Computer Science*, 2117:373–390, 2001.

[14] Paul Dourish, Annette Adler, Victoria Bellotti, and Austin D. Henderson. Your place or mine? learning from long-term use of audio-video communication. *Computer Supported Cooperative Work*, 5(1):33–62, 1996.

[15] Aimee L. Drolet and Michael W. Morris. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50, January 2000.

[16] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage.* W. W. Norton & Company, New York, August 2001.

[17] Jim Gemmell, Kentaro Toyama, Lawrence C. Zitnick, Thomas Kang, and Steven Seitz. Gaze awareness for video-conferencing: a software approach. *Multimedia, IEEE*, 7(7):26–35, October-December 2000.

[18] James J. Gibson and Anne D. Pick. Perception of another person's looking behavior. *American Journal of Psychology*, 76(3):386–394, Sept 1963.

[19] Russel Hardin. *Trust and Trustworthiness.* Russell Sage Foundation, New York, 2002.

[20] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[21] Wojciech Matusik and Hanspeter Pfister. 3d tv: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Trans. Graph.*, 23(3):814–824, 2004.

[22] Andrew F. Monk and Caroline Gale. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, 33(3):257–278, 2002.

[23] Ken I. Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. Multiparty videoconferencing at virtual social distance: MAJIC design. In *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 385–393, New York, NY, USA, 1994. ACM Press.

[24] Eric Paulos and John Canny. PRoP: personal roving presence. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 296–303, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[25] D. I. Perrett, J. K. Hietanen, M. W. Oram, and P. J. Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions: Biological Sciences*, 335(1273):23–30, January 1992.

[26] Willian Poundstone. *A Two Person Dilemma.* Anchor, New York, 1992.

[27] Jens Riegelsberger, Angela M. Sasse, and John D. Mccarthy. The researcher's dilemma: evaluating trust in computer-mediated communication. *International Journal of Human-Computer Studies*, 58(6):759–781, June 2003.

[28] Elena Rocco. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 496–502, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[29] Abigail Sellen, Bill Buxton, and John Arnott. Using spatial cues to improve videoconferencing. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 651–652, New York, NY, USA, 1992. ACM Press.

[30] John Short, Ederyn Williams, and Bruce Christie. *The Social Psychology of Telecommunications*. John Wiley and Sons Ltd, September 1976.

[31] Roel Vertegaal, Gerrit van der Veer, and Harro Vonsfect. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface*, pages 95–102, Montreal, Canada, 2000. Human-Computer Communications Society, Morgan Kaufmann Publishers.

[32] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 521–528, New York, NY, USA, 2003. ACM.

[33] Dhanraj Vishwanath, Ahna R. Girshick, and Martin S. Banks. Why pictures look right when viewed from the wrong place. *Nature Neuroscience*, 8(10):1401–1410, September 2005.