

BiDi Screen: Depth and Lighting Aware Interaction and Display

by

Matthew W. Hirsch

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

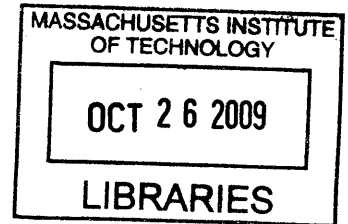
Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[September 2009]
August 2009

ARCHIVES



© Massachusetts Institute of Technology 2009. All rights reserved.

Author_____

Program in Media Arts and Sciences
August 13, 2009

Certified by_____

U U

Henry Holtzman
Research Scientist
Media Laboratory
Thesis Supervisor

Accepted by_____

Deb Roy
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

BiDi Screen: Depth and Lighting Aware Interaction and Display

by

Matthew W. Hirsch

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 13, 2009, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

In this thesis, I describe a new type of interactive display that supports both on-screen multi-touch interactions and off-screen hover-based gestures. This BiDirectional (BiDi) screen, capable of both image capture and display, is inspired by emerging LCDs that use embedded optical sensors to detect multiple points of direct contact. The key contribution of this thesis is to exploit the spatial light modulation capability of LCDs to allow dynamic mask-based scene capture without interfering with display functionality. A large-format image sensor is placed slightly behind the liquid crystal layer. By alternately switching the liquid crystal between a display mode showing traditional graphics and a capture mode in which the backlight is disabled and a pinhole array or an equivalent tiled-broadband code is displayed, the BiDi Screen can recover multi-view orthographic imagery while functioning as a 2D display. The recovered imagery is used to passively estimate the depth of scene points from focus. I discuss the design and construction of a prototype to demonstrate these capabilities in two motivating applications: a hybrid touch plus gesture interaction and a light-gun mode for interacting with external light-emitting widgets. The working prototype simulates the large format light sensor with a camera and diffuser, supporting interaction up to 50 cm in front of a modified 20.1 inch LCD.

Thesis Supervisor: Henry Holtzman
Title: Research Scientist, Media Laboratory

BiDi Screen: Depth and Lighting Aware Interaction and Display

by


Matthew W. Hirsch

The following people served as readers for this thesis:

Thesis Reader _____

Ramesh Raskar
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader _____


Joseph Paradiso
Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Acknowledgments

This work was conceived and developed in close collaboration with Prof. Ramesh Raskar, MIT Media Lab, and Douglas Lanman, a Ph.D. candidate at Brown University. Much of the theoretical analysis presented herein is reproduced from our SIGGRAPH Asia 2009 paper titled *BiDi Screen: Depth and Lighting Aware Interaction and Display* for which Douglas Lanman contributed the analysis. I sincerely thank Henry Holtzman, Prof. Ramesh Raskar and Douglas Lanman for their support and advice on completing this thesis. I also thank my friends and colleagues in the Information Ecology and Design Ecology groups, who have helped me keep my head in the clouds. Finally many thanks to Louise, who resents me only a little for all the late nights at the lab.

Contents

1	Introduction	13
1.1	A BiDirectional Display	14
1.1.1	New Interaction Modes	14
1.1.2	Video Chat and Appearance Capture	14
1.2	Problem and Approach	15
1.3	Contributions	16
2	Background	19
2.1	Multi-touch and Gesture	19
2.1.1	Light Sensitive Displays	22
2.1.2	Depth Sensitive and Light-field Cameras	22
2.2	Bi-directional Displays	24
2.3	Light-fields	24
2.3.1	Mask-Based Light-field Photography	27
3	Bidirectional Display Design	29
3.1	Design Parameters	29
3.2	LCD Overview	31
3.3	Hardware Design	32
3.4	Optical Design	35
3.4.1	Pinhole Arrays	35
3.4.2	Broadband Masks	39
3.5	Gesture Detection	40
3.5.1	Depth Extraction	40
3.5.2	Hand Tracking	41
3.5.3	Touch Estimation	42
4	Prototype: Construction and Performance	45
4.1	Implementation	45
4.1.1	Hardware	45
4.1.2	Software	47
4.2	Limitations	53
4.3	Validation	54
4.3.1	Resolution	54
4.3.2	Depth Resolution	55

4.3.3	Touch and Hover Discrimination	55
4.4	User Experience	57
5	Interaction Modes	59
5.1	Multi-Touch and Hover	59
5.1.1	Demonstrated	59
5.1.2	Future Work	60
5.2	Lighting Sensitive Interaction	62
5.2.1	Demonstrated	63
5.2.2	Future Work	63
5.3	Video and Appearance Capture	64
5.3.1	Demonstrated	65
5.3.2	Future Work	65
6	Conclusions	67
6.1	Future Work	67
6.1.1	Hardware	67
6.1.2	Variable Masks	68
6.1.3	Video	68
6.1.4	Scanning	69
6.1.5	Handheld	69
6.2	Synopsis	69
A	Optimizing Heterodyne Mask Properties	71
A.1	Forward	71
A.2	Pinhole Array Mask Configuration	72
A.3	Tiled-Broadband Mask Configuration	73

List of Figures

2-1	BiDirectional Screen Concept	21
2-2	Light-field: Single Plane Parametrization	25
2-3	Light-field: Two Plane Parametrization	26
3-1	Liquid Crystal Display Components	33
3-2	BiDi Screen Layout	34
3-3	Pinhole Camera Design	35
3-4	Multi-view Orthographic Imagery	37
3-5	Effective Spatial Resolution	38
3-6	Spatial and Angular Resolution Tradeoffs	39
3-7	Synthetic Aperture Refocusing	42
4-1	Prototype Output	48
4-2	Software Flow Control Diagram	49
4-3	Pinhole Calibration	50
4-4	Thread Queue Job Ordering	52
4-5	Performance Validation Results	54
4-6	Performance Validation Tests	56
5-1	Multi-touch Demonstration	60
5-2	World Navigator Demonstration	61
5-3	Model Viewer Demonstration	62
5-4	Lighting Demonstration	63
5-5	Image Capture Result	64
A-1	Effective Spatial Resolution	73
A-2	Tiled Broadband Code Theory	74

Chapter 1

Introduction

A novel method for using light sensors to detect multiple points of contact with the surface of liquid crystal displays (LCDs) is emerging. Sharp Corporation [7] and Planar Systems, Inc. [1] have demonstrated LCDs with arrays of optical sensors interlaced within the pixel grid. The location of a finger or stylus can be determined from the spatial position of occluded sensors that receive less light. For objects pressed directly against such screens, photographic imaging should be possible, but objects moved further away quickly become blurred as the light reflecting off any portion of the object is spread across many pixels.

In this thesis I describe how to modify traditional LCDs to allow both image capture and display. By using the LCD to display a pinhole array, or an equivalent tiled-broadband code [29], the angle and intensity of light entering a co-located sensor array may be captured. By correlating data from multiple views, it is possible to image objects, such as fingers, that are located beyond the display's surface and measure their distance from the display. This thesis describes the construction of a prototype device, in which imaging is performed in real-time, enabling the detection of off-screen gestures. When used with a light-emitting wand, the prototype can determine not only where the wand is aimed, but also the incidence angle of light cast on the display surface.

1.1 A BiDirectional Display

In this thesis I propose a *BiDirectional (BiDi)* screen, such that the entire surface of a thin, LCD-like device functions both as a display and a image capture device. The key component of a BiDi screen is a sensor array located slightly behind the spatial light modulating layer of a conventional LCD. The BiDi screen alternately switches between two modes: a display mode, where the backlight and liquid crystal spatial light modulator function as normal to display the desired output on the screen, and a capture mode where the backlight is disabled and the light modulator displays an array of pinholes or a tiled-broadband code.

I demonstrate a working prototype of a BiDi screen, substituting a diffuser and conventional cameras for the sensor array. This prototype shows the BiDi screen in two motivating applications to demonstrate its depth and lighting-aware abilities: a hybrid touch plus gesture interaction, and a light-gun mode for interaction using a light-emitting widget.

1.1.1 New Interaction Modes

While earlier light-sensing display designs have focused on enabling touch interfaces, the BiDi screen enhances the field by seamlessly transitioning from on-screen multi-touch to off-screen hover and gesture-based interaction. Thus, the proposed device alternates between forming the displayed image and capturing a modulated light field through a liquid crystal spatial light modulator.

The interaction capabilities of the prototype BiDi Screen are presented in four demonstrations. A model viewer application (Figure 5.1), a world viewer (Figure 5.1), light-emitting widget interaction (Figure 5.2), and multi-touch demonstration (see Figure 5.1)

1.1.2 Video Chat and Appearance Capture

In this thesis I primarily explore the BiDi screen as an interaction device. However, the ability of this device to produce video data from a source coincident with a display creates

the opportunity for new applications in video chat. Current state-of-the-art systems for video chat employ a camera that is necessarily offset from a display screen. This creates an eye-gaze problem, whereby a user looking at the display screen (and not at the camera) appears not to make eye contact with the other party in the conversation. At a basic level, the collocated camera of the BiDi screen can be used to provide a realistic remote interaction between two people, free from the eye-gaze problem previously described. The BiDi screen is situated among other methods for addressing this problem in Chapter 2. Further applications for a coincident camera and display include interactive virtual mirrors and augmented reality displays.

Capturing a light-field in a wide-baseline sensor means that the BiDi screen could function as a 3D appearance capture device. An object could be held up to the screen and rotated in order to record a light-field map of its appearance.

1.2 Problem and Approach

The BiDi screen will enable a new dimension in interaction, allowing a seamless transition from a traditional multi-touch display to a new hover-based interactive device (Section 5.1). The BiDi screen is further capable of dynamically relighting virtual scenes with real-world light sources (Section 5.2). This technique clears the way for novel mixed-reality rendering applications. By collocating an image capture device with a display device, the BiDi screen potentially solves the problem of establishing eye-contact during a video chat (Section 5.3). This ability will also allow for novel dynamic mirror-world renderings, in which the BiDi screen functions as a mirror, optionally altering the world before displaying it on the screen. Mirror-world rendering will enable interesting games and information displays that use the player's own body as a canvas. The BiDi screen further functions as a wide-baseline light-field sensor. This means that the screen can function in an appearance-capture mode, in which the 3D physical appearance of an object can be captured in a single shot (also Section 5.3).

The BiDi screen solves these problems by building on a trend in multi-touch technology.

Some manufacturers [1] [7] are creating optical multi-touch displays, which place an optical sensor in each pixel of an LCD device. Such displays can produce an image of objects in contact with the surface of the screen, enabling a multi-touch interface. These devices contain no optics or equivalent so they cannot bring objects at any distance from the screen into sharp focus.

The BiDi screen modifies an optical multi-touch screen by translating the transparent sensor plane a short distance behind the LCD display (Figure 3-2). If a tiled broadband mask is displayed on the LCD, the sensor layer will record a modulated light-field, as described by the spatial heterodyning technique [51]. In the BiDi screen the LCD is then put to double duty, in that the display and modulation functions are time multiplexed. The LCD alternately displays an image for the viewer while enabling the backlight, and displays a tiled broadband mask for light-field modulation while disabling the backlight, and recording from the sensor. When the display mode of the LCD is time multiplexed above the flicker fusion frequency of the human visual system [56], users will perceive a steady image on the screen.

In the prototype BiDi screen constructed for this thesis, a diffuser and camera system was used in place of a transparent sparse photosensor array. While the array would be preferable, practical constraints on time and budget required that the prototype be build from commodity hardware.

1.3 Contributions

The emphasis of this work is on demonstrating novel techniques for optical sensing enabled when an LCD and diffuse light-sensing grid are placed proximate to each other. As devices combining display and capture elements in close proximity are currently being developed commercially for multi-touch interaction, one goal is to influence the design of these displays by exploring design choices and illustrating additional benefits and applications that can be derived. This thesis only touches upon the interaction techniques enabled, leaving additional assessment to future work.

Earlier light sensing display designs have focused on promoting touch interfaces. The BiDi screen design enhances the field by supporting both on-screen multi-touch interactions and off-screen hover and gesture-based interfaces. A base contribution is that the LCD can be put to double duty; it can alternate between its traditional role in forming the displayed image and a new role in acting as an optical mask. It is shown that achieving per-pixel, depth and light aware interactions requires a small displacement between the sensing plane and the display plane. Furthermore, this work demonstrates a method that can maximize the display and capture frame rates using optimally light-efficient mask patterns.

In this thesis I describe a thin, lensless light field camera composed of an optical sensor array and a spatial light modulator. The performance of pinhole arrays and tiled-broadband masks for light field capture from primarily reflective, rather than transmissive, scenes is evaluated. I describe key design issues, including mask selection and the critical importance of angle-limiting materials.

I demonstrate that a BiDi screen can recognize on-screen as well as off-screen gestures, and its ability to detect light-emitting widgets, showing novel interactions between displayed images and external lighting.

I describe how the system can be expanded to support object appearance capture and novel video interactions with depth keying and eye contact between video chat participants.

Chapter 2

Background

The BiDi screen makes contributions in a diverse set of domains. In this chapter, I will situate the contributions of this thesis in the fields of touch and gesture interaction, specifically with respect to light sensing displays, depth sensing technology, collocated camera and display, and mask-based light-field photography.

2.1 Multi-touch and Gesture

Touch screens capable of detecting a single point of touch first appeared in a product in 1983 with Hewlett Packard's introduction of the HP-150 computer, although touch and pen systems existed earlier than this. Single-touch systems typically used transparent resistive or capacitive grids overlaid on a screen to sense touch. The pressure of a finger on the screen would induce a change in the resistance or capacitance of a few wires in the grid, allowing accompanying electronics to determine the position, and possibly pressure, of the touch.

Recently, much emphasis has been placed on enabling the detection of multiple simultaneous touches on a screen, which has enabled new types of interaction. The Frustrated Total Internal Reflection (FTIR) multi-touch wall [19], HoloWall [39], and Microsoft Surface use a camera to detect infrared light reflected from fingers in contact with a screen. The Visual

Touchpad [38], uses a visible light camera to track hands. In Tactex’s MTC Express [36] an array of pressure sensors is used to locate the position at which a membrane is depressed. Hillis [22] forms a 2D pressure sensing grid using force-sensitive resistors. A popular approach to multi-touch sensing of fingers and hand positions is through the use of capacitive arrays, described by Lee et al. [30] and made popular with the iPhone from Apple, Inc., following Fingerworks iGesturePad, both based on the work of Westerman and Elias [53]. The SmartSkin [44], DiamondTouch [12], and DTLens [16] also use capacitive arrays in various configurations.

The development of more sophisticated cameras and improved computational power has made real-time gestural interaction a possibility. Some of the most influential work in free-space gestural interaction includes the TouchLight [55], which uses a specialized depth-sensing IR camera to track gesture, and Oblong Industries g-speak, which uses a camera array to track tags that may be affixed to hands and other objects.

In a work closely-related to my own, the ThinSight [25] places a compact IR emitter and detector array behind a traditional LCD. This approach begins in the direction of combining multi-touch and free-space gestural interaction. Benko and Ishak [3] use a DiamondTouch system as well as 3D tracked gloves to achieve mixed multi-touch and gesture interaction. In contrast to capacitive and direct-contact sensors, a variety of methods have emerged for imaging through a display surface. Izadi et al. [26] introduced SecondLight as a modified rear-projection display with an electronically-switchable diffuser. In their design, off-screen gestures are imaged by one or more cameras when the diffuser is in the clear state. While supporting high-resolution image capture, SecondLight significantly increases the thickness of the display—placing several projectors and cameras far behind the diffuser. Similarly, DepthTouch [4] places a depth-sensing camera behind a rear-projection screen. While producing inferior image quality, the BiDi screen has several unique benefits and limitations with respect to such direct-imaging designs. Foremost, with a suitable large-format sensor, the proposed design eliminates the added thickness in current *projection-vision* systems, at the cost of decreased image quality.



Figure 2-1: Towards a BiDirectional (BiDi) screen for walk-up 3D interaction with flat screens. (Left) A conceptual sketch of public use with direct multi-touch plus non-contact hover gestures. Embedding a large-format optical sensor backplane is becoming possible. (Right, Top) Emerging LCDs [7] with co-located optical sensors (Right, Bottom) capable of capturing sharp images of fingers in direct contact, but blurred for hovering parts.

2.1.1 Light Sensitive Displays

In work that has deeply inspired the design of the BiDi screen, Sharp and Planar have demonstrated LCD prototypes with integrated optical sensors co-located at each pixel for inexpensive multi-touch interaction. These optical touch screens are the first example of a thin, portable display technology that can measure incident light on a per-pixel basis.

Coarser lighting sensitive displays have been used for a range of purposes prior to this. Some portable electronics, including laptops and mobile phones, use ambient light sensors to adjust the brightness of the display depending on the lighting environment. Nayar et al. [41] proposed creating spatially-selective lighting sensitive displays (LSD) by placing optical sensors within the display bezel and altering the rendered imagery to accurately reflected ambient lighting conditions. Cossairt et al. [10] implemented a light field transfer system, capable of co-located capture and display, to facilitate real-time relighting of synthetic and real-world scenes. Fuchs et al. [17] achieved a passive lighting sensitive display capable of relighting pre-rendered scenes printed on static masks. Unlike their design, the BiDi screen works with directional light sources located in front of the display surface and can support relighting of dynamic computer-generated scenes.

2.1.2 Depth Sensitive and Light-field Cameras

Many depth sensing techniques exist which could be for the purpose of gestural interaction. Passive optical ranging methods include multi-view stereo [46] and depth from focus [40]. Active optical ranging methods include laser striping [6], time-of-flight cameras (e.g., Canesta’s CANESTAVISION chip [23] and 3DV’s Z-Cam [24]), depth from defocus [52], and structured lighting [45].

A wide variety of passive and active techniques are available to estimate scene depth in real-time. The BiDi screen records an incident light field [32] using a variety of attenuating patterns equivalent to a pinhole array. A key benefit is that the image is formed without refractive optics. Similar lensless systems with coded apertures are used in astronomical

and medical imaging to capture X-rays and gamma rays. Zomet and Nayar [59] describe a lensless imaging system composed of a bare sensor and several attenuating layers, including a single LCD. This system does not capture a light-field, but rather explores the space of modified apertures in camera systems. Liang et al. [34] use temporally-multiplexed attenuation patterns, also displayed with an LCD, to capture light fields. They use a translated pinhole to capture a light-field. In contrast to the broadband mask approach used in the BiDi screen, this would not be capable of capturing images in realtime. Both Liang et al [34]. and Zoment and Nayar [59] focus on building devices in traditional camera bodies, which are not intended to be coupled with display or used for interaction. Zhang and Chen [58] recover a light field by translating a bare sensor. Levin et al. [31] and Farid [13] use coded apertures to estimate intensity and depth from defocused images. Vaish et al. [50] discuss related methods for depth estimation from light fields. In a closely-related work, Lanman et al. [29] demonstrated a large-format lensless light field camera using a family of attenuation patterns, including pinhole arrays, conceptually similar to the heterodyne camera of Veeraraghavan et al. [51]. I use the tiled-broadband codes introduced in those works to reduce the required exposure time. Unlike Veeraraghavan et al. [51] and Lanman et al. [29], the design presented here exploits a mask implemented with a modified LCD panel. In addition, the BiDi screen uses reflected light with uncontrolled illumination.

Pinholes and masks can be used at other positions in an optical train to achieve different results. For example, a pair of pinhole apertures are used in confocal microscopy to increase contrast and scan a 3D target by eliminating rays that originate from points on the target that are out of focus. Coded aperture imaging, also discussed by Veeraraghavan et al. [51], places a coded mask in the aperture of a traditional camera system to enable deconvolution of focus blur after an image has been captured, or measurement of scene depth, depending on the type of code used.

2.2 Bi-directional Displays

Attempts to create a system with a coincident camera and display have aimed to solve the gaze-direction problem in video chat, described in Section 5.3. One example of such a system is patented by Apple Computer[21], and uses a spinning wheel with transparent and mirrored sections to switch a screen between the optical path of a projector and a camera. Systems that use beam splitters or mechanical time multiplexers typically extend the optical path length of the system significantly. The system described by Apple Computer is not able to measure depth. Another coaxial configuration comes from a second Apple Computer patent [48], describing an LCD screen with an array of conventional cameras behind it, or interleaved between rows of display pixels. This configuration has never been manufactured on a wide scale, and no details have been publicly released by Apple about the performance or other specifications of any actual prototypes that have been built. It should be noted that this device should in theory be capable of capturing the incident light field, though it will not sample it as uniformly as the BiDi screen. Since each of the sensors in the Apple device requires a lens, the portion of the screen occupied by each small camera cannot be used for display causing unpleasant visual artifacts in the displayed image.

Software attempts to simulate the effect of coaxial imaging and display devices for the specific case of video-conference gaze correction[11] are fundamentally limited to correcting the specific off-axis imaging problem for which they were designed. Software correction approaches have limitations on the types of images they can correct, and the accuracy of the correction, as there is information lost at the time of capture that cannot be recovered in software.

2.3 Light-fields

One simplification that can often be applied to optical systems is to consider a ray-based light transport system. In the realm of geometric optics, the propagation of light through a system can be described by considering the plenoptic function [2], lumigraph [18], or light-

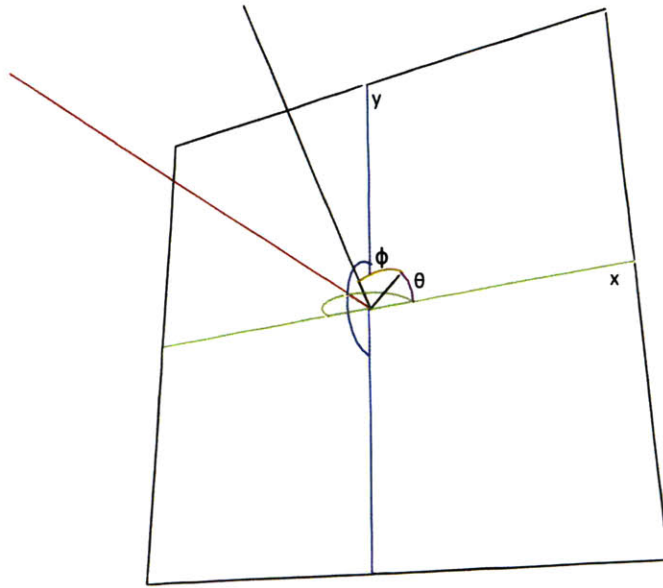


Figure 2-2: Single plane parametrization of the light-field. The position of intersection of a ray with the $x - y$ plane, defined above by the x -axis (green) and y -axis (blue), and the angle in two dimensions θ_x (orange), and θ_y (purple), describes the set of rays in 3D space.

field[32]. These functions parametrize the radiance of light along rays. The contemporary conception of the light-field was first described in 1996 in the papers by Levoy [32] and Gortler [18]. Earlier references to the idea of capturing a full description of set of rays crossing a plane date to 1908, when Lippmann [35] described a system consisting of an array of pinhole cameras which was able to make measurements of a light-field similar to those recorded by the BiDi screen.

In 3D-space, when occluders are neglected, the light-field can be parametrized with four variables. It is useful to consider two such parametrization. In the case of a single plane parametrization, a ray can be uniquely identified by its point of intersection with a 2D plane, (x, y) , and by its angle of intersection with the plane in two directions, θ_x and θ_y , shown in Figure 2-2.

A second method of parametrize the light-field in 3D-space employs two parallel, two-dimensional, planes. The ray is uniquely identified in space by its point of intersection with the first plane, (u, v) , and its point of intersection with the second plane, (s, t) . This also results in a four dimensional function of radiance, as shown in Figure 2-3.

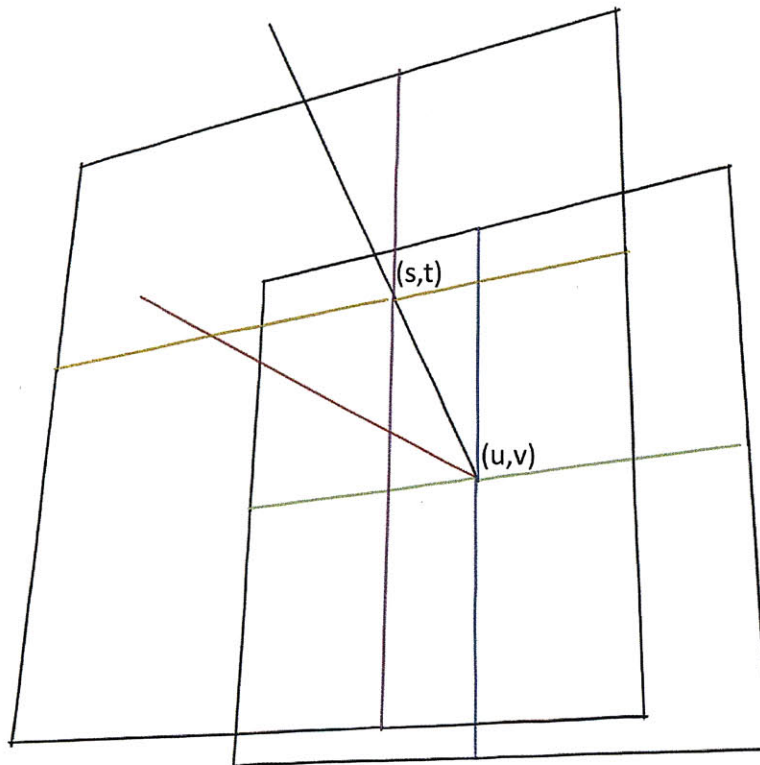


Figure 2-3: Two plane parametrization of the light-field. The position of intersection of a ray with the $u - v$ plane, defined above by the u -axis (green) and v -axis (blue), and the $s - t$ plane, defined above by the s -axis (orange) and t -axis (purple), describes the set of rays in 3D space.

The concept of a light-field is used for many computer graphics rendering applications, as described by Levoy [33]. In the field of computational photography, there have arisen numerous methods for sampling the light-field generated by real-world objects. For example, Wilburn [54], and others [57], describe arrays of cameras capable of sampling a light-field, and Vaish [49] describes a computer system capable of refocusing the captured light-field in real-time. Ng [42] describes a single, hand-held camera which uses an array of lenses, or lenslets, on top of a sensor to enable a range of light-field sampling patterns. Lenslet arrays have now become a popular way to sample the light-field. Methods that do not employ additional lenses have also been described. Zomet uses an LCD in front of a camera sensor to capture a light field [60]. This design can scan a single transparent pixel across the LCD, effectively time-multiplexing an array of pinhole cameras. A class of light-field sampling devices exists which does not employ lenses and does not rely on time-multiplexing: mask-based light-field photography.

2.3.1 Mask-Based Light-field Photography

Mask-based light-field capture uses a tiled broadband mask offset from the sensor plane to create shifted copies of the light-field frequency spectrum. The principle exploited by this technique is the same one used by AM radio to shift audio-frequency signals (20kHz) to broadcast frequency signals (eg. 1000kHz). For this reason, Veeraraghavan, first to describe the technique, refers to this work as Spatial Heterodyning [51]. The flatland case is depicted in Figure A-2.

Though Veeraraghavan used a lens in their mask-based imaging system, it can be shown that a sensor and mask alone can capture a light-field that is sufficiently band limited. Lanman demonstrates a system that captures a light-field-like quantity without the use of lenses [29]. Lanman et. al. also show the optimality of the tiled-Modified Uniform Redundant Array (MURA) code, for light efficiency.

The BiDi screen builds on the work of Veeraraghavan and Lanman to create a wide-baseline lensless light-field sensor, capable of sampling the light-field generated by reflected light from

an arbitrary scene. The BiDi screen is the first to use a dynamic mask (an LCD screen, in this case) to perform mask-based light-field photography.

Chapter 3

Bidirectional Display Design

It is increasingly common for consumer electronics devices that have the ability to display images to also be able to capture them. Four basic goals motivated the design of the BiDi screen:

1. Capture 3-D to enable depth and light aware interaction.
2. Prevent image capture from interfering with image display.
3. Support walk-up interaction (i.e., no implements or markers).
4. Allow for mobility and portability.

3.1 Design Parameters

After careful consideration of the related work and possible image capture options, I believe that the approach taken toward the BiDi screen is uniquely positioned to satisfy the above goals. In this section I contrast the chosen approach to other possible methods.

A core design decision was to use optical sensing rather than capacitive, resistive, or acoustic modalities. While such technologies have been effectively used for multi-touch input, they do not allow for 3-D capture and off-screen gestures. Some capacitive solutions permit the

detection of an approaching off-screen finger or hand, but they can not accurately or repeatably determine the distance of the approaching object. Nor do any of these technologies support lighting aware interactions.

Optical sensing can be achieved in various ways. In many of the cases of previous work, cameras image the space in front of the display. The result is either a specially crafted environment, similar to g-speak, where multiple infrared cameras are positioned around the user and special gloves with high contrast markers must be worn; or, the display housing becomes large to accommodate the cameras, such as with Microsoft's Surface.

Another issue with using standard cameras is the trade-off between placing them behind, to the side, or in front of the display. If the camera is behind the display, then it will interfere with the backlighting, casting shadows and causing noticeable variations in the display brightness. Han's FTIR sensor, SecondLight, and DepthTouch all avoid this problem by using rear projection onto a diffuser rather than an LCD display, at the cost of increased display thickness. If the camera is located in front of the display or to the side, then it risks being occluded by users. Placing a camera in front of the display is also problematic as it demands a fixed installation—violating the design goals. Cameras could be placed in the bezel, looking sideways across the display, but that would also increase the display thickness and still suffer from user self-occlusion events. A half-silvered beam-splitting mirror, placed at a 45 degree angle to the display surface, could be used to give a side-mounted camera a clear view of the interaction space, but this would both greatly increase the depth of the display and also prevent the user from directly touching the screen.

In contrast, the approach described here is to use a sparse array of photodetectors, located behind the individual display pixels. Because the array is sparse, it will not block much of the light from the backlight and any attenuation will be evenly distributed across the screen. Being behind the display, it does not suffer from occlusions caused by the users. The detector layer can be extremely thin and potentially optically transparent (using thin film manufacturing processes), supporting the goal of portability. These are all design attributes we share with the multi-touch displays being contemplated by Sharp and Planar. However, the display additionally requires a small gap between the spatial light modulating and light

detecting planes. This gap, combined with the techniques described in Section 3.3, allows for the measurement of the angle of incident light, as well as its intensity, and thereby the capture of 3-D data.

It should be noted that for the purpose of the prototype, I sacrificed the goal of portability in exchange for ease of fabrication with commodity hardware. In particular, I simulated the sparse array of photo detectors by using a diffuser screen imaged by standard CCD cameras from behind.

3.2 LCD Overview

An LCD is composed of two primary components: a backlight and a spatial light modulator. A typical backlight consists of a cold cathode fluorescent lamp (CCFL), a light guide, a diffuser, and several brightness enhancing films (BEF). The overall function of these layers is to condition the light produced by the CCFL such that it is spatially uniform and collimated [28]. A key role is played by the backlight diffuser; by randomizing both the polarization state and angular variation of transmitted and reflected rays, the diffuser greatly increases the efficiency of the backlight, allowing light rays to be “recycled” by reflecting between the various layers until they satisfy the necessary collimation and polarization conditions.

The spatial light modulator of an LCD is composed of three primary components: a pair of crossed linear polarizers and a layer of liquid crystal molecules sandwiched between glass substrates with embedded electrode arrays [56]. The polarizer closest to the backlight functions to select a single polarization state. When a variable electric field is applied to an individual electrode (i.e., a single display pixel), the liquid crystal molecules are reconfigured so that the incident polarization state is rotated. The polarizer closest to the viewer attenuates all but a single polarization state, allowing the pixel to appear various shades of gray depending on the degree of rotation induced within the liquid crystal layer. Color display is achieved by embedding a spatially-varying set of color filters within the glass substrate. To achieve wide-angle viewing in ambient lighting, a final diffuser, augmented

with possible anti-reflection and anti-glare films, is placed between the last polarizer and the viewer.

3.3 Hardware Design

As shown in Figure 3-2, the BiDi screen is formed by repurposing typical LCD components such that image capture is achieved without hindering display functionality. The first step is to exclude certain non-essential layers, including the CCFL/light guide/reflector components, the various brightness enhancing films, and the final diffuser between the LCD and the user. In a manner similar to [29], a large-aperture, multi-view image capture device is created by using the spatial light modulator to display a pinhole array or tiled-broadband mask. The key insight is that, for simultaneous image capture and display using an LCD, the remaining backlight diffuser must be moved *away* from the liquid crystal. In doing so, a coded image equivalent to an array of pinhole images is formed on the diffuser, which can be photographed by one or more cameras placed behind the diffuser. The backlight display functionality is restored by including an additional array of LEDs behind the diffuser.

One important consideration is that an angle-limiting material or other source of vignetting is critical to achieving image capture using the BiDi screen. In practice, the reflected light from objects in front of the screen will vary continuously over the full hemisphere of incidence angles. However, as I describe in the following sections, the proposed image capture scheme assumes light varies only over a limited range of angles—although this range can be arbitrarily large. An angular-limiting film could be placed in front of the BiDi screen, however such a film would also limit the field of view of the display. In the prototype design, the cameras are placed about one meter behind the diffuser. Since the diffuser disperses light into a narrow cone, the diffuser and cameras act together to create a vignetting effect equivalent to an angle-limiting film.

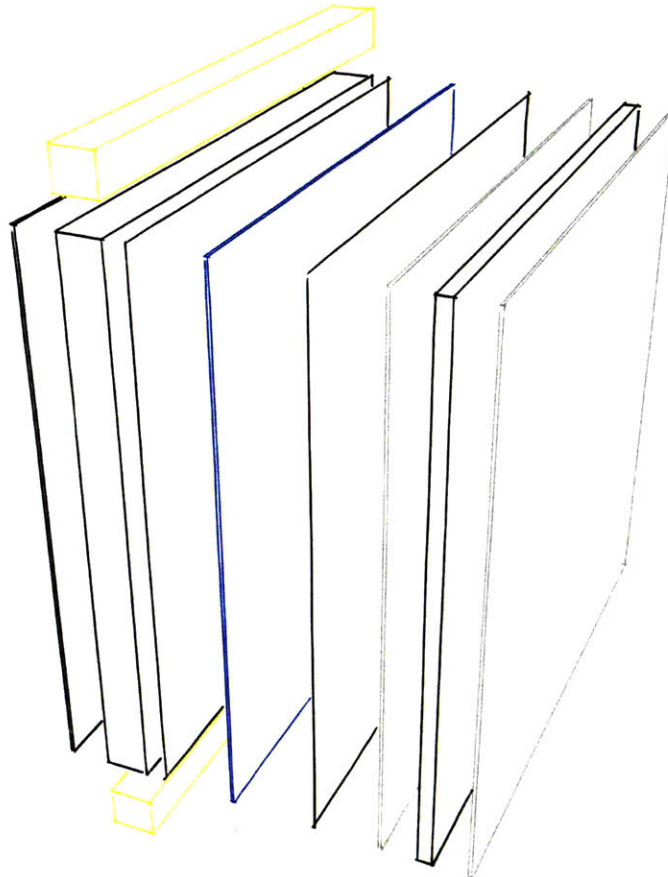


Figure 3-1: Typical LCD components. An LCD is disassembled to display the arrangement of the individual optical components. From left to right: (a) rear reflector, (b) the CCFL/light guide illumination source, (c) strong diffuser, (d) brightness enhancing prism film, (e) weak diffuser, (f) reflective rear polarizer (attached with adhesive to liquid crystal layer) (g) the liquid crystal spatial light modulator, and (h) the front polarizer/diffuser (attached with adhesive to liquid crystal layer). While not identical, most modern LCDs have a similar set of layered components.

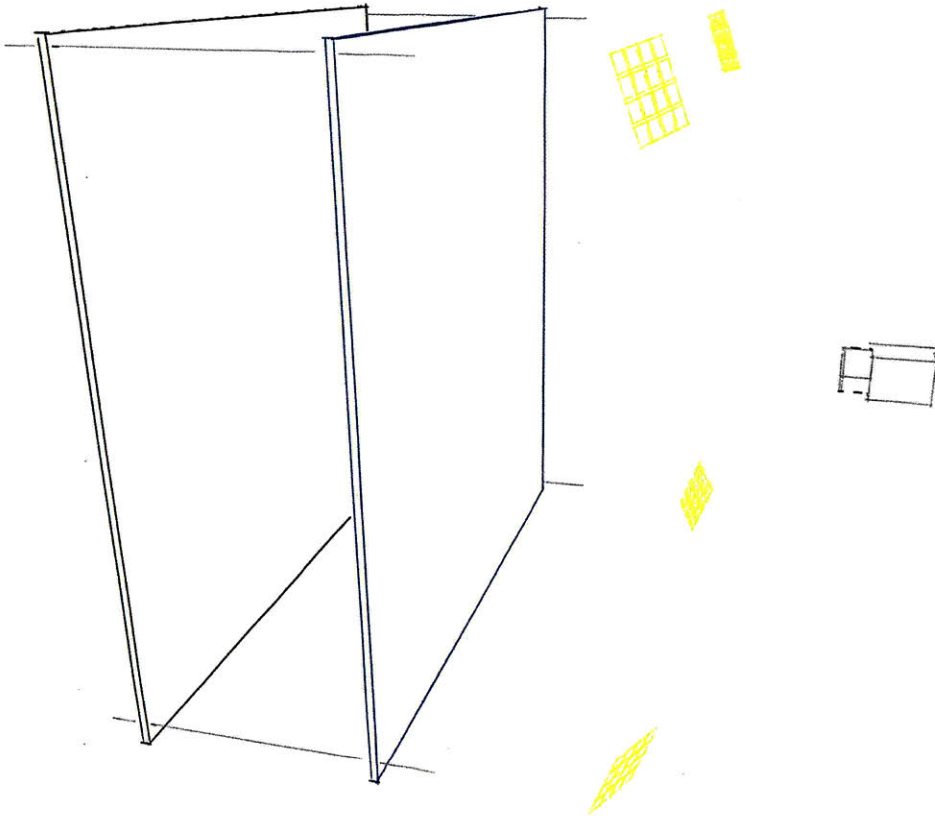


Figure 3-2: This figure depicts two possible BiDi screen configurations. In both configurations an LCD spatial light modulator is placed at the front plane (black). In the configuration used in the BiDi screen prototype described in Chapter 4, the back plane (blue) is a diffuser, imaged by a camera (gray). In an alternate configuration, the back plane (blue) is a wide-area sensor, and the camera is not required.

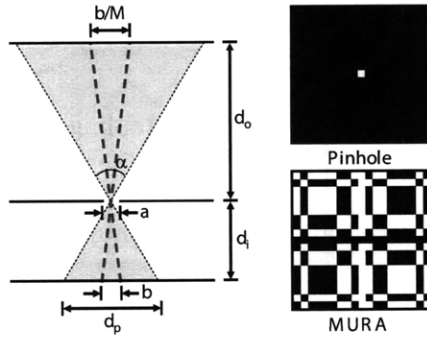


Figure 3-3: Design of a pinhole camera. (Left) The PSF width b is given by Equation 3.1 as a function of sensor-pinhole separation d_i , object distance d_o , and the aperture a . The PSF width is magnified by $M = d_i/d_o$ in the plane at d_o . (Right, Top) A single pinhole comprises an opaque set of 19×19 cells, with a central transparent cell. (Right, Bottom) We increase the light transmission by replacing the pinhole with a MURA pattern composed of a 50% duty cycle arrangement of opaque and transparent cells. As described by Lanman et al. [29] and earlier by Fenimore and Cannon [14], this pattern yields an equivalent image as a pinhole.

3.4 Optical Design

The above design goals require sufficient image resolution to estimate the 3-D position of points located in front of the screen, as well as the variation in position and angle of incident illumination. As described by [51], the trade-off between spatial and angular resolution is governed by the pinhole spacing (or the equivalent size of a broadband tile) and by the separation between the spatial light modulator and the image plane (i.e., the diffuser). As with any imaging system, the ultimate spatial and angular resolution will be limited by the optical point spread function (PSF). In this section I analyze the optimization of a BiDi screen for both on-screen and off-screen interaction modes under these constraints for the case of a pinhole array mask. In Section 3.4.2 I extend this analysis to the case of tiled-broadband masks.

3.4.1 Pinhole Arrays

Multi-View Orthographic Imagery: As shown in Figure 3-4, a uniform array of pinhole images can be decoded to produce a set of multi-view orthographic images. Consider the

orthographic image formed by the set of optical rays perpendicular to the display surface. This image can be generated by concatenating the samples directly below each pinhole on the diffuser plane. Similar orthographic views, sampling along different angular directions from the surface normal of the display, can be obtained by sampling a translated array of points of the diffuser-plane image offset from the center pixel under each pinhole.

On-screen Interaction: For multi-touch applications, only the spatial resolution of the imaging device in the plane of the display is of interest. For a pinhole mask, this is simply the total number of displayed pinholes. Thus, to optimize on-screen interactions the pinhole spacing should be reduced as much as possible (in the limit displaying a fully transparent pattern) and the diffuser brought as close as possible to the spatial light modulator. This is precisely the configuration utilized by the existing optical touch sensing displays by Brown et al. [7] and Abileah et al. [1].

Off-screen Interaction: To allow depth and lighting aware off-screen interactions, additional angular views are necessary. First, in order to passively estimate the depth of scene points, angular diversity is needed to provide a sufficient baseline for triangulation. Second, in order to facilitate interactions with an off-screen light-emitting widget the captured imagery must sample a wide range of incident lighting directions. As a result, spatial and angular resolution must be traded to optimize the performance for a given application. Off-screen rather than on-screen interaction is the driving factor behind the decision to separate the diffuser from the spatial light modulator, allowing increased angular resolution at the cost of decreased spatial resolution with a pinhole array mask.

Spatio-Angular Resolution Trade-off: Consider the design of a single pinhole camera shown in Figure 3-3, optimized for imaging at wavelength λ , with circular aperture diameter a , and sensor-pinhole separation d_i . The total width b of the optical point spread function, for a point located a distance d_o from the pinhole, can be approximated as

$$b(d_i, d_o, a, \lambda) = \frac{2.44\lambda d_i}{a} + \frac{a(d_o + d_i)}{d_o}. \quad (3.1)$$

Note that the first and second terms correspond to the approximate blur due to diffraction

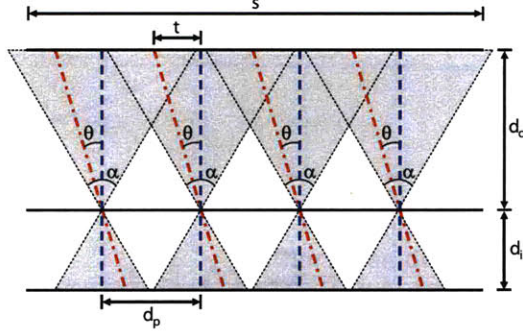


Figure 3-4: Multi-view orthographic imagery from pinhole arrays. A uniform array of pinhole images (each field of view shaded gray) is resampled to produce a set of orthographic images, each with a different viewing angle θ with respect to the surface normal of the display. The set of optical rays perpendicular to the display surface (shown in blue) is sampled underneath the center of each pinhole. A second set of parallel rays (shown in red) is imaged at a uniform grid of points offset from the center pixels under each pinhole.

and the geometric projection of the pinhole aperture onto the sensor plane, respectively [20]. If it is assumed that each pinhole camera has a limited field of view, given by α , then the minimum pinhole spacing d_p is

$$d_p(d_i, d_o, a, \lambda, \alpha) = 2d_i \tan\left(\frac{\alpha}{2}\right) + b(d_i, d_o, a, \lambda). \quad (3.2)$$

Note that a smaller spacing would cause neighboring pinhole images to overlap. As previously described, such limited fields of view could be due to vignetting or achieved by the inclusion of an angle-limiting film. Since, in the proposed design, the number of orthographic views $N_{angular}$ is determined by the resolution of each pinhole image, one can conclude that the angular resolution of the system is limited to the width of an individual pinhole image (equal to the minimum pinhole spacing d_p) divided by the PSF width b as follows.

$$N_{angular}(d_i, d_o, a, \lambda, \alpha) = \frac{d_p(d_i, d_o, a, \lambda, \alpha)}{b(d_i, d_o, a, \lambda)}. \quad (3.3)$$

Now consider an array of pinhole cameras uniformly distributed across a screen of width s and separated by a distance d_p (see Figure 3-4). Note that a limiting field of view is necessary to prevent overlapping of neighboring images. As described in Section 3.5.1, the BiDi screen uses a depth from focus method to estimate the separation of objects from the

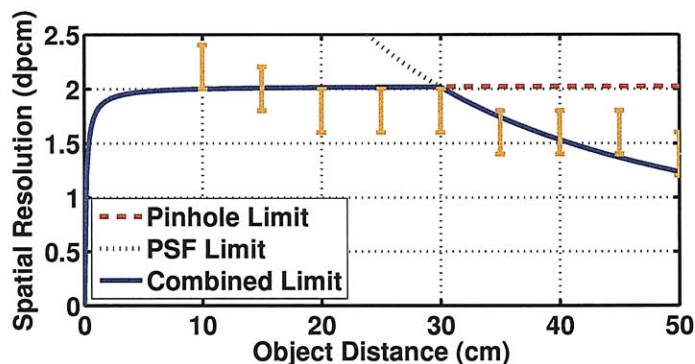


Figure 3-5: Effective spatial resolution as a function of distance d_o from the display. The effective spatial resolution in a plane at d_o (given in dots per cm) is evaluated using Equation 3.4. System parameters correspond with the prototype. Orange error bars denote the experimentally-estimated spatial resolution described in Section 4.3. Note that, using either dynamically-shifted masks or a higher-quality image sensor, the spatial resolution could significantly increase near the display (approaching the higher limit imposed by the optical PSF).

display surface. As a result, the system components should be placed in order to maximize the effective spatial resolution in a plane located a distance d_o from the camera. The total number of independent spatial samples $N_{spatial}$ in this plane is determined by the total number of pinholes and by the effective PSF for objects appearing in this plane, and given by

$$N_{spatial}(d_i, d_o, a, \lambda, \alpha; d_p, b) = \min \left(\frac{s}{d_p}, \frac{d_i s}{d_o b} \right), \quad (3.4)$$

where the first argument is the total number of pinholes and the second argument is the screen width divided by the magnified PSF evaluated in the plane at d_o . Thus, the effective spatial resolution is given by $N_{spatial}/s$. Note that, since the BiDi screen is orthographic, it is assumed the object plane at d_o is also of width s .

As shown in Figure 3-5, the effective spatial resolution in a plane at d_o varies as a function of the object distance from the pinhole array. For small values of d_o , the resolution monotonically increases as the object moves away from pinholes; within this range, the spatial resolution is approximately equal to the total number of pinholes divided by the screen width. For larger values of d_o , the resolution monotonically decreases; intuitively, when objects are located far from the display surface, neighboring pinholes produce nearly identical images. As described in Appendix A, the sensor-mask (or diffuser-mask) sepa-

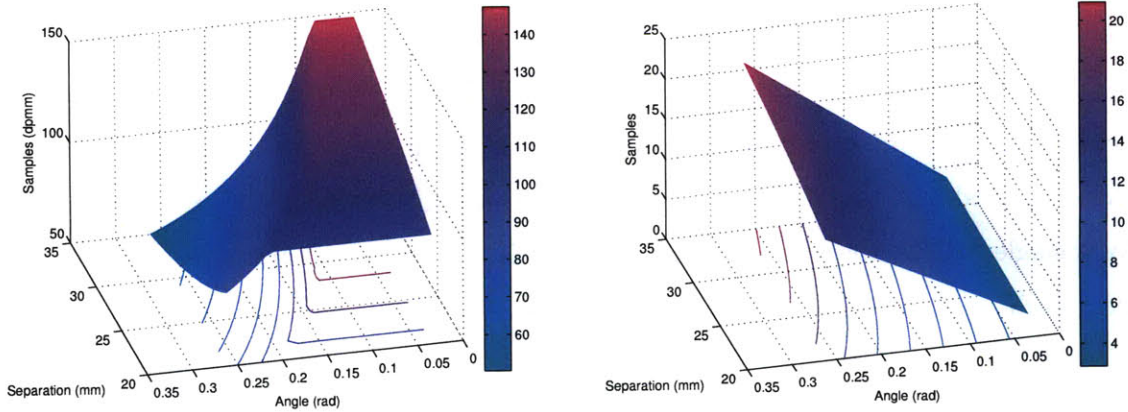


Figure 3-6: (Left) The number of spatial samples measured and (Right) the number of angular samples per pixel measured by varying d_i and α , with $d_o = 250\text{mm}$, $a = 256\ \mu\text{m}$, and $s = 487.2\text{mm}$. In the above plots, the mask spacing d_p was maximized for a given α . The parameter space in this problem is large, with tradeoffs existing between screen (mask) resolution, sensor resolution, mask pattern size (pinhole spacing), sensor-mask separation, and spatial and angular resolution (ability to resolve depth) and field of view. Note that regions of parameter space above that maximize spatial samples reduce angular samples and vice-versa.

ration is selected to maximize the effective spatial resolution located within 50 cm of the display surface. Note that, in Figure 3-5, the resolution close to the pinhole array drops dramatically according to theory. However, in practice the resolution close to the display remains proportional to the number of pinholes. This is due to that fact that, in the prototype, the pinhole separation d_p is held constant (as opposed to the variable spacing given in Equation 3.4). Practically, the vignetting introduced by the diffuser and camera's field of view prevents overlapping views even when an object is close to the screen—allowing for a fixed pinhole spacing.

3.4.2 Broadband Masks

The primary limitation of a pinhole array is severe attenuation of light. For example, in the proposed system a pinhole array is created by separating each pinhole by 18 LCD pixels, both horizontally and vertically. As a result, only approximately 0.2% of incident light reaches the diffuser (or sensor array in an ideal design). To overcome this attenuation,

extremely bright external lighting would be required for real-time interaction. Such lighting would significantly impair image display, due to strong glare and reflections. Fortunately, the LCD can be used to display arbitrary 24-bit RGB mask patterns. As a result, the generalized tiled-broadband masks described by Lanman et al. [29] are used. Specifically, the code used is a tiled-MURA code, as shown in Figure 3-3. Each pinhole is replaced by a single MURA tile of size 19×19 LCD pixels. Because the MURA pattern is binary (i.e., each pixel is either completely transparent or opaque) with a 50% duty cycle, the tiled-MURA mask transmits 50% of incident light. Assuming the cameras have a linear radiometric response, a tiled-MURA mask allows the external lighting to be dimmed by a factor of 180 (in comparison to pinhole array masks).

The heterodyne decoding method of Veeraraghavan et al. [51] is used to decode the diffuser-plane image, yielding orthographic multi-view imagery equivalent to that provided by a pinhole array mask. The use of tiled-broadband codes, however, does require additional computation to decode the diffuser-plane image and introduces additional noise from the decoding process. Note that the spatio-angular resolution trade-off for such tiled-broadband codes are similar to those described in the previous section for pinhole arrays—yielding a multi-view orthographic image array with similar spatial and angular sampling rates. Additional details on such design issues are provided in Appendix A.

3.5 Gesture Detection

3.5.1 Depth Extraction

As described in Chapter 2, there are a wide variety of methods to passively estimate depth from multi-view imagery. The BiDi screen employs a depth from focus method inspired by [40]. In their approach, a focal stack is collected by focusing at multiple depths within the scene. Afterwards, a per-pixel focus measure operator is applied to each image in the focal stack, with the assumption that each image patch will appear with highest contrast when the camera is focused at the depth of the patch. In the implementation described

in Chapter 4 a simple smoothed gradient magnitude focus measure was used. Finally, a coarse depth map can be obtained by evaluating the maximum value of the focus measure for each pixel. While modern depth from focus/defocus methods include more sophisticated focus operators, the approach used here can easily be evaluated in real-time on commodity hardware (see Chapter 4).

In order to obtain the set of refocused images (i.e., the focal stack), I apply methods from synthetic aperture photography [50]. As shown in Figure 3-4, when considering the intersection of the optical rays with a plane at distance d_o , each orthographic view, whether captured using pinhole arrays or tiled-broadband codes, is translated from the central view by a fixed amount. For an orthographic view rotated by an angle θ from the display's surface normal, the translation $t(\theta)$ will be given by

$$t(d_o, \theta) = d_o \tan(\theta). \quad (3.5)$$

In order to synthetically focus at a display d_o , I follow the computationally-efficient approach of Ng [43]; rather than directly accumulating each orthographic view, shifted by $-t(d_o, \theta)$, the Fourier Projection-Slice Theorem is applied to evaluate refocused images as 2D slices of the 4D Fourier transform of the captured light field. Typical refocusing results are shown in Figure 3-7. Because the refocusing operation is performed on a two-times unsampled grid, the process achieves a small degree of super-resolution through refocusing. The results shown indicate that web-cam quality images are possible from a device matching the specifications of the prototype constructed (See Chapter 4).

3.5.2 Hand Tracking

Hand tracking is performed using the depth map described in the previous section as input. The OpenCV blob-tracking library is used to segment regions of low distance values in the depth map corresponding to hands from the rest of the depth map. The Mean Shift method [9], combined with an automatically updated foreground mask was found to give the best results. The OpenCV library refers to this as the MSFG method. A Kallman filter, provided

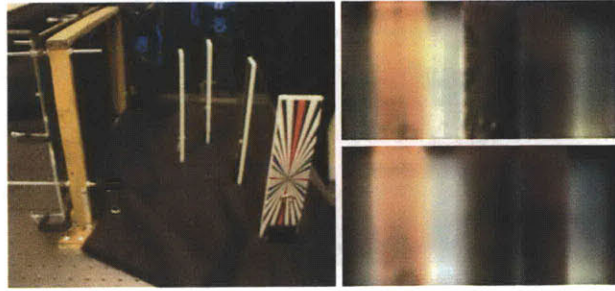


Figure 3-7: Synthetic aperture refocusing with orthographic imagery. (Left) A scene composed of three textured cards (center card has a printed resolution test chart shown facing camera on the left). (Right) Synthetically refocused images at a distance d_o of 10 cm and 15 cm shown on the top and bottom, respectively.

by the OpenCV blob-tracking library, was used to filter the resulting position data.

The method described here has some disadvantages over other types of gesture tracking. It cannot determine the pose or orientation of the hand, only the position over the screen. This method would likely be used as the first step in a more sophisticated gesture tracking pipeline. The method has the advantage that it easily runs in realtime, and was already implemented in free software. The method works sufficiently well in the prototype described in Chapter 4 to drive the demonstrations described in Chapter 5.

3.5.3 Touch Estimation

There exist a wide variety of methods for extracting touch events from a device like the BiDi screen. It would certainly be possible to cover the surface of the BiDi screen with a resistive or capacitive touch screen. However, this would lead to increased complexity and cost. To reduce cost and complexity, it will be most efficient to use the optical sensor of the BiDi screen to locate touch events. A rudimentary method of touch estimation is to find the average depth value of each blob used in hand tracking (Section 3.5.2), and determine touch events when this average value falls below a threshold.

For a detected blob with mean position (x, y) in the depth map, ROI R , and number of

non-zero elements, d , contained in R , N_R , the mean depth value, D , is given by

$$D = \frac{1}{N_R} \sum_{i \in R} d_i \quad (3.6)$$

A touch is detected when $D \leq D_{\text{thr}}$, where D_{thr} is a depth threshold that can be tuned. It should be noted that a gap exists between the front of the LCD and the transparent acrylic sheet used to protect the surface of the LCD from direct touches. This gap prevents fingers that touch the screen from occluding all ambient lighting.

The above approach is useful in its extreme simplicity, and worked sufficiently well for the prototype discussed in Chapter 4. However, it has significant disadvantages. The center of a detected blob is rarely aligned with the extended finger performing the actual touch, meaning that this method further reduces the resolution of the depth map, allowing for only very coarse touch detection. Also, using this method, touch events are dependant not only on the distance of the tip of a finger to the surface of the screen, but also on the distance of the rest of the hand. This is undesirable, since it is an unexpected result for users (see the discussion of user experience in Section 4.4).

A plausible extension to the method described here would resolve the above issues without incurring significantly more processing cost. Once candidate blobs have been selected for touch detection using the threshold applied to Equation 3.6, the raw modulated light field data can be used as input to a touch detection algorithm inspired by touchlib (a library for creating multi-touch interaction surfaces by tracking blobs of reflected light) or similar libraries. Without the influence of the mask function, which can be removed with a low-pass filter, the user's hand is effectively viewed through a diffuser. This is consistent with many of the optical touch detection approaches described in Chapter 2. The depth map blob could provide an ROI into the raw data to further reduce computation time and spurious results.

Chapter 4

Prototype: Construction and Performance

4.1 Implementation

4.1.1 Hardware

As shown in Figure 3-2, the prototype BiDi screen was constructed by modifying a Sceptre X20WG-NagaII 20.1 inch LCD with a 2 ms response time. The spatial light modulator was separated from the backlight, and the front diffuser/polarizer was removed. The weak diffuser was retained from the backlight and placed at $d_i=2.5$ cm from the liquid crystal layer on the side opposite the user. The front polarizer of the LCD was replaced with a linear polarizing polyvinyl alcohol-iodine (PVA) filter placed in direct contact with the diffuser. Commercial LCD screens typically combine the front polarizer with a diffusing layer, as was done on the X20WG. A diffuser in the plane of the spatial light modulator would interfere with the image capture mechanism. In order to easily mount the replacement polarizer on the correct side of the screen, the LCD was mounted backwards, so that the side typically facing the user was instead facing inward towards the backlight. The backlight functionality was restored by replacing the CCFL/light guide/reflector component with a

set of 16 Luxeon Endor Rebel cool white LEDs, each producing 540 lumens at 700 mA, arranged evenly behind the LCD. The LEDs were strobed via the parallel port to allow them to be shut off during the capture frame.

A pair of Point Grey Flea2 video cameras were placed 1 m behind the diffuser, each imaging approximately half of the diffuser while recording a 1280x960 16-bit grayscale image at 7 fps. For some interaction sessions, the cameras were operated in 8-bit grayscale mode. The camera shutters were triggered from the parallel port in order to correctly synchronize image capture with the LCD frame updates and LED strobing. Image capture and display was performed on an Intel Xeon 8 Core 2.66 GHz processor with 4 GB of system RAM and an NVIDIA Quadro FX 570 graphics card. The refocusing, depth estimation, and lighting direction estimation pipeline was capable of processing raw imagery at up to 7.5 fps.

The prototype was constructed on a standard optical bench, which provided a surface to precisely mount each of the elements of the optical stack described above. A custom frame was constructed to hold the naked LCD screen. The frame consists of two wooden beams bolted vertically to the optical bench. Slits were cut into the inner faces of the beams, allowing the screen to be slid in between them. Bolts were run through the beams in the direction into the plane of the LCD screen. These bolts were used to suspend the remainder of the optical stack at the desired distance from the LCD. A protective clear acrylic sheet was suspended between the user and the LCD screen to protect the delicate rear-polarizer from touches. This protective layer serves a double purpose, also preventing the LCD from being fully shaded from external lighting when a user is in contact with the screen.

External lighting was provided by overhead halogen lamps when the MURA code was used. Capturing images with a pinhole mask required an additional halogen stage lamp placed above the region in front of the LCD. This lighting was sufficient for imaging gestures and objects placed far from the display (e.g., the textured cards in Figure 3-7).

Both pinhole arrays and tiled-MURA codes were displayed on the LCD, with the latter used for real-time interaction and the former for static scene capture. Both the pinholes and MURA tiles repeated every 19×19 LCD pixels, such that $d_p = 4.92$ mm with a square

pinhole aperture of $a = 256 \mu\text{m}$. Following the derivation in Section 3.4, the acquired light field had a maximum spatial resolution of 88×55 samples (in the plane of the LCD) and an angular resolution of 19×19 samples spanning ± 5.6 degrees perpendicular to the display surface. The actual spatial resolution recorded was 73×55 samples, due to redundant measurements in the area of the screen viewed by both cameras. While narrow, this field of view and the limited spatial resolution was sufficient to provide robust refocusing and depth estimation (see Figures 3-7 and 4-5).

During interactive operation, three frames were sequentially displayed: a MURA code, and two display frames. The screen was refreshed at an average rate of 21 Hz and images were captured by the cameras each time a MURA frame was displayed. This results in the 7 fps capture rate described above. For static scene capture a sequence of two frames, a pinhole mask, and a “black” background frame were captured. Because timing was unimportant in a static scene, the frame rate of the pinhole capture sequence was adjusted according to scene lighting to allow for a sufficiently long camera exposure time. Background subtraction was used to mitigate the effects of the limited contrast achieved by the spatial light modulator for large incidence angles [56].

4.1.2 Software

As described above, the BiDi screen is controlled by a desktop computer running Windows XP 64-bit. The software controlling the BiDi screen, known as bidicam, was written in C++ and compiled with Microsoft Visual C++. The rendering and image capture from the cameras are controlled from a single class called `Video`. The graphics generated by the BiDi screen are rendered in OpenGL [47]. The Glut toolkit [27], provided by NVIDIA, is used to provide the rendering context for OpenGL, and the main loop for the bidicam application. Images are read from the cameras using the Point Grey flycapture library, version 2.

A Glut timer callback function orchestrates the backlight, camera triggering, and screen updates. The cameras and lights are both triggered by a signal sent from the parallel port



Figure 4-1: 3D interaction with thin displays. Using the methods presented in this thesis I describe a novel modification of an LCD to allow collocated image capture and display. (Left) An example of a mixed on-screen touch and off-screen hover interaction. Virtual models are controlled by the user’s hand movement. Touching a model brings it forward from the menu, or puts it away if already selected. Once selected, free space gestures control model rotation and scale. (Middle) Multi-view orthographic imagery recorded in real-time using a mask displayed by the LCD. (Right, Top) Image refocused at the depth of the hand on the right side. Note that the other hand, which is closer to the screen, is defocused. (Right, Bottom) Real-time depth map, with near and far objects shaded green and blue, respectively.

of the computer. When a MURA mask is required, it is generated on the fly using NVIDIA’s Cg shader language [15]. The basic structure of the software is shown in Figure 4-2.

The OpenCV and FFTW libraries are used to perform the image processing required in decoding the modulated light field, extracting a depth map, and tracking gesture. These calculations take place in a processing thread, which is launched as the application starts. As captured frames become available for processing, the processing thread is activated and performs heterodyne decoding. Depending on the interaction mode, the processing thread can be configured to perform image refocusing, and depth from focus using a maximum contrast operator, in order to create a depth map. For gestural interaction, the processing thread can be configured to additionally perform blob tracking, using the OpenCV library’s Mean Shift with Foreground (MSFG) method [9]. The processed results, such as spatial coordinates of detected hands, are stored to be used as an input to the OpenGL rendering.

The processing thread for the pinhole decoding has an analogous structure. However, instead of performing a frequency domain heterodyne reconstruction, the pinhole pipeline performs a simple re-binning operation, driven by a calibration file.

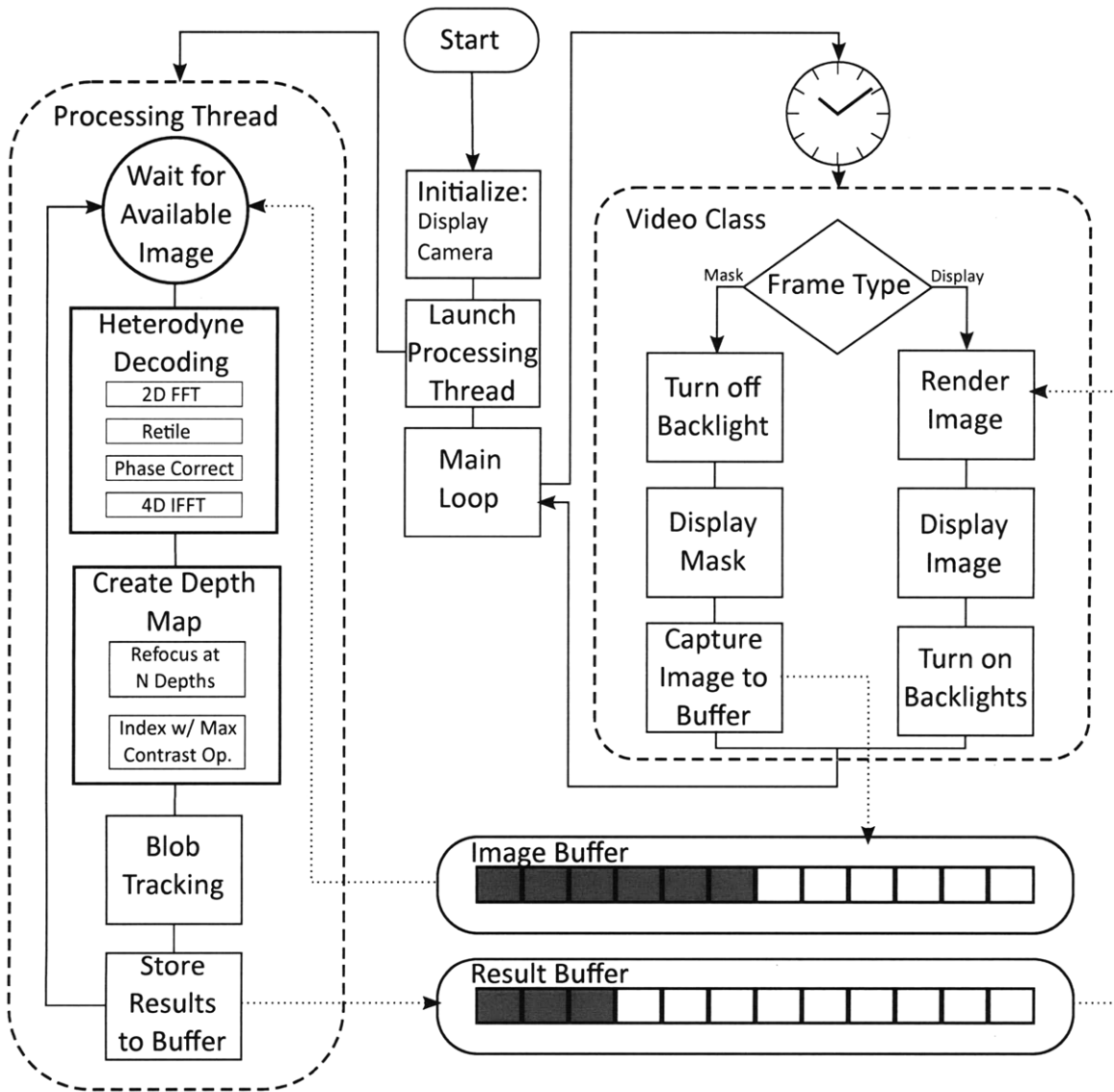


Figure 4-2: This schematic drawing closely depicts the structure of the software that controls the operation of the BiDi Screen. The main GLUT loop coordinates the elements on the right of the diagram. An processing thread runs in parallel to support the calculations shown on the left.

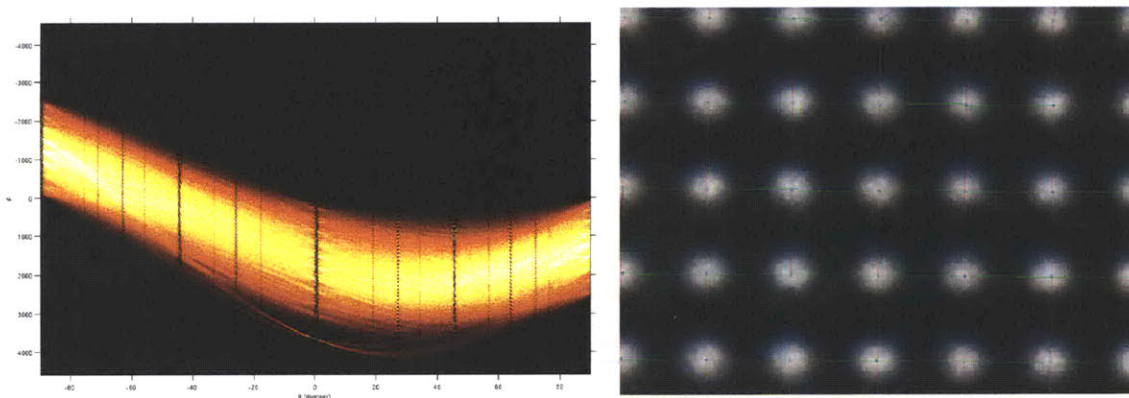


Figure 4-3: (Left) The Hough transform of the lightbox imaged by a pinhole array. Maximums occur along at points representing rows and columns of lightbox images. (Right) An annotated lightbox image from a pinhole array. The centers of the pinholes are located at the intersections of the row and column maximums in the Hough transform.

Calibration

Calibration for both the pinhole and heterodyne decoding pipelines is achieved using a semi-automatic scripted MATLAB procedure. When calibrating the system, a light box is placed very near the front of the BiDi screen, and an image is captured from the cameras. In the case of the pinhole array mask, a series of dots are produced, due to per-pinhole vignetting. The maximum intensity of each dot will appear under the location of each pinhole in the pinhole array mask.

The pinhole calibration MATLAB script performs a linear Hough transform to extract the most likely positions of vertical and horizontal lines of pinholes (Figure 4-3). The intersections of these lines are assumed to be the locations of the pinholes in the array. This method automatically orders the located pinholes. The coordinates obtained from this MATLAB script are saved into a MATLAB `.mat` file, which is read directly by the `bidicam` application. This calibration method can correct for rotation and linear scale distortions, but cannot correct for radial (pincushion or barrel) distortion. In the constructed prototype a high quality, 16mm lens is used, which results in negligible radial distortion. This method is robust to small deviations in lightbox angle, as the lightbox is a wide-angle uniform source.

The heterodyne calibration MATLAB script operates on the transformed modulated light-box image. The measured signal will be a convolution of the broadband lightbox signal with the tiled broadband MURA mask. By the duality property of the Fourier transform, the transformed image will contain a tiled delta function. The calibration script requires the user to hand tune rotation and scale parameters in order to visually match a point cloud to the tiled delta function. The calibration additionally requires the user to choose a phase shift which creates a centered vignetting function in the reconstructed image of the light box. This method is also robust to small changes in lightbox orientation. The lightbox is a broadband signal which is insensitive to a phase shift.

Heterodyne Decoding

The heterodyne decoding section of the processing thread follows directly from the technique described by Lanman [29], and is shown in Figure A-2. The raw input image is transformed using a 2D forward FFT. The heterodyne technique produces a constellation of slices of shifted copies of the light field spectrum. In the frequency domain, tiles are chosen surrounding each shifted spectral copy slice. The location of the tile center is chosen based on the calibration described in Section 4.1.2. The tiles are rearranged into a 4D function, representing the transform of the light-field spectrum. The tiles are phase corrected to remove the influence of the MURA mask and produce the calibrated vignetting function. This 4D function is then inverse transformed using a backward 4D FFT to obtain the light-field.

Parallelism

In order to obtain results in real-time, it was necessary to exploit the parallelism available in modern multi-core processors. This was achieved in various ways throughout the processing pipeline. The pinhole path was not used for realtime interaction, as it required long camera exposures. Therefore, the optimizations described here only apply to the heterodyne

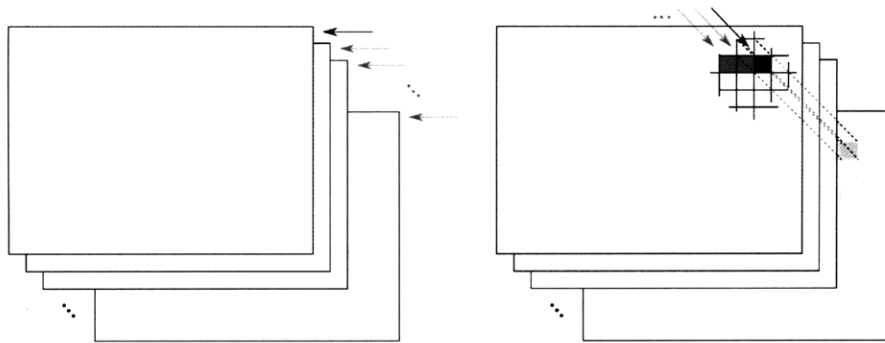


Figure 4-4: (Left) Jobs that process an entire image at a time are submitted such that each image in the focal stack is processed in parallel. (Right) Jobs that compare pixels through the focal stack are submitted such that the pixel columns through the stack are processed in parallel.

pipeline. The structure of the code that implements the parallelism described here is shown in Figure 4-2.

The simplest place to take advantage of parallelism was in the FFT and IFFT required by the heterodyne decoding procedure. The FFTW library used to perform these transforms natively supports multiple threads with proper configuration. The FFTW library can be configured to test many permutations of the division of labor between different threads to solve the transform requested to find the most efficient permutation for a particular piece of hardware.

Thread queues were used to take advantage of parallelism in the depth from focus estimation routine. A thread queue creates a pool of worker threads which may be assigned jobs from a pool of work items. The first pool of work items calls an OpenCV function for each refocused image to find the gradient magnitude of the image. These can be called in parallel, as the results do not depend on one another (ordering depicted in Figure 4-4, left). The same process is repeated to smooth the gradient magnitude image with a local averaging box filter. Once these jobs have completed, the job queue is loaded with jobs which travel through the stack of smoothed gradient magnitude images (ordering depicted in Figure 4-4, right), determining first the focus confidence (standard deviation) and then assembling a depth map.

4.2 Limitations

The proposed system is currently constrained to operate within the range of specifications for available consumer off-the-shelf technology, which places a lower limit on the size of the pixels used in the LCD and the sensor. In turn these components limit the maximum angular and spatial resolutions of the device, as described in Section 4.1. Experimentally, the diffuser PSF did not significantly limit the effective sensor resolution. Note, however, that the diffraction term in Equation 3.1 was a significant factor. The $256 \mu\text{m}$ pixels in the display, each color sub-pixel being a third of this width, resulted in a diffraction width of about $400 \mu\text{m}$ in the diffuser plane.

The BiDi Screen is optimized for real-time interaction, rather than single-shot image capture. The current implementation uses a pair of synchronized video cameras and a diffuser to simulate the performance of an embedded optical sensor array. The frame rate of the current apparatus is limited to 7.5 fps by the performance of currently available video cameras and the maximum transfer rate allowed by the 1394b firewire bus.

External lighting was required to provide sufficient object light during image capture. Efficient mask patterns, such as the tiled-MURA pattern, allow external lighting to be reduced. The intensity of external lighting required by the system reduces contrast during image display due to glare and reflections. Objects imaged close to the display can be shadowed from ambient lighting sources, causing less accurate measurements. In contrast to transmission-mode light field capture, such as in [29], the design requires the inclusion of an angle-limiting element (either a thin film or other source of vignetting), further reducing the light reaching the optical sensors. The LCD spatial light modulator has limited contrast, which is further reduced at large incidence angles. When capturing data with a tiled-MURA mask, this contrast reduction can be compensated for algorithmically. However to compensate for low contrast when using a pinhole mask, a background image must be captured with the LCD set to fully opaque, and subtracted from the image recorded as the mask is displayed. Capturing the background image further reduces the frame rate possible when using the pinhole mask. Additionally the layered components of the design introduce some artifacts

from reflections and scattering.

4.3 Validation

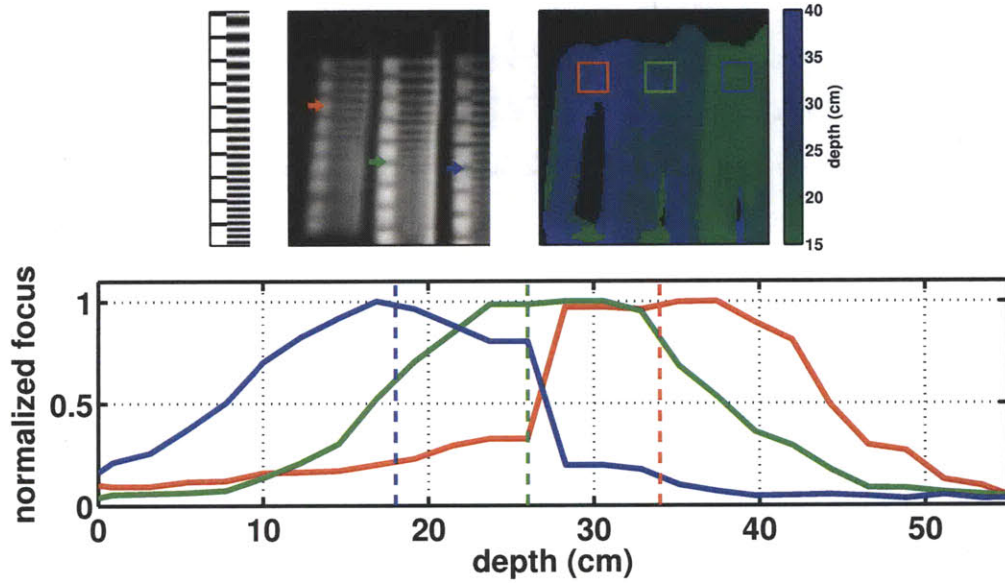


Figure 4-5: Experimental analysis of depth and spatial resolution. (Top, Left) A linear sinusoidal chirp, over the interval $[0.5, 1.5]$ cycles/cm with marks on the left margin indicating 0.1 cycles/cm increments in the instantaneous frequency. Similar to Figure 3-7, three copies of the test pattern were placed parallel to the screen, at distances of $d_o = \{18, 26, 34\}$ cm (from right to left). (Top, Middle) All-in-focus image obtained by refocusing up to 55 cm from the display. As predicted by Equation 3-5, the spatial resolution is approximately 2 cycles/cm near the display, and falls off beyond 30 cm. Note that the colored arrows indicate the spatial cut-off frequencies predicted by Equation 3.4. (Top, Right) The recovered depth map, with near and far objects shaded green and blue, respectively. (Bottom) Focus measure operator response, for the inset regions in the depth map. Note that each peak corresponds to the depth of the corresponding test pattern (true depth shown with dashed lines).

4.3.1 Resolution

A test pattern consisting of a linear sinusoidal chirp, over the interval $[0.5, 1.5]$ cycles/cm, was used to quantify the spatial resolution (in a plane parallel to the display) as a function of distance d_o . In the first experiment, three test patterns were placed at various depths

throughout the interaction volume (see Figure 4-5). Each chart was assessed by plotting the intensity variation from the top to the bottom. The spatial cut-off frequency was measured by locating the point at which fringes lost contrast. As predicted by Equation 3-5, the measured spatial resolution was ≈ 2 cycles/cm near the display; for $d_o < 30$ cm, the pattern lost contrast halfway through (where fringes were spaced at the Nyquist rate of 1 cycle/cm). In a second experiment, the test pattern was moved through a series of depths d_o using a linear translation stage (see Figure 4-6). The experimentally-measured spatial resolution confirms the theoretically-predicted trend in Figure 3-5. In a third experiment, a point light source was translated parallel to the display surface at a fixed separation of 33 cm. The image under a single pinhole (or equivalently a single MURA tile) was used to estimate the lighting angle, confirming a field of view of ≈ 11 degrees. In a fourth experiment, an oscilloscope connected to the GPIO camera trigger recorded an image capture rate of 6 Hz and a display update rate of 20 Hz.

4.3.2 Depth Resolution

The depth resolution was quantified by plotting the focus measure operator response as a function of object distance d_o . For each image pixel this response corresponds to the smoothed gradient magnitude evaluated over the set of images refocused at the corresponding depths. As shown in Figure 4-5, the response is compared at three different image points (each located on a different test pattern). Note that the peak response corresponds closely with the true depth. As described by Nayar and Nakagawa [40], an accurate depth map can be obtained by fitting a parametric model to the response curves. However, for computational efficiency the per-pixel depths are assigned at the per-pixel maximum, leading to more outliers than a parametric model would produce.

4.3.3 Touch and Hover Discrimination

The prototype can discriminate touch events from non-contact gesture motions. Each object in front of the scene is considered to be touching if the median depth is less than 3 cm. In

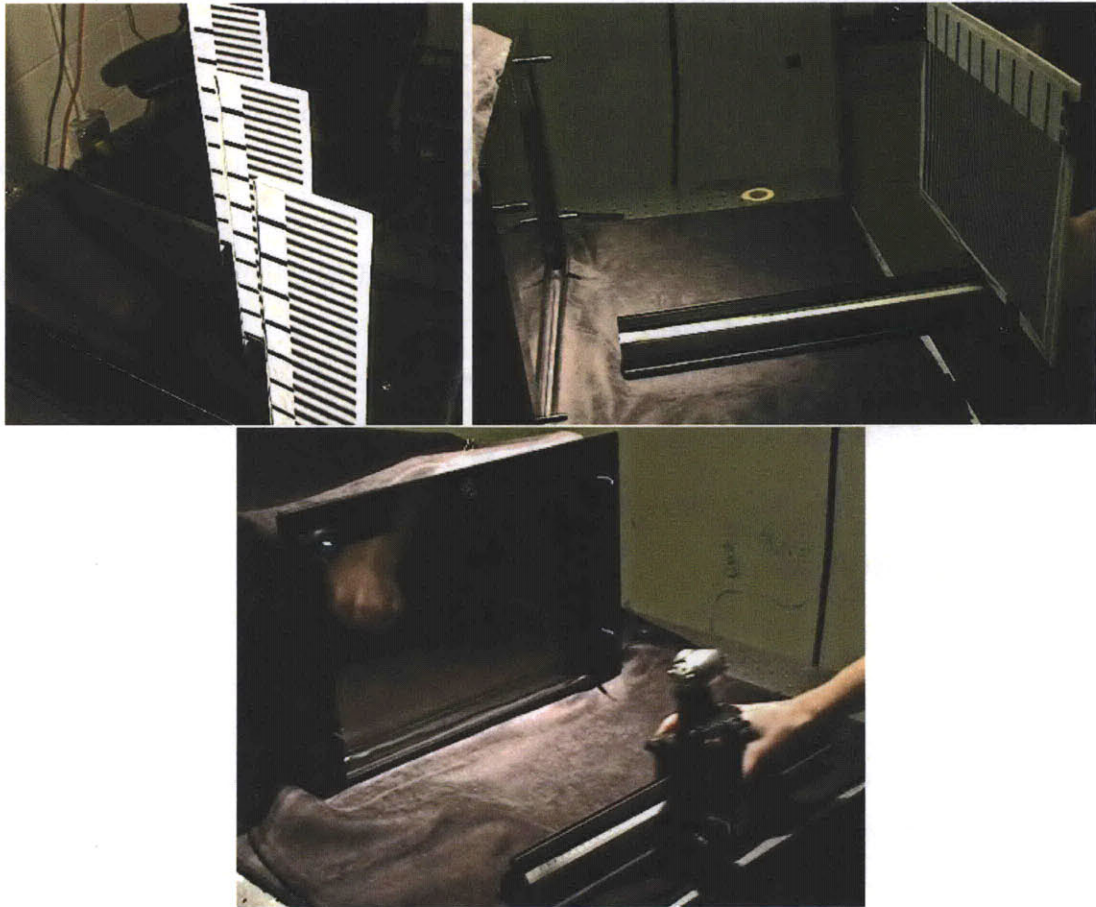


Figure 4-6: The depth and spatial resolution validation was performed as shown here. (Top,Left) Resolution targets were placed at various depths, creating the images shown in Figure 4-5. (Top,Right) a resolution target is translated towards the screen using a translation stage. These tests allow us to validate the spatial resolution obtained as a function of distance from the screen, also shown in Figure 4-5. (Bottom) Angular resolution and field of view is validated by moving a point light along a translation stage.

the following section, I report the experience of several users with the system.

4.4 User Experience

A user experience evaluation was conducted to identify the strengths and weaknesses of the BiDi screen. Four different users were presented with two of the demonstrations seen in the supplementary video. The world viewer demo, and the model viewer demo, described in Section 4.3 were used in all cases, presented in varying order. Three of the four users were able to use both applications successfully, after an initial learning curve. Only one of the users was able to figure out how to control both demos without explanation. Each of the four users experienced two common problems with the depth tracking: moving outside the range of measurement, and moving too rapidly, such that the system lost track of the user's hand. A common problem was also encountered when attempting to use the touch interface. Users typically touch the screen such that the palm of the hand is placed below the point where the fingers touch. The system would register the touch below the point where the fingers were placed due to the palm measurement. One user complained about the flicker rate, one complained about the intensity of the lighting, and two users commented that the world viewer demo was unintuitive. During the course of the model viewer demonstration it was found that the depth threshold used for distinguishing touch had to be adjusted on a per user basis, indicating that a calibration may be required. Despite the problems, three of the four users commented positively about their experience.

The demonstrations used in this evaluation were coded primarily to display the measurement capabilities of the BiDi Screen, rather than to present an intuitive use of those capabilities. The results of this evaluation support the assertion that this method can support a new class of interactive display devices, as the problems encountered did not prevent the users from accomplishing their goals, and are specific to the prototype, such as low frame rate and simplistic interpretation of measurements in the demonstrations.

Chapter 5

Interaction Modes

In this section I describe three proof-of-concept interaction modes supported by the BiDi screen. The first two examples, gestural interaction (multi-touch and hover), and lighting sensitive interaction, are supported by the constructed prototype described in Chapter 4.

5.1 Multi-Touch and Hover

The BiDi screen supports on-screen multi-touch and off-screen hover gestures by providing a real-time depth map, allowing 3D tracking of objects in front of the display. There are few devices that can support both multi-touch interaction and free-space gesture. The multitude of touch devices and gesture devices are described in Chapter 2. Figure 5.1 shows a sample multi-touch to hover interaction.

5.1.1 Demonstrated

As shown in Figure 5.1, a model viewer application is controlled using the 3D position of a user's hand. Several models are presented along the top of the screen. When the user **touches** a model, it is brought to the center of the display. Once selected, the model can



Figure 5-1: Multi-touch demonstration, showing an image being manipulated by two fingers. In this demonstration the image can be scaled and rotated. If a hand is lifted, the image can be tilted up, demonstrating the hover capabilities of the BiDi screen.

be manipulated with **touch-free hover** gestures. The model can be rotated along two axes by moving the hand left to right and up and down. Scaling is controlled by the **distance** between the hand and the screen. Touching the model again puts it away. As shown in Figure 5.1, a world navigator application controls an avatar in a virtual environment. Moving the hand left and right turns, whereas moving the hand up and down changes gaze. Reaching towards or away from the screen affects movement of the avatar. More than one hand can be tracked by the BiDi screen, allowing multi-touch gestures to be implemented as well.

5.1.2 Future Work

The technology behind the BiDi screen is promising for use in mobile devices in that with an area sensor, the screen can be made with approximately the thickness of a typical LCD screen. In the world of mobile computing, where small screens and small keyboards or button pads are a necessity, enabling interaction in a volume in front of the device could be a liberating advancement. On-screen touches easily obscure a large portion of the display on a small device. Hover interaction will solve this problem. Free space gesture offers a greater degree of flexibility for articulating a vocabulary of meaningful gestures than gesture

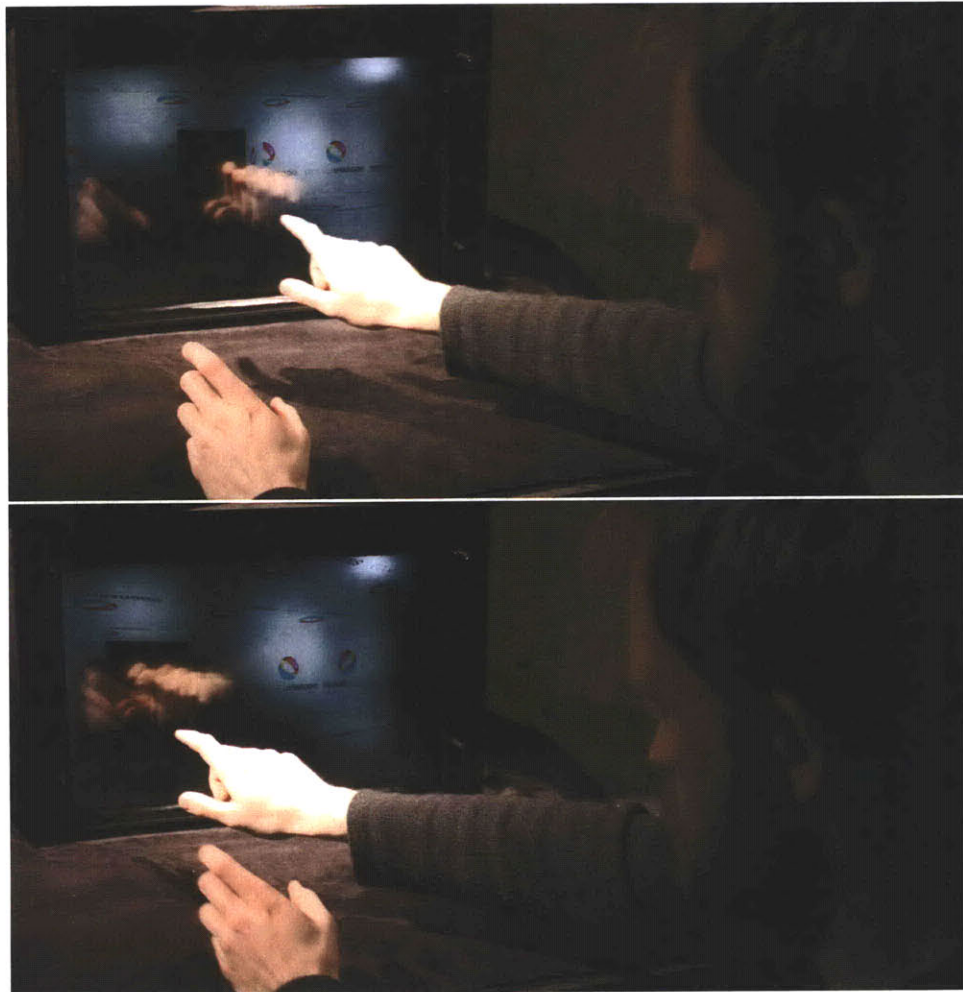


Figure 5-2: A virtual world navigated by tracking a users hand. Moving the hand left, right, up, and down changes the users heading. Reaching towards or away from the screen moves. The figure shows a the world view changing as the hand moves from the center of the screen (Top) to the left (Bottom).

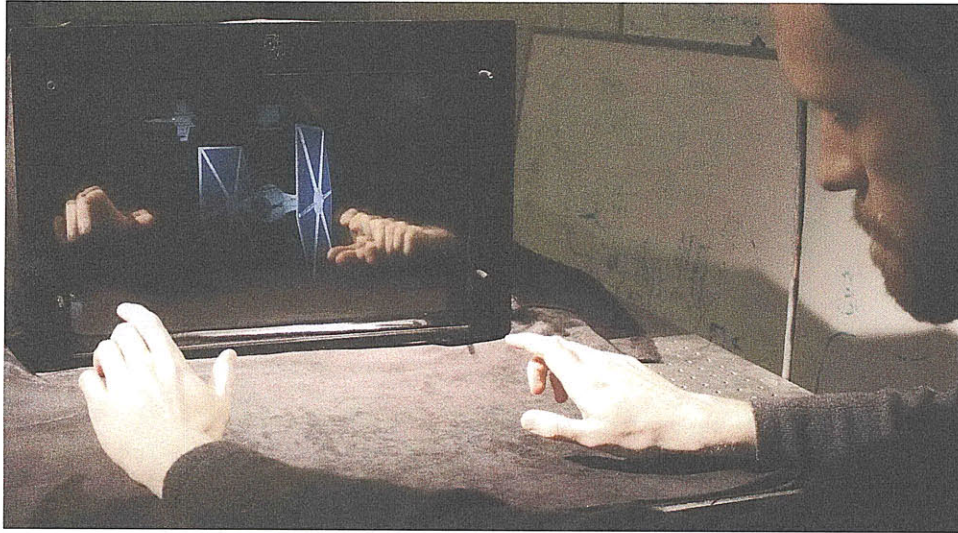


Figure 5-3: (Top) Model viewer demonstration, showing a ship model being manipulated by free space gesture, an example of a mixed on-screen touch and off-screen hover interaction. Virtual models are controlled by the user's hand movement. Touching a model brings it forward from the menu, or puts it away if already selected. Once selected, free space gestures control model rotation and scale.

constrained to the surface of a screen.

The ability of the BiDi screen to create a depth map means that, in addition to tracking the user's hands, it will be possible to track the user's face. This means that a BiDi screen enabled mobile device could take full advantage of its display by providing virtual parallax, whereby the user's shifting his head from side to side would reveal different views of a virtual canvas. This approach treats the screen of the device like a window that the user looks through, rather than as a drawing surface.

5.2 Lighting Sensitive Interaction

Another mode for interacting with the BiDi screen is to intentionally alter the light striking it to alter the rendering of a virtual scene. The BiDi screen creates a window into a virtual world through which users can interact by altering lighting conditions.

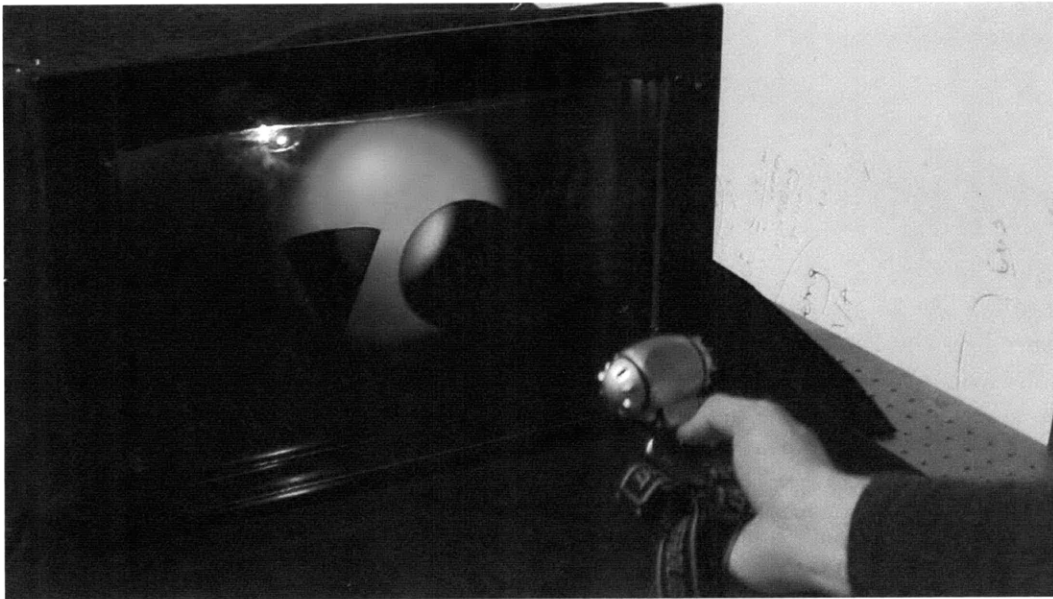


Figure 5-4: A relighting application controlled with a real flashlight. The light field estimates the flashlight position. A similar virtual light is created as if the real flashlight was shining into the virtual world.

5.2.1 Demonstrated

A model lighting application was implemented to allow interactive relighting of virtual scenes (see Figure 5.2). In this interaction mode, the user translates a flashlight in front of the display. For a narrow beam of light, a single pinhole (or MURA tile) may be illuminated. Below this region, a subset of the light sensors will be activated by the beam. The position of the pinhole, in combination with the position of the sensors that are illuminated, can be used to determine the direction along which light entered the screen. A virtual light source is then used to light the simulated scene—as if the viewer was shining light directly into the virtual world.

5.2.2 Future Work

With a sufficiently sensitive screen, it would be possible to transfer real-world lighting conditions into a rendered scene. More exotic applications, such as shining a real projector into a virtual world, are also theoretically possible.



Figure 5-4: A relighting application controlled with a real flashlight. The light field estimates the flashlight position. A similar virtual light is created as if the real flashlight was shining into the virtual world.

5.2.1 Demonstrated

A model lighting application was implemented to allow interactive relighting of virtual scenes (see Figure 5.2). In this interaction mode, the user translates a flashlight in front of the display. For a narrow beam of light, a single pinhole (or MURA tile) may be illuminated. Below this region, a subset of the light sensors will be activated by the beam. The position of the pinhole, in combination with the position of the sensors that are illuminated, can be used to determine the direction along which light entered the screen. A virtual light source is then used to light the simulated scene—as if the viewer was shining light directly into the virtual world.

5.2.2 Future Work

With a sufficiently sensitive screen, it would be possible to transfer real-world lighting conditions into a rendered scene. More exotic applications, such as shining a real projector into a virtual world, are also theoretically possible.

5.3.1 Demonstrated

Although the prototype constructed does not have sufficient spatial resolution to support the video interaction described here, design modifications such as screen and sensor resolution improvements, angle limiter changes and screen-sensor separation changes will enable them. It is possible to increase the spatial resolution of the images captured by the BiDi screen through a simple super-resolution technique. By creating, on an upsampled grid, a refocused spatial image from a captured light-field (see Chapter 3), I show in Figure 5.3 that web-cam quality images may be obtained.

5.3.2 Future Work

One common problem in video chat applications is that the users of the system are unable to make eye contact. Many users find this an unpleasant or unnerving experience. The eye contact, or eye gaze problem, is a result of the fact that video chat applications record the participants with a camera which is spatially separated from the display upon which the participants are imaged. When a participant is looking at the display (at the other participant in the video chat) he or she is not looking at the camera recording him or her. The BiDi screen solves this problem by collocating the display and camera spatially. In order to use this feature, the prototype I constructed would need to be enhanced to perform at higher spatial resolution, and to capture a light-field in color. The current prototype captures only in monochrome.

In a video chat scenario, background segmentation can be performed even in dynamic scenes because scene depth can be extracted from a captured light-field image (See Chapter 3). The video-rate depth map available from the BiDi screen would also make it possible to render virtual objects into the real captured scene. This ability opens up the possibility of novel virtual mirrors built using BiDi screens. A virtual mirror might show the user an image of himself, but render additional objects such as clothing or animated creatures into the scene. The rendered objects could be lit with real-world lighting, as described earlier in this chapter.

Chapter 6

Conclusions

Displays containing arrays of light sensors are already surfacing as research prototypes and poised to enter the market. This work demonstrates a new geometry and technique which add a long list of new features to optically sensitive devices. As this transition occurs, the BiDi screen may inspire the inclusion of some of the features described in this thesis into these devices. In particular, many of the works discussed in Chapter 2 enabled either only multi-touch or pure relighting applications. The contribution of a potentially-thin device for multi-touch and gesture detection is unique. For such interactions, it is not enough to have an embedded array of omnidirectional sensors; instead, by including a sparse array of low-resolution virtual cameras (e.g., through the multi-view orthographic imagery in the design shown here), the increased angular resolution directly facilitates depth and lighting aware interaction.

6.1 Future Work

6.1.1 Hardware

The prototype described in this thesis uses a diffuser and camera pair to simulate an area sensor that will be available in future optically sensitive displays. A camera-diffuser pair is

not the ideal hardware for the BiDi screen, as it creates an unnecessarily large device. In the future it will be interesting to use an area sensor to create a truly thin BiDi screen. There are many options for this sensor, such as thin-film transistors (TFT) as are transparent and used in LCD screens, printed transistors on a diffuse substrate, or traditional CMOS or CCD sensors, to name a few. Note that using TFT sensors is compatible with existing LCD manufacturing techniques.

6.1.2 Variable Masks

One of the key benefits of the LCD-based design presented here is that it transforms a liquid crystal spatial light modulator to allow both image capture and display. Unlike many existing mask-based imaging devices, the system is capable of dynamically updating the mask. Promising directions of future work include reconfiguring the mask based on the properties of the scene (e.g., locally optimizing the spatial vs. angular resolution trade-off), and temporally multiplexing masks to gain higher resolution (See Section 6.1.4 on Scanning).

6.1.3 Video

As higher-resolution video cameras and LCD screens become available, the presented design should scale to provide photographic-quality images – enabling demanding videoconferencing, gaze tracking, and foreground/background matting applications. The prototype simulated a two megapixel sensor and used a one megapixel display. The prototype produced an image with approximately 150x100 resolution (See Figure 5.3). A factor of two improvement would produce web-cam-quality video. Higher frame rates should allow flicker-free viewing and more accurate tracking. In order to achieve higher-resolution imagery for these applications, recent advances [5, 37] in light field super-resolution could be applied to orthographic multi-view imagery captured by my prototype.

6.1.4 Scanning

The use of dynamic lensless imaging systems in the consumer market is another potential direction of future work. A promising direction of future work is to apply the BiDi screen for high-resolution photography of still objects by using translated pinhole arrays; however, such dynamic masks would be difficult to extend to moving scenes. The display could additionally be used in a feedback loop with the capture mode to directly illuminate gesturing body parts or enhance the appearance of nearby objects [10], as currently achieved by SecondLight [26].

6.1.5 Handheld

The space of handheld and other portable devices can derive substantial benefits from incorporating BiDi screens as their displays. The interfaces of small portable devices are constrained by their size. Expanding the interactive region from the area of the device's screen to a volume in front of the screen using gestural interaction will afford users greater input bandwidth to control this class of device. Technologies to allow robust depth discrimination in a form factor suitable for handheld devices have not been available or cost effective until now. Beyond providing mixed reality mirrors and seamless videoconferencing on handheld devices, the ability to track the head will allow more efficient use of limited display area by allowing the device to simulate parallax with a 2D display.

6.2 Synopsis

In this thesis I have described a method for extending a typical LCD screen to support multi-touch and free-space gestural interaction. The LCD accomplishes this by using the liquid crystal layer as a spatial light modulator, coupled with a wide-area sensor. Spatial heterodyning is performed to enable real-time light-field capture. From a captured light-field, scene depth may be extracted, and scene lighting may be used to render virtual scenes. Finally, the approach presented enables a new class of applications that take advantage of

a collocated camera and display. These applications include video chat that enables eye-contact between participants, and mixed reality virtual mirrors.

I have presented a prototype device, capable of supporting real-time gestural and touch interaction with bare hands and light emitting widgets. The prototype substitutes a diffuser and camera combination for the proposed large-area sensor layer. This substitution allows the prototype to be constructed entirely from consumer, off-the-shelf parts.

Appendix A

Optimizing Heterodyne Mask Properties

A.1 Forward

This appendix is taken directly from the BiDi Screen paper submitted by myself, Douglas Lanman, Ramesh Raskar, and Henry Holtzman to SIGGRAPH Asia 2009. I include it here as an appendix because the analysis performed herein was contributed by Lanman, but is important to have on hand in order to fully explain the work done for this thesis.

In this paper we propose a lensless light field capture device composed of a single attenuating mask placed slightly in front of an optical sensor array. As in our prototype, a diffuser and one or more cameras can be substituted for the sensor array and an LCD can be used for the mask. A user will be primarily concerned with maximizing the spatial and angular resolution of the acquired multi-view imagery. In this appendix we describe how to choose the sensor-mask (or diffuser-mask) separation and specific mask patterns to optimize image capture, first for pinhole arrays and then for tiled-broadband codes.

As with other mask-based and lenslet-based light field cameras, the total number of light field samples, given by the product of the spatial and angular samples, can be no greater

than the number of camera pixels [43]. In our system the discretization due to the LCD further limits the mask resolution, restricting the total number of light field samples to be approximately equal to the total number of pixels in the display (i.e., $1680 \times 1050 = 1.764 \times 10^6$ pixels). Thus, as described in Section 4.1, we achieve a spatial resolution of 88×55 samples and an angular resolution of 19×19 samples with a pinhole or MURA tile spacing of $d_p = 4.92$ mm and a mask separation of $d_i = 2.5$ cm. However, by adjusting the spacing and separation, the spatio-angular resolution trade-off can be adjusted.

A.2 Pinhole Array Mask Configuration

As shown in Figure 3-4, each pinhole must be separated by a distance $d_p = 2d_i \tan(\alpha/2)$ if diffraction is negligible (otherwise Equation 3.2 must be used). Thus, the necessary pinhole array separation d_i is given by

$$d_i = \frac{d_p}{2 \tan(\alpha/2)}. \quad (\text{A.1})$$

The field of view α , shown in Figure 3-3, is either determined by the vignetting of each sensor pixel (e.g., that due to the diffuser and camera's field of view in our system) or by a user-selected angle-limiting film. Wider fields of view may be desirable for some applications. However, for a fixed field of view, the user must still choose the mask separation d_i to optimize the effective spatial resolution within the region in front of the display. Thus, Equation 3.4 can be used to maximize $N_{spatial}$ as a function of d_i . In our design we choose a fixed object distance of $d_o = 25$ cm (centered within our working distance $0 \text{ cm} < d_o < 50 \text{ cm}$). As an alternative to Figure 3-5, we can plot the effective spatial resolution as a function of the mask separation d_i (see Figure A-1). Note that the selected distance $d_i = 2.5$ cm is close to the maximum, allowing slightly higher angular resolution (via Equation 3.3) without a significant reduction in spatial resolution.

A.3 Tiled-Broadband Mask Configuration

The placement and design of tiled-broadband masks was described in [29]. However, their design was for a transmission-mode system with a uniform array of LEDs placed a fixed distance in front of the sensor. Our reflection-mode system requires the mask be placed at a different distance from the sensor to allow light field capture. In this section, the notation and derivation mirrors that paper. We describe 2D light fields and 1D sensor arrays, however the extension to 4D light fields and 2D sensor arrays arrives at a similar mask separation d_m .

As shown in Figure A-2, consider the two-plane parameterization [8], where u denotes the horizontal coordinate (along the sensor or diffuser) and s denotes the horizontal position of intersection (in the local frame of u) of an incident ray with a plane that is a fixed distance $d_{ref} = 1$ cm away from, and parallel to, the first plane. A mask separated by d_m from the sensor creates a *shield field* that acts as a volumetric occlusion function $o(u, s) = m(u + (d_m/d_{ref})s)$. Thus, each ray parameterized by coordinates (u, s) is attenuated by the mask's attenuation function $m(\xi)$ evaluated at $\xi = u + (d_m/d_{ref})s$. Taking the 2D Fourier transform, with respect to u and s , we conclude that the mask's shield field spectrum $O(f_u, f_s)$ is given by

$$O(f_u, f_s) = M(f_u)\delta(f_s - (d_m/d_{ref})f_u), \quad (\text{A.2})$$

where $M(f_\xi)$ is the 1D Fourier transform of $m(\xi)$. As described in [51], the optical action of the mask can be modeled by convolving the incident light field spectrum $L_{incident}(f_u, f_s)$

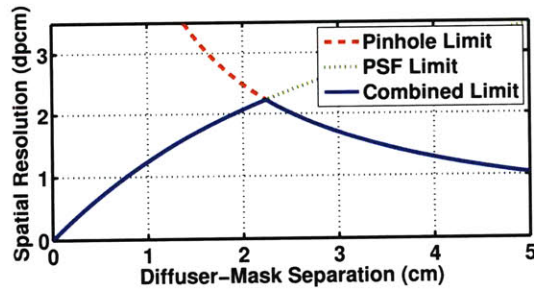


Figure A-1: Effective spatial resolution as a function of diffuser-mask separation d_i for a pinhole array, given by Equation 3.4. System parameters correspond with the prototype in Chapter 4.

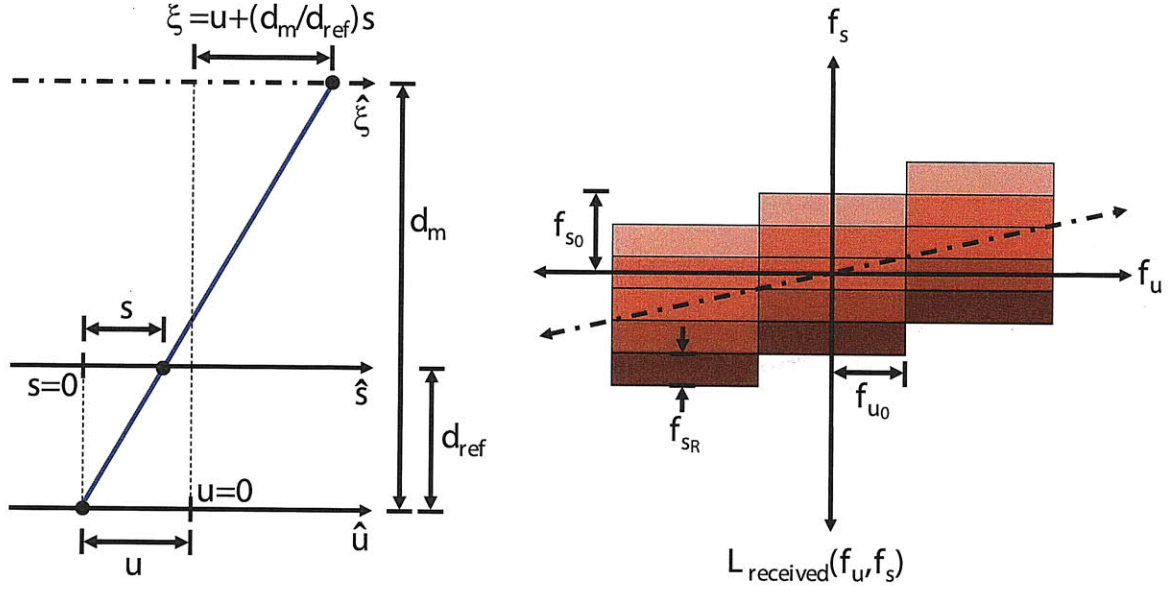


Figure A-2: Geometric derivation of tiled-broadband mask separation. (Left) The two-plane parameterization (u, s) of the optical ray shown in blue. The ray intersects the mask at $\xi = u + (d_m/d_{ref})s$. (Right) The received light field spectrum contains multiple spectral replicas shown in shades of red. The shield field spectrum must lie along the dashed line (i.e., $f_s/f_u = (d_m/d_{ref}) = f_{sR}/(2f_{u0})$).

with the shield field spectrum $O(f_u, f_s)$. As they show, this implies that the mask spectrum must be composed of a uniform series of impulses, with the tiled-MURA mask being one such valid pattern when the tile dimensions are equal to the pinhole spacing d_p .

At this point the mask separation d_m must be determined such that the received image can be decoded to recover the incident light field. Assume that $L_{incident}(f_u, f_s)$ is bandlimited to f_{u0} and f_{s0} , as shown in Figure A-2. Since Equation A.2 implies that the mask spectrum lies along the line $f_s = (d_m/d_{ref})f_u$, then we conclude that

$$d_m = \frac{d_{ref} f_{sR}}{2f_{u0}} = \frac{d_p}{2 \tan(\alpha/2)}, \quad (\text{A.3})$$

where $f_{sR} = 1/(2d_{ref} \tan(\alpha/2))$ and $f_{u0} = 1/(2d_p)$. Note that Equations A.1 and A.3 imply that the pinhole array and tiled-broadband codes are placed the same distance away from the sensor.

Bibliography

- [1] Adiel Abileah, Willem den Boer, Richard T. Tuenge, and Terrance S. Larsson. Integrated optical light sensitive active matrix liquid crystal display. United States Patent 7009663, Assignee: Planar Systems, Inc., March 2006.
- [2] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.
- [3] Hrvoje Benko and Edward W. Ishak. Cross-dimensional gestural interaction techniques for hybrid immersive environments. In *Proc. of IEEE Virtual Reality*, pages 209–216, 327, 2005.
- [4] Hrvoje Benko and Andrew D. Wilson. Depthtouch: Using depth-sensing camera to enable freehand interactions on and above the interactive surface. *Tech. Report MSR-TR-2009-23*, 2009.
- [5] Tom Bishop, Sara Zanetti, and Paolo Favaro. Light field superresolution. In *Proc. IEEE International Conference on Computational Photography*, 2009.
- [6] François Blais. Review of 20 years of range sensor development. *Journal of Electronic Imaging*, 13(1):231–240, 2004.
- [7] Chris J. Brown, Hiromi Kato, Kazuhiro Maeda, and Ben Hadwen. A continuous-grain silicon-system lcd with optical input function. *IEEE Journal Of Solid State Circuits*, 42(12), 2007.
- [8] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proc. of ACM SIGGRAPH*, pages 307–318, 2000.
- [9] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 142–149, 2000.
- [10] Oliver Cossairt, Shree Nayar, and Ravi Ramamoorthi. Light field transfer: global illumination between real and synthetic objects. *ACM Trans. Graph.*, 27(3):1–6, 2008.
- [11] A. Criminisi, J. Shotton, A. Blake, C. Rother, and P. H.S. Torr. Efficient dense stereo and novel-view synthesis for gaze manipulation in one-to-one teleconferencing, 2003.

- [12] Paul Dietz and Darren Leigh. Diamondtouch: a multi-user touch technology. In *Proc. ACM Symposium on User Interface Software and Technology*, pages 219–226, 2001.
- [13] H. Farid. *Range Estimation by Optical Differentiation*. PhD thesis, University of Pennsylvania, 1997.
- [14] E. Fenimore and T. Cannon. Coded aperture imaging with uniformly redundant arrays. *Appl. Optics*, 17(3):337–347, 1978.
- [15] Randima Fernando and Mark J. Kilgard. *The Cg Tutorial: The Definitive Guide to Programmable Real-Time Graphics*. Addison-Wesley Professional, 2003.
- [16] Clifton Forlines and Chia Shen. Dtlens: multi-user tabletop spatial data exploration. In *Proc. ACM Symposium on User Interface Software and Technology*, pages 119–122, 2005.
- [17] Martin Fuchs, Ramesh Raskar, Hans-Peter Seidel, and Hendrik P. A. Lensch. Towards passive 6d reflectance field displays. *ACM Trans. Graph.*, 27(3):1–8, 2008.
- [18] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, New York, NY, USA, 1996. ACM.
- [19] J Y. Han. Low-cost multi-touch sensing through frustrated total internal reflection. *Proc. ACM Symposium on User Interface Software and Technology*, 2005.
- [20] Eugene Hecht. *Optics (4th Edition)*. Addison Wesley, 2001.
- [21] Douglas L. Heirich. System and method for image capture and display utilizing time sharing across a single, two-way optical path. United States Patent 5801758, Assignee: Apple Computer, Inc., Sep. 1998.
- [22] W. Daniel Hillis. A High-Resolution Imaging Touch Sensor. *International Journal of Robotics Research*, 1(2):33–44, 1982.
- [23] S. Hsu, S. Acharya, A. Rafii, and R. New. Performance of a time-of-flight range camera for intelligent vehicle safety applications. *Advanced Microsystems for Automotive Applications*, 2006.
- [24] G. J. Iddan and G. Yahav. Three-dimensional imaging in the studio and elsewhere. *Three-Dimensional Image Capture and Applications IV*, 4298(1):48–55, 2001.
- [25] Shahram Izadi, Steve Hodges, Alex Butler, Alban Rrustemi, and Bill Buxton. Thin-sight: integrated optical multi-touch sensing through thin form-factor displays. In *Proc. Workshop on Emerging Display Technologies*, page 6, 2007.
- [26] Shahram Izadi, Steve Hodges, Stuart Taylor, Dan Rosenfeld, Nicolas Villar, Alex Butler, and Jonathan Westhues. Going beyond the display: a surface technology with an electronically switchable diffuser. In *Proc. ACM Symposium on User Interface Software and Technology*, pages 269–278, 2008.

- [27] Mark J. Kilgard. *OpenGL Programming for the X Window System*. Addison-Wesley Professional, 1996. Describes GLUT.
- [28] Shunsuke Kobayashi, Shigeo Mikoshiba, and Sungkyoo Lim, editors. *LCD Backlights*. Display Technology. Wiley, 2009.
- [29] Douglas Lanman, Ramesh Raskar, Amit Agrawal, and Gabriel Taubin. Shield fields: modeling and capturing 3d occluders. *ACM Trans. Graph.*, 27(5):1–10, 2008.
- [30] SK Lee, William Buxton, and K. C. Smith. A multi-touch three dimensional touch-sensitive tablet. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 21–25, 1985.
- [31] A Levin, R Fergus, F Durand, and W T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.*, 26(3):70, 2007.
- [32] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proc. of ACM SIGGRAPH*, pages 31–42, 1996.
- [33] Marc Levoy and Pat Hanrahan. Light field rendering, 1996.
- [34] Chia-Kai Liang, Tai-Hsu Lin, Bing-Yi Wong, Chi Liu, and Homer H. Chen. Programmable aperture photography: multiplexed light field acquisition. *ACM Trans. Graph.*, 27(3):1–10, 2008.
- [35] G Lippmann. Epreuves reversible donnant la sensation du relief. *Journal of Physics*, 7(4):821–825, 1908.
- [36] David M. Lokhorst and Sathya R Alexander. Pressure sensitive surfaces. United States Patent 7077009, Assignee: Tactex Controls Inc., 2004.
- [37] Andrew Lumsdaine and Todor Georgiev. The focused plenoptic camera. In *Proc. IEEE International Conference on Computational Photography*, 2009.
- [38] Shahzad Malik and Joe Laszlo. Visual touchpad: a two-handed gestural input device. In *Proc. International Conference on Multimodal Interfaces*, pages 289–296, New York, NY, USA, 2004. ACM.
- [39] Nobuyuki Matsushita and Jun Rekimoto. Holowall: designing a finger, hand, body, and object sensitive wall. In *Proc. of ACM Symposium on User Interface Software and Technology*, pages 209–210, 1997.
- [40] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(8):824–831, 1994.
- [41] Shree K. Nayar, Peter N. Belhumeur, and Terry E. Boult. Lighting sensitive display. *ACM Trans. Graph.*, 23(4):963–979, 2004.
- [42] R Ng, Mark Levoy, M Bredif, G Duval, M Horowitz, and P Hanrahan. Light field photography with a hand-held plenoptic camera. *Tech Report, Stanford University*, 2005.

- [43] Ren Ng. Fourier slice photography. In *Prof. of ACM SIGGRAPH*, pages 735–744, 2005.
- [44] Jun Rekimoto. Smartskin: an infrastructure for freehand manipulation on interactive surfaces. In *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, pages 113–120, 2002.
- [45] J. Salvi, J. Pages, and J. Batlle. Pattern codification strategies in structured light systems. In *Pattern Recognition*, volume 37, pages 827–849, April 2004.
- [46] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 519–528, 2006.
- [47] Dave Shreiner, Mason Woo, Jackie Neider, and Tom Davis. *OpenGL Programming Guide: The Official Guide to Learning OpenGL*. Addison-Wesley Professional, sixth edition, 2007. Red Book.
- [48] Michael Uy. United states patent application us 2006/0007222, a1 assignee: Apple computer, inc., integrated sensing display. United States Patent Application US 2006/0007222 A1, Assignee: Apple Computer, Inc., Jan. 2006.
- [49] V. Vaish, G. Garg, E. Talvala, E. Antunez, B. Wilburn, M. Horowitz, and M. Levoy. Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In *Computer Vision and Pattern Recognition - Workshops, 2005*, pages 129–129, 2005.
- [50] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C. L. Zitnick, and Sing Bing Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 2331–2338, 2006.
- [51] A Veeraraghavan, Ramesh Raskar, R Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69, 2007.
- [52] Masahiro Watanabe and Shree K. Nayar. Rational filters for passive depth from defocus. *Int. J. Comput. Vision*, 27(3):203–225, 1998.
- [53] Wayne Westerman and John G. Elias. Multi-touch: A new tactile 2-d gesture interface for human-computer interaction. In *Proc. of Human Factors And Ergonomics Society*, pages 632–636, 2001.
- [54] Bennett Wilburn, Michal Smulski, Kelin Lee, and Mark A. Horowitz. The light field video camera. In *in Media Processors 2002*, pages 29–36, 2002.
- [55] Andrew D. Wilson. Touchlight: an imaging touch screen and display for gesture-based interaction. In *Proc. International Conference on Multimodal Interfaces*, pages 69–76, 2004.

- [56] Shin-Tson Wu and Deng-Ke Yang. *Fundamentals of Liquid Crystal Devices*. Display Technology. Wiley, 2006.
- [57] Jason C. Yang, Matthew Everett, Chris Buehler, and Leonard McMillan. A real-time distributed light field camera. In *Thirteenth Eurographics Workshop on Rendering*, pages 77–86, 2002.
- [58] Cha Zhang and Tsuhan Chen. Light field capturing with lensless cameras. In *Proc. IEEE International Conference on Image Processing*, pages 792–795, 2005.
- [59] A. Zomet and S.K. Nayar. Lensless imaging with a controllable aperture. *Proc. IEEE Computer Vision and Pattern Recognition*, 1:339–346, 2006.
- [60] A. Zomet and S.K. Nayar. Lensless Imaging with a Controllable Aperture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2006.