

BiDi Screen: A Thin, Depth-Sensing LCD for 3D Interaction using Light Fields

Matthew Hirsch¹

Douglas Lanman²

Henry Holtzman¹

Ramesh Raskar¹

¹MIT Media Lab

²Brown University

<http://www.bidiscreen.com>

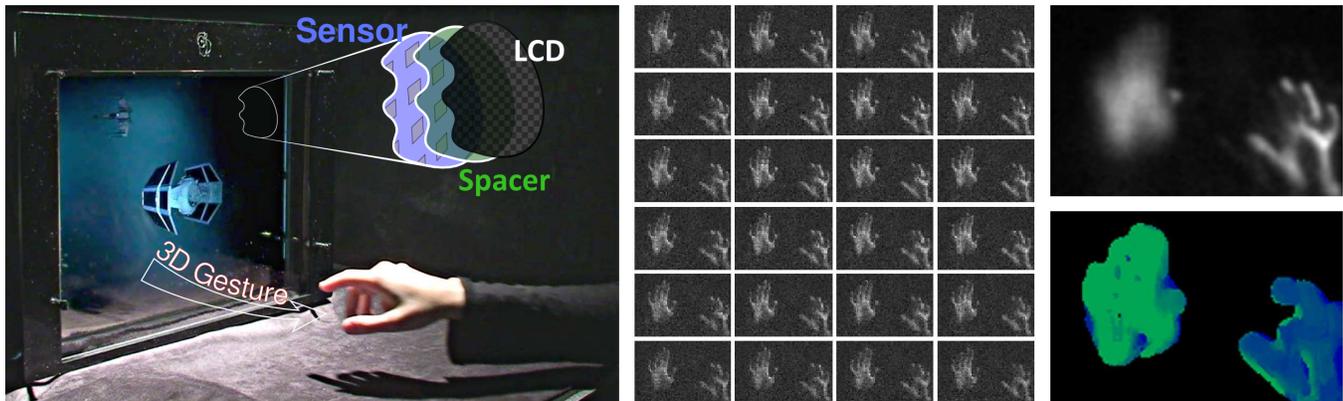


Figure 1: 3D interaction with thin displays. We modify an LCD to allow co-located image capture and display. (Left) Mixed on-screen 2D multi-touch and off-screen 3D interactions. Virtual models are manipulated by the user’s hand movement. Touching a model brings it forward from the menu, or puts it away. Once selected, free-space gestures control model rotation and scale. (Middle) Multi-view 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; 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.

Abstract

We transform an LCD into a display that supports both 2D multi-touch and unencumbered 3D gestures. Our 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 contact. Our key contribution is to exploit the spatial light modulation capability of LCDs to allow lensless imaging without interfering with display functionality. We switch between a display mode showing traditional graphics and a capture mode in which the backlight is disabled and the LCD displays a pinhole array or an equivalent tiled-broadband code. A large-format image sensor is placed slightly behind the liquid crystal layer. Together, the image sensor and LCD form a mask-based light field camera, capturing an array of images equivalent to that produced by a camera array spanning the display surface. The recovered multi-view orthographic imagery is used to passively estimate the depth of scene points. Two motivating applications are described: a hybrid touch plus gesture interaction and a light-gun mode for interacting with external light-emitting widgets. We show a working prototype that simulates the image sensor with a camera and diffuser, allowing interaction up to 50 cm in front of a modified 20.1 inch LCD.

Keywords: LCD, 3D interaction, light field, 3D reconstruction, depth from focus, image-based relighting, lensless imaging

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 [Brown et al. 2007] and Planar Systems, Inc. [Abileah et al. 2006] have demonstrated LCDs with arrays of optical sensors interlaced within the pixel grid. The location of a finger or stylus is determined from the spatial position of occluded sensors that receive less light. For objects pressed directly against such screens, photographic imaging is possible, but objects moved further away quickly become blurred as the light reflecting off any portion of the object is spread across many sensors.

In this paper we describe how to modify LCDs to allow both image capture and display. By using the LCD to display a pinhole array, or an equivalent tiled-broadband code [Lanman et al. 2008], we capture the angle and intensity of light entering a co-located sensor array. By correlating data from multiple views, we image objects (such as fingers) that are located beyond the display’s surface and measure their distance from the display. In our prototype imaging is performed in real-time, enabling the detection of off-screen gestures. When used with a light-emitting implement, our screen determines not only where the implement is aimed, but also the incidence angle of light cast on the display surface.

We propose the *BiDirectional (BiDi) screen*. 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 image, and a capture mode where the backlight is disabled and the light modulator displays an array of pinholes or a tiled-broadband code. Together, the image sensor and LCD form a mask-based light field camera. We have built a working prototype, substituting a diffuser and conventional cameras for the sensor array. We show the BiDi screen in two motivating applications: a hybrid touch plus gesture interaction, and a light-gun mode for interaction using a light-emitting widget.

1.1. Contributions

Thin, Depth-Sensing LCDs: Earlier light-sensing displays focused on achieving touch interfaces. Our design advances the field by supporting both on-screen 2D multi-touch and off-screen, unencumbered 3D gestures. Our key contribution is that the LCD is put to double duty; it alternates between its traditional role in forming the displayed image and a new role in acting as an optical mask. We show that achieving depth- and lighting-aware interactions requires a small displacement between the sensing plane and the display plane. Furthermore, we maximize the display and capture frame rates using optimally light-efficient mask patterns.

Lensless Light Field Capture: We describe a thin, lensless light field camera composed of an optical sensor array and a spatial light modulator. We evaluate the performance of pinhole arrays and tiled-broadband masks for light field capture from primarily reflective, rather than transmissive, scenes. We describe key design issues, including: mask selection, spatio-angular resolution trade-offs, and the critical importance of angle-limiting materials.

Unencumbered 3D Interaction: We show novel interaction scenarios using a BiDi screen to recognize on- and off-screen gestures. We also demonstrate detection of light-emitting widgets, showing novel interactions between displayed images and external lighting.

1.2. Benefits and Limitations

The BiDi screen has several benefits over related techniques for imaging the space in front of a display. Chief among them is the ability to capture multiple orthographic images, with a potentially thin device, without blocking the backlight or portions of the display. Besides enabling lighting direction and depth measurements, these multi-view images support the creation of a true mirror, where the subject gazes into her own eyes, or a videoconferencing application in which the participants have direct eye contact [Rosenthal 1947]. At present, however, the limited resolution of the prototype does not produce imagery competitive with consumer webcams.

The BiDi screen requires separating the light-modulating and light-sensing layers, complicating the display design. In our prototype an additional 2.5 cm was added to the display thickness to allow the placement of the diffuser. In the future a large-format sensor could be accommodated within this distance, however the current prototype uses a pair of cameras placed about 1 m behind the diffuser—significantly increasing the device dimensions. Also, as the LCD is switched between display and capture modes, the proposed design will reduce the native frame rate. Image flicker will result unless the display frame rate remains above the flicker fusion threshold [Izadi et al. 2008]. Lastly, the BiDi screen requires external illumination, either from the room or a light-emitting widget, in order for its capture mode to function. Such external illumination reduces the displayed image contrast. This effect may be mitigated by applying an anti-reflective coating to the surface of the screen.

2. Related Work

2.1. Multi-Touch and 3D Interaction

Sharp and Planar have demonstrated LCDs with integrated optical sensors co-located at each pixel for inexpensive multi-touch interaction. The Frustrated Total Internal Reflection (FTIR) multi-touch wall [Han 2005], TouchLight [Wilson 2004], Microsoft Surface, Oblong Industries g-speak, Visual Touchpad [Malik and Laszlo 2004], and the HoloWall [Matsushita and Rekimoto 1997] use various specialized cameras to detect touch and gestures. In a closely-related work, ThinSight [Izadi et al. 2007] places a compact IR emitter and detector array behind a traditional LCD. In Tactex’s MTC Express [Lokhorst and Alexander 2004] an array of pressure sensors localize where a membrane is depressed. Hillis [1982]

forms a 2D pressure-sensing grid using force-sensitive resistors. A popular approach to multi-touch sensing is through the use of capacitive arrays, described by Lee et al. [1985] and made popular with the iPhone from Apple, Inc., following Fingerworks iGesturePad, both based on the work of Westerman and Elias [Westerman and Elias 2001]. The SmartSkin [Rekimoto 2002], DiamondTouch [Dietz and Leigh 2001], and DTLens [Forlines and Shen 2005] also use capacitive arrays. Benko and Ishak [Benko and Ishak 2005] use a DiamondTouch system and 3D tracked gloves to achieve mixed multi-touch and gesture interaction.

Recent systems image directly through a display surface. Izadi et al. [2008] introduce SecondLight as a 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 [Benko and Wilson 2009] 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 might eliminate the added thickness in current *projection-vision* systems, at the cost of decreased image quality.

2.2. Sensing Depth

A wide variety of passive and active techniques are available to estimate scene depth in real-time. Our prototype records an incident light field [Levoy and Hanrahan 1996] using 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 [2006] describe a system composed of a bare sensor and several attenuating layers, including a single LCD. Liang et al. [2008] use temporally-multiplexed attenuation patterns, also displayed with an LCD, to capture light fields. Zhang and Chen [2005] recover a light field by translating a bare sensor. Levin et al. [2007] and Farid [1997] use coded apertures to estimate intensity and depth from defocused images. Vaish et al. [2006] discuss related methods for depth estimation from light fields. In a closely-related work, Lanman et al. [2008] demonstrate 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. [2007]. We use the tiled-broadband codes from those works to reduce the exposure time in our system. Unlike these systems, our design exploits a mask implemented with a modified LCD panel. In addition, we use reflected light with uncontrolled illumination.

2.3. Lighting-Sensitive Displays

Lighting-sensitive displays have emerged in the market in recent years; most 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. [2004] propose creating lighting-sensitive displays (LSD) by placing optical sensors within the display bezel and altering the rendered imagery to accurately reflect ambient lighting conditions. Cossairt et al. [2008] implement 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. [2008] achieve a passive lighting-sensitive display capable of relighting pre-rendered scenes printed on static masks. Unlike their design, our system works with directional light sources located in front of the display surface and can support relighting of dynamic computer-generated scenes.

3. Bidirectional Screen Design

3.1. Design Goals

It is increasingly common for devices that have the ability to display images to also be able to capture them. In creating the BiDi screen we have four basic design goals:

1. Capture 3D to enable depth- and lighting-aware interaction.
2. Prevent image capture from interfering with image display.
3. Support walk-up interaction (i.e., no implements or markers).
4. Achieve these goals with a portable, thin form factor device.

3.2. Comparison of Design Alternatives

After considering related work and possible image capture options, we believe that the BiDi screen is uniquely positioned to satisfy our design goals. In this section we compare our approach to others.

Capacitive, Resistive, or Acoustic Modalities: A core design decision was to use optical sensing rather than capacitive, resistive, or acoustic modalities. While such technologies are effective for multi-touch, they cannot capture 3D gestures. Some capacitive solutions detect approaching fingers or hands, but cannot accurately determine their distance. Nor do these technologies support lighting-aware interaction. Optical sensing can be achieved in various ways. In many prior works, cameras image the space in front of the display. The result is typically a specially-crafted environment, similar to g-speak, where multiple cameras track special gloves with high contrast markers; or, the display housing is enlarged to accommodate the cameras, as with Microsoft's Surface.

Cameras Behind, To the Side, or In Front of the Display: Another issue is the trade-off between placing a small number of cameras at various points around the display. A camera behind the display interferes with backlighting, casting shadows and causing variations in the display brightness. Han's FTIR sensor, Second-Light, and DepthTouch all avoid this problem by using rear projection onto a diffuser, 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. Cameras placed in the bezel, looking sideways across the display, increase the display thickness and suffer from user self-occlusion. Furthermore, any design incorporating a small number of cameras cannot capture the incident light field, prohibiting certain relighting applications and requiring computationally-intensive multi-view stereo depth estimation, rather than relatively simple depth from focus analysis.

Photodetector Arrays: In contrast, our approach uses an array of photodetectors located behind the LCD (see Figure 2). This configuration will not obscure the backlight and any attenuation will be evenly-distributed. Being behind the display, it does not suffer from user self-occlusion. The detector layer can be extremely thin and optically transparent (using thin film manufacturing), supporting our goal of portability. These are all design attributes we share with multi-touch displays being contemplated by Sharp and Planar. However, we emphasize that our display additionally requires a small gap between the spatial light modulating and light detecting planes. This critical gap allows measuring the angle of incident light, as well as its intensity, and thereby the capture of 3D data.

Camera Arrays: Briefly, we note that a dense camera array placed behind an LCD is equivalent to our approach. However, such tiled cameras must be synchronized and assembled, increasing the engineering complexity compared to the bare sensor in a BiDi screen. In addition, the sensors and lenses required by each camera introduce backlight non-uniformity. Designs incorporating dense camera arrays must confront similar challenges as the BiDi screen, including light absorption (by various LCD layers) and image flicker (due to switching between display and capture frames).

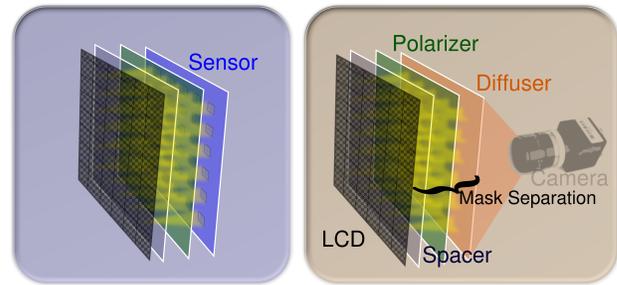


Figure 2: Image capture and display can be achieved by rearranging the optical components within an LCD. A liquid crystal spatial light modulator is used to display a mask (either a pinhole array or equivalent tiled-broadband code). A large-format sensor, placed behind the spatial light modulator, measures the angle of incident light, as well as its intensity. (Left) The modulated light is captured on a sensor array for decoding. (Right) With no large-area sensor available, a camera images a diffuser to simulate the sensor array. In both cases, LEDs restore the backlight function.

4. Designing a Thin, Depth-Sensing LCD

The preferred formulation of the BiDi screen would contain an optically-transparent thin film sensor array embedded within the backlight. In this section we discuss an alternate implementation that substitutes a diffuser and camera array for the sensor array, which is commercially unavailable today. While sub-optimal, many of the practical benefits and limits of a BiDi screen can be explored with our low-cost prototype.

4.1. Overview of LCD Components

We briefly review the functionality of key components included in modern LCDs to provide context for the modifications we describe.

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 rear reflecting surface covering the light guide, a diffuser, and several brightness enhancing films. The overall function of these layers is to condition the light produced by the CCFL such that it is spatially uniform, collimated, and polarized along a single axis before reaching the spatial light modulator. 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. 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.

4.2. Hardware Design

As shown in Figure 2, our BiDi screen is formed by repurposing typical LCD components such that image capture is achieved without hindering display functionality. We begin by excluding 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 [Lanman et al. 2008], we then create a large-aperture, multi-view image capture device by using the spatial light modulator to display a pinhole array or tiled-broadband mask. Our key insight is that, for simultaneous image capture and display, 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 is restored by including an additional array of LEDs behind the diffuser.

We note that an angle-limiting material or other source of vignetting is critical to achieve 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, the proposed image capture scheme assumes light varies only over a limited range of angles—although this range can be arbitrarily large. An angle-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 our design we place the cameras 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.

4.3. Optical Design with Pinhole Arrays

Our design goals require sufficient image resolution to estimate the 3D position of points located in front of the screen, as well as the variation in position and angle of incident illumination. As described by [Veeraraghavan et al. 2007], 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 spatial and angular resolution will be determined by the point spread function (PSF). In this section we optimize the BiDi screen for both on-screen and off-screen interaction under these constraints for the case of a pinhole array mask. In Section 4.4 we extend this analysis to tiled-broadband masks.

Multi-View Orthographic Imagery: As shown in Figure 3, 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 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. [2007] and Abileah et al. [2006].

Off-screen Interaction: To allow hands-free 3D interaction, additional angular views are necessary. First, to estimate the depth of scene points, angular diversity is needed to provide a sufficient

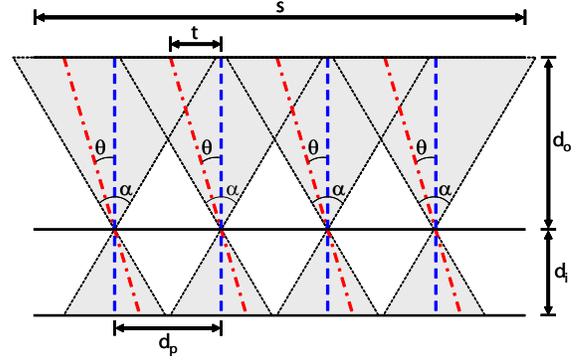


Figure 3: Multi-view orthographic imagery from pinhole arrays. A uniform array of pinhole images (each field of view shaded gray) is resampled to produce a set of orthographic images, each with a different viewing angle θ with respect to the surface normal of the display. The set of optical rays perpendicular 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.

baseline for reconstruction. Second, to facilitate interactions with off-screen light-emitting widgets, the imagery must sample a wide range of incident lighting directions. We conclude that 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 our 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 a single pinhole camera shown in Figure 4, 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, is modeled as

$$b(d_i, d_o, a, \lambda) = \frac{2.44\lambda d_i}{a} + \frac{a(d_o + d_i)}{d_o}. \quad (1)$$

Note that the first and second terms correspond to the approximate blur due to diffraction and the geometric projection of the pinhole aperture onto the sensor plane, respectively [Hecht 2001]. If we now assume that each pinhole camera has a limited field of view, given by α , then the minimum pinhole spacing d_p is

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

Note that a smaller spacing would cause neighboring pinhole images to overlap. As previously described, a limited field of view could be due to vignetting or achieved by the inclusion of an angle-limiting film. Since, in our design, the number of orthographic views $N_{angular}$ is determined by the resolution of each pinhole image, we conclude that the angular resolution of our 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.

$$N_{angular}(d_i, d_o, a, \lambda, \alpha) = \frac{d_p(d_i, d_o, a, \lambda, \alpha)}{b(d_i, d_o, a, \lambda)}. \quad (3)$$

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). Note that a limited field of view is necessary to prevent overlapping of neighboring images. As described in Section 4.5, we use a depth from focus method to estimate the separation of objects from the display surface. As a result, the system components

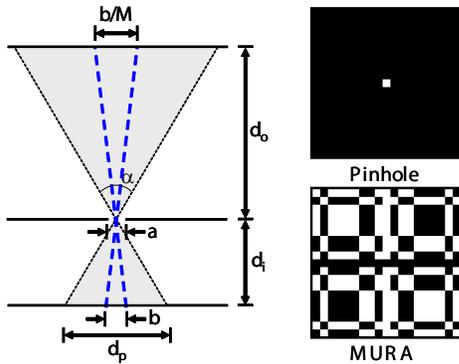


Figure 4: Design of a pinhole camera. (Left) The PSF width b is given by Equation 1 as a function of sensor-pinhole separation d_i , object distance d_o , and the aperture a . The PSF width is magnified by $M = d_i/d_o$ in the plane at d_o . (Right, Top) A single pinhole comprises an opaque set of 19×19 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. [2008] and earlier by Fenimore et al. [1978; 1989], this pattern yields an equivalent image as a pinhole after decoding.

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 is by

$$N_{spatial}(d_i, d_o, a, \lambda, \alpha; d_p, b) = \min\left(\frac{s}{d_p}, \frac{d_i s}{d_o b}\right), \quad (4)$$

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 our system is orthographic, we assume the object plane at d_o is also of width s .

As shown in Figure 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. Note that, in Figure 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 our prototype, the pinhole separation d_p is held constant (as opposed to the variable spacing given in Equation 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.

Optimizing the Sensor-Mask Separation: As with other light field cameras, the total number of samples (given by the product of the spatial and angular resolutions) cannot exceed the number of sensor pixels. Practical considerations, such as LCD discretization, may further limit the mask resolution (see Section 6.1) and restrict the total number of light field samples to be equal to the total number of pixels in the display. However, by adjusting the spacing of pinholes and the sensor-mask (or diffuser-mask) separation, the spatio-angular resolution trade-off can be adjusted.

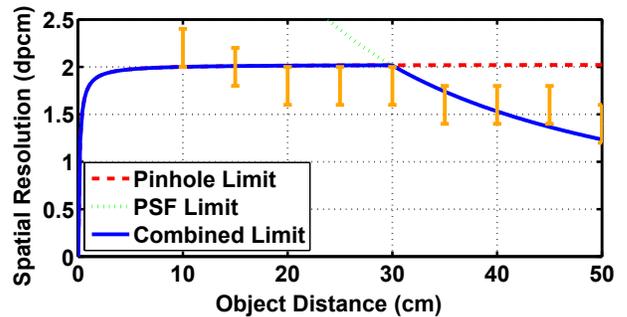


Figure 5: Effective spatial resolution as a function of distance d_o from the display. The effective spatial resolution in a plane at d_o is evaluated using Equation 4. System parameters correspond with the prototype. Orange error bars denote the experimentally-estimated spatial resolution described in Section 6.3. Note that, using either dynamically-shifted masks or higher-quality components, the spatial resolution could significantly increase near the display (approaching the higher limit imposed by the optical PSF).

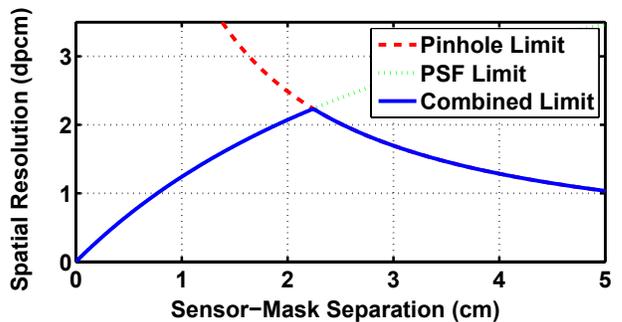


Figure 6: Effective spatial resolution as a function of sensor-mask (or diffuser-mask) separation d_i , as given by Equation 4. System parameters correspond with the prototype in Section 6.1.

We optimize the sensor-mask separation d_i to maximize the effective spatial resolution for objects located in the interaction volume (i.e., within 50 cm of the display). Equation 4 is used to maximize $N_{spatial}$ as a function of d_i . We assume an average object distance of $d_o = 25$ cm. As an alternative to Figure 5, we plot the effective spatial resolution as a function of the mask separation d_i (see Figure 6). Note that the selected distance $d_i = 2.5$ cm is close to the maximum, allowing slightly higher angular resolution (via Equation 3) without a significant reduction in spatial resolution.

4.4. Optical Design with Tiled-Broadband Masks

The primary limitation of a pinhole array is severe attenuation of light. For example, in our 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. 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, we use the generalized tiled-broadband masks described by Lanman et al. [2008]. Specifically, we use a tiled-MURA code, as shown in Figure 4. 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).

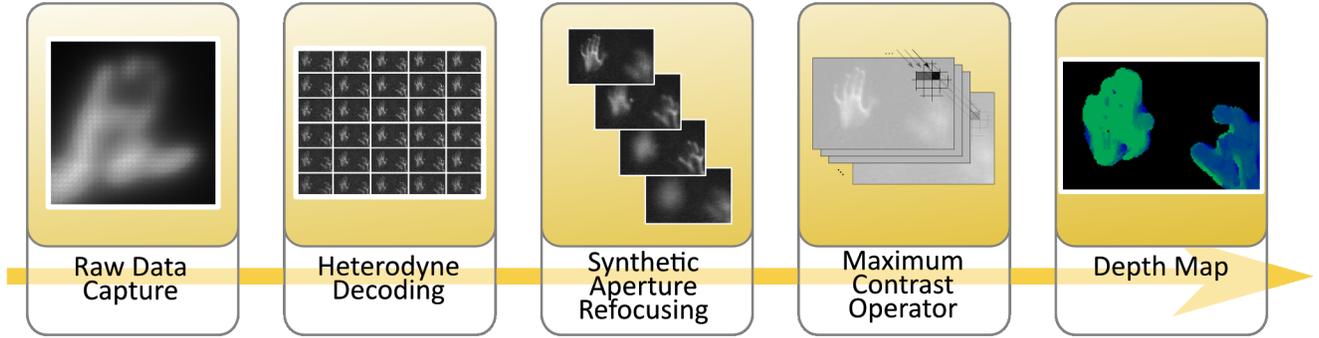


Figure 7: Depth estimation using a BiDi screen. The image captured on the sensor (or diffuser), as modulated by the mask pattern displayed by the LCD, is decoded to recover the incident light field [Veeraraghavan et al. 2007]. Afterwards, synthetic aperture refocusing [Ng 2005] generates a focal stack, from which the depth map is estimated by applying a maximum contrast operator [Watanabe and Nayar 1998].

The heterodyne decoding method of Veeraraghavan et al. [2007] is used to interpret the diffuser-plane image, yielding orthographic multi-view imagery equivalent to a pinhole array mask. The decoding algorithm, however, does require additional computation and introduces noise. We note that the spatio-angular resolution trade-off for such tiled-broadband codes is similar to that described in the previous section for pinhole arrays—yielding a multi-view orthographic image array with similar spatial and angular sampling rates. Furthermore, as derived in Appendix A (included with the supplementary material), a tiled-broadband mask is placed the same distance away from the sensor as an equivalent pinhole array.

4.5. Multi-View Processing

A wide variety of methods exist to estimate depth from multi-view imagery (see Section 2). As shown in Figure 7, we employ a depth from focus method inspired by [Nayar and Nakagawa 1994]. 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, with the assumption that a patch will appear with greatest contrast when the camera is focused at the depth of the patch. In our implementation a simple smoothed gradient magnitude focus measure is used. A coarse depth map is 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, our approach can easily be evaluated in real-time on commodity hardware (see Section 6).

In order to obtain the set of refocused images (i.e., the focal stack), we apply methods from synthetic aperture photography [Vaish et al. 2006]. As shown in Figure 3, when considering the intersection of the optical rays with a plane at distance d_o , each orthographic view,

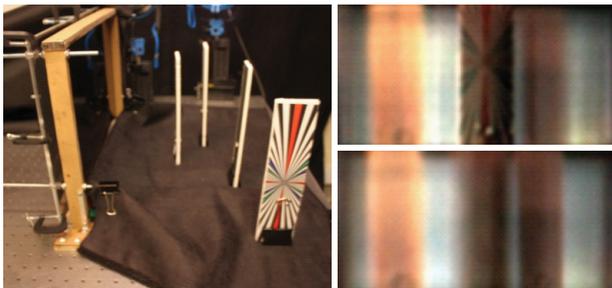


Figure 8: Synthetic aperture refocusing with orthographic imagery. (Left) A scene composed of three textured cards. Note that the center card has a similar radial texture as shown on the card facing the camera. (Right) Refocused images at a distance d_o of 10 cm and 15 cm, shown on the top and bottom, respectively.

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

$$t(d_o, \theta) = d_o \tan(\theta). \quad (5)$$

In order to synthetically focus at a distance d_o , we follow the computationally-efficient approach of Ng [2005]; 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. Refocusing results are shown in Figure 8.

5. Interaction Modes

In this section we describe three proof-of-concept interaction modes supported by the BiDi screen. Examples of user experiences are included in the supplementary video.

5.1. Multi-Touch and 3D Interaction

The BiDi screen supports on-screen multi-touch and off-screen gestures by providing a real-time depth map, allowing 3D tracking of objects in front of the display. As shown in Figure 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 is 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 9, 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. As shown in the supplementary material, more than one hand can be tracked, allowing multi-handed gestures as well.

5.2. Lighting-Sensitive Interaction

Another interaction mode involves altering the light striking the screen. A **model lighting** application allows interactive relighting of virtual scenes (see Figure 9). In this interaction mode, the user translates a flashlight in front of the display. For a narrow beam, a single pinhole (or MURA tile) is illuminated. Below this region, a subset of light sensors is activated. The position of the pinhole, in combination with the position of the illuminated sensors, determines the direction along which light entered the screen. A similar light source is then created to illuminate the simulated scene—as if the viewer was shining light directly into the virtual world.

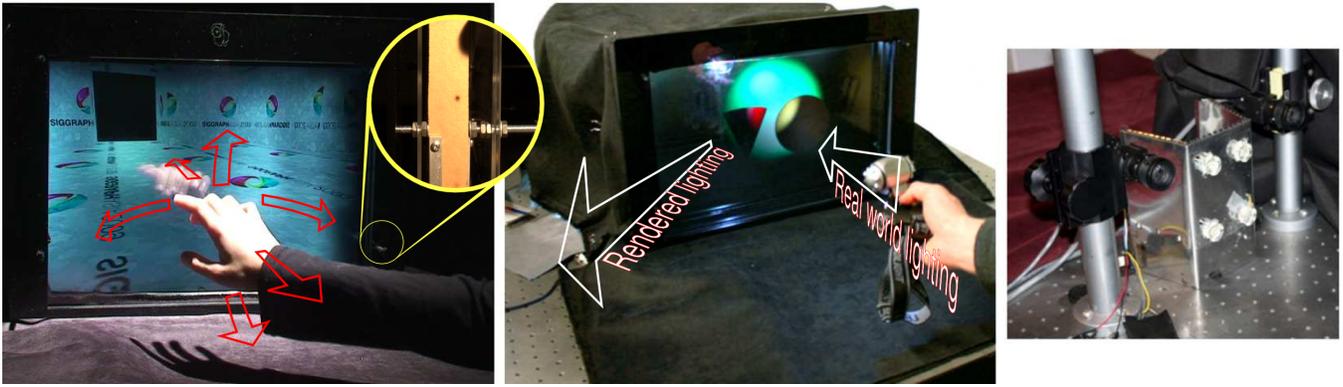


Figure 9: Additional interaction modes. (Left) A virtual world navigated by tracking a user’s hand. Moving the hand left, right, up, and down changes the avatar’s heading. Reaching towards or away from the screen moves. The layers of the prototype are shown in the circled inset, including from left to right: the decorative cover, LCD (in a wooden frame), and diffuser. (Middle) A relighting application controlled with a real flashlight. The flashlight is tracked using the captured light field. A similar virtual light is created, as if the real flashlight was shining into the virtual world. (Right) A pair of cameras and multiple LEDs placed 1 m behind the diffuser (shown in profile in Figure 8).

6. Performance

6.1. Implementation

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 LCDs typically combine the front polarizer with a diffusing layer, as was done with our screen. A diffuser in the plane of our spatial light modulator would interfere with image capture. 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. The CCFL/light guide/reflector backlight was replaced with 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 image capture.

A pair of Point Grey Flea2 cameras was placed 1 m behind the diffuser, each imaging half of the diffuser while recording a 1280×960 16-bit grayscale image at 7 fps (satisfying the Nyquist criterion for the 1680×1050 LCD). For interaction sessions, the cameras were operated in 8-bit grayscale mode. The shutters were triggered from the parallel port to synchronize image capture with LCD refresh and LED strobing. Image capture and display were 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 CPU-based refocusing, depth estimation, and lighting direction estimation pipeline processed raw imagery at up to 7.5 fps.

External lighting was provided by overhead halogen lamps when the tiled-MURA pattern was used. Pinhole masks required an additional halogen 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 8).

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 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 4.3, the acquired light field has 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 overlap between the fields of view of each camera. While limited, the field of view and spatial resolution were sufficient for refocusing and depth estimation (see Figures 8 and 10).

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

6.2. Limitations

The proposed system is constrained to operate within the limits of consumer off-the-shelf technology, which places a lower limit on the pixel sizes in the LCD and the sensor. In turn, these components limit the maximum angular and spatial resolution, as described in Section 6.1. Experimentally, we did not observe that the diffuser PSF limited the effective system resolution. However, the diffraction term in Equation 1 was a significant factor. The $256 \mu\text{m}$ pixels in our display, each color sub-pixel being a third of this width, resulted in a PSF width of about $400 \mu\text{m}$ in the diffuser plane.

Our prototype captures 19×19 orthographic views, each 73×55 pixels. Consumer devices typically provide high spatial or high temporal resolution, but not both simultaneously. The BiDi screen is optimized for real-time interaction, rather than high-resolution photography. The prototype uses a pair of synchronized video cameras and a diffuser to simulate the performance of an embedded optical sensor array. The frame rate is limited to 7.5 fps by the performance of current video cameras and the maximum transfer rate allowed by the 1394b FireWire bus. While the spatial resolution of the depth map is limited, it proves sufficient for tracking individual hands, both in contact and removed from the display surface. Furthermore, individual fingers are resolved in the refocused imagery (see Figure 1), indicating that more sophisticated processing could allow higher-fidelity touch and gesture recognition.

External lighting is required to provide sufficient illumination during image capture. Efficient mask patterns, such as tiled-MURA, allow lighting to be dimmed. However, any external lighting will reduce display contrast due to glare and reflections. Inclusion of an anti-reflection coating may mitigate this effect. Objects close to the display can be occluded from ambient lighting sources, reducing tracking accuracy. In contrast to transmission-mode light field capture, such as in [Lanman et al. 2008], our design requires the inclusion of an angle-limiting element, 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 displaying a tiled-MURA mask, this contrast reduction can be compensated for algorithmically. However, to compensate for low contrast when using a pinhole array, a background image also must be recorded. Capturing the background image further reduces the frame rate when using a pinhole array. An additional lighting pitfall is caused by the layered components of our design, which may introduce artifacts from reflections and scattering.

6.3. Validation

Spatial/Angular/Temporal Resolution: A chart containing 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 a first experiment, three charts were placed at various depths throughout the interaction volume (see Figure 10). 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 position at which fringes lost contrast. As predicted by Equation 4, the 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, a chart was moved through a series of depths d_o using a linear translation stage (for details, see the supplementary material). The experimentally-measured spatial resolution confirms the theoretically-predicted trend in Figure 5. In a third experiment, an LED was translated parallel to the display 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 a capture rate of 6 Hz and a display refresh rate of 20 Hz.

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 10, the response is compared at three different image points (each located on a different chart). Note that the peak response corresponds closely with the true depth. As described by Nayar and Nakagawa [1994], an accurate depth map can be obtained by fitting a parametric model to the response curves. However, for computational efficiency, we assign a quantized depth corresponding to the per-pixel maximum response—leading to more outliers than with a parametric model.

Touch vs. Hover Discrimination: As shown in the supplementary video, the prototype can discriminate touch events from non-contact gesture motions. Each object in front of the screen is considered to be touching if the median depth is less than 3 cm.

Light Efficiency: A Canon EOS Digital Rebel XSi camera was used as a light meter to quantify attenuation for patterns displayed by the LCD. The camera was placed behind the diffuser and percent transmission was measured with respect to the LCD in the fully transparent state. An opaque “black” screen resulted in 5.5% transmission, whereas the tiled-MURA pattern yielded 50% transmission—corresponding with theory. Transmission for

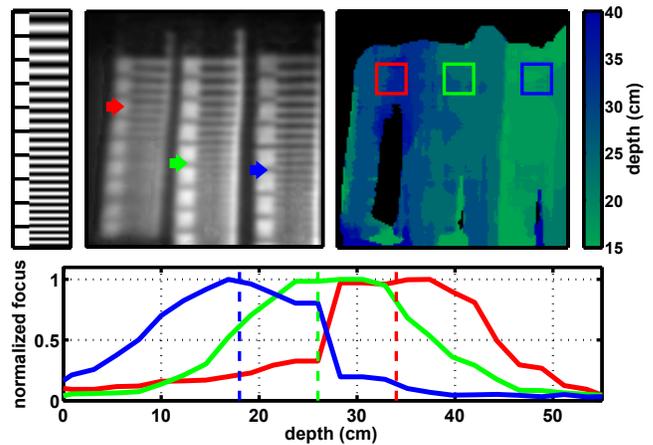


Figure 10: Experimental analysis of depth and spatial resolution. (Top, Left) A linear sinusoidal chirp, over the interval $[0.5, 1.5]$ cycles/cm with marks on the left margin indicating 0.1 cycles/cm increments in the instantaneous frequency. Similar to Figure 8, three copies of the test pattern were 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 4, 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 4. (Top, Left) 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).

the pinhole array fell below the light meter’s sensitivity to distinguish from the opaque state. This underscores the necessity of background subtraction and the light-efficient tiled-MURA pattern. We note, however, that the LCD and diffuser only achieve 3.8% transmission, as compared to photographing a lightbox directly.

7. Discussion and Future Directions

One of the key benefits of our LCD-based design is that it transforms a liquid crystal spatial light modulator to allow both image capture and display. Unlike many existing mask-based imaging devices, our 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). As higher-resolution video cameras and LCD screens become available, our design should scale to provide photographic-quality images — enabling demanding videoconferencing, gaze tracking, and foreground/background matting applications. Higher frame rates should allow flicker-free viewing and more accurate tracking. In order to achieve higher-resolution imagery for these applications, recent advances in light field superresolution [Bishop et al. 2009; Lumsdaine and Georgiev 2009] could be applied to our orthographic multi-view imagery.

The use of dynamic lensless imaging systems in the consumer market is another potential direction of future work. A promising direction 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 ability to track multiple points in free-space could allow identification and response to multiple users, although higher-resolution imagery would be required than currently produced by the prototype. Finally, the display could be used in a feedback loop with the capture mode to directly illuminate gesturing body parts or enhance

the appearance of nearby objects [Cossairt et al. 2008], as currently achieved by SecondLight [Izadi et al. 2008].

8. Conclusion

Light-sensing displays are emerging as research prototypes and are poised to enter the market. As this transition occurs we hope to inspire the inclusion of some BiDi screen features in these devices. Many of the early prototypes discussed in Section 2 enabled either only multi-touch or pure relighting applications. We believe our contribution of a potentially-thin device for multi-touch and 3D interaction is unique. For such interactions, it is not enough to have an embedded array of omnidirectional sensors; instead, by including an array of low-resolution cameras (e.g., through multi-view orthographic imagery in our design), the increased angular resolution directly facilitates unencumbered 3D interaction with thin displays.

Acknowledgements

We thank the reviewers for insightful feedback, the Camera Culture and Information Ecology groups for support, and Gabriel Taubin for useful discussions. Support is provided for Douglas Lanman by the National Science Foundation under Grant CCF-0729126 and for Ramesh Raskar by an Alfred P. Sloan Research Fellowship.

References

- ABILEAH, A., DEN BOER, W., TUENGE, R. T., AND LARSSON, T. S., 2006. Integrated optical light sensitive active matrix liquid crystal display. United States Patent 7,009,663.
- BENKO, H., AND ISHAK, E. W. 2005. Cross-dimensional gestural interaction techniques for hybrid immersive environments. In *IEEE VR*, 209–216, 327.
- BENKO, H., AND WILSON, A. D. 2009. DepthTouch: Using depth-sensing camera to enable freehand interactions on and above the interactive surface. *Tech. Report MSR-TR-2009-23*.
- BISHOP, T., ZANETTI, S., AND FAVARO, P. 2009. Light field superresolution. In *IEEE ICCP*.
- BROWN, C. J., KATO, H., MAEDA, K., AND HADWEN, B. 2007. A continuous-grain silicon-system LCD with optical input function. *IEEE J. of Solid-State Circuits* 42, 12.
- COSSAIRT, O., NAYAR, S., AND RAMAMOORTHI, R. 2008. Light field transfer: Global illumination between real and synthetic objects. *ACM Trans. Graph.* 27, 3.
- DIETZ, P., AND LEIGH, D. 2001. DiamondTouch: A multi-user touch technology. In *ACM UIST*, 219–226.
- FARID, H. 1997. *Range Estimation by Optical Differentiation*. PhD thesis, University of Pennsylvania.
- FENIMORE, E. E., AND CANNON, T. M. 1978. Coded aperture imaging with uniformly redundant arrays. *Appl. Optics* 17, 3, 337–347.
- FORLINES, C., AND SHEN, C. 2005. DTLens: Multi-user tabletop spatial data exploration. In *ACM UIST*, 119–122.
- FUCHS, M., RASKAR, R., SEIDEL, H.-P., AND LENSCH, H. P. A. 2008. Towards passive 6D reflectance field displays. *ACM Trans. Graph.* 27, 3.
- GOTTESMAN, S. R., AND FENIMORE, E. E. 1989. New family of binary arrays for coded aperture imaging. *Appl. Optics* 28, 20, 4344–4352.
- HAN, J. Y. 2005. Low-cost multi-touch sensing through frustrated total internal reflection. *ACM UIST*, 115–118.
- HECHT, E. 2001. *Optics (4th Edition)*. Addison Wesley.
- HILLIS, W. D. 1982. A high-resolution imaging touch sensor. *Int'l. J. of Robotics Research* 1, 2, 33–44.
- IZADI, S., HODGES, S., BUTLER, A., RRUSTEMI, A., AND BUXTON, B. 2007. ThinSight: Integrated optical multi-touch sensing through thin form-factor displays. In *Workshop on Emerging Display Technologies*.
- IZADI, S., HODGES, S., TAYLOR, S., ROSENFELD, D., VILLAR, N., BUTLER, A., AND WESTHUES, J. 2008. Going beyond the display: A surface technology with an electronically switchable diffuser. In *ACM UIST*, 269–278.
- LANMAN, D., RASKAR, R., AGRAWAL, A., AND TAUBIN, G. 2008. Shield fields: Modeling and capturing 3D occluders. *ACM Trans. Graph.* 27, 5.
- LEE, S., BUXTON, W., AND SMITH, K. C. 1985. A multi-touch three dimensional touch-sensitive tablet. In *ACM SIGCHI*, 21–25.
- LEVIN, A., FERGUS, R., DURAND, F., AND FREEMAN, W. T. 2007. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.* 26, 3, 70.
- LEVOY, M., AND HANRAHAN, P. 1996. Light field rendering. In *ACM SIGGRAPH*, 31–42.
- LIANG, C.-K., LIN, T.-H., WONG, B.-Y., LIU, C., AND CHEN, H. H. 2008. Programmable aperture photography: Multiplexed light field acquisition. *ACM Trans. Graph.* 27, 3.
- LOKHORST, D. M., AND ALEXANDER, S. R., 2004. Pressure sensitive surfaces. United States Patent 7,077,009.
- LUMSDAINE, A., AND GEORGIEV, T. 2009. The focused plenoptic camera. In *IEEE ICCP*.
- MALIK, S., AND LASZLO, J. 2004. Visual touchpad: A two-handed gestural input device. In *Int'l. Conf. on Multimodal Interaction*, 289–296.
- MATSUSHITA, N., AND REKIMOTO, J. 1997. HoloWall: Designing a finger, hand, body, and object sensitive wall. In *ACM UIST*, 209–210.
- NAYAR, S. K., AND NAKAGAWA, Y. 1994. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.* 16, 8, 824–831.
- NAYAR, S. K., BELHUMEUR, P. N., AND BOULT, T. E. 2004. Lighting sensitive display. *ACM Trans. Graph.* 23, 4, 963–979.
- NG, R. 2005. Fourier slice photography. In *ACM SIGGRAPH*, 735–744.
- REKIMOTO, J. 2002. SmartSkin: An infrastructure for freehand manipulation on interactive surfaces. In *ACM SIGCHI*, 113–120.
- ROSENTHAL, A. H., 1947. Two-way television communication unit. United States Patent 2,420,198, May.
- VAISH, V., LEVOY, M., SZELISKI, R., ZITNICK, C. L., AND KANG, S. B. 2006. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *IEEE CVPR*, 2331–2338.
- VEERARAGHAVAN, A., RASKAR, R., AGRAWAL, R., MOHAN, A., AND TUMBLIN, J. 2007. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.* 26, 3, 69.
- WATANABE, M., AND NAYAR, S. K. 1998. Rational filters for passive depth from defocus. *Int. J. Comput. Vision* 27, 3.
- WESTERMAN, W., AND ELIAS, J. G. 2001. Multi-Touch: A new tactile 2-D gesture interface for human-computer interaction. In *Human Factors And Ergonomics Society*, 632–636.
- WILSON, A. D. 2004. TouchLight: An imaging touch screen and display for gesture-based interaction. In *Int'l. Conf. on Multimodal Interfaces*, 69–76.
- ZHANG, C., AND CHEN, T. 2005. Light field capturing with lensless cameras. In *IEEE ICIP*, 792–795.
- ZOMET, A., AND NAYAR, S. 2006. Lensless imaging with a controllable aperture. *IEEE CVPR I*, 339–346.