Joint 5D Pen Input for Light Field Displays

James Tompkin[1,2] **Samuel Muff**[1] **Jim McCann**[1,3] **Hanspeter Pfister**[1,2]
Jan Kautz[4,5] **Marc Alexa**[1,6] **Wojciech Matusik**[7]
[1]**Disney Research** [2]**Harvard SEAS** [3]**Adobe** [4]**UCL** [5]**NVIDIA** [6]**TU Berlin** [7]**MIT CSAIL**

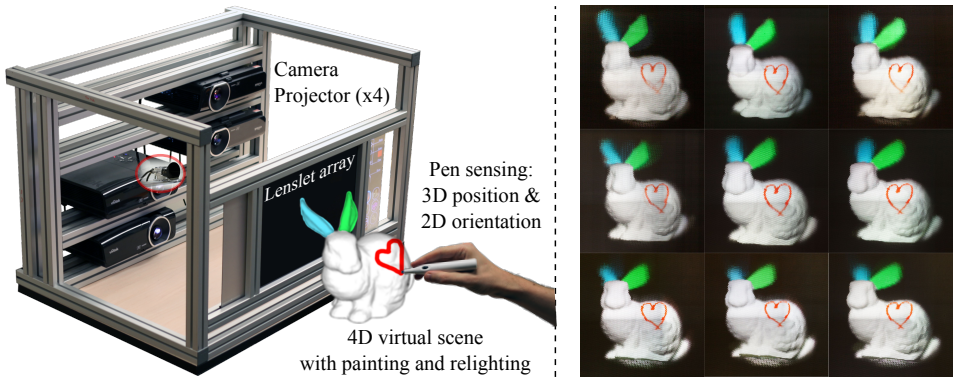

Figure 1. A lenslet array refracts outgoing light from projectors to form a 4D light field display. Simultaneously, the lenslet array refracts incoming IR pen light to a IR camera for 4D light field sensing. This allows us to recover the 3D position and 2D orientation of the pen quickly, to use in interactive applications like object relighting and free-form painting. The camera is behind a projector from this view, and so is made visible in the red ellipse.

ABSTRACT

Light field displays allow viewers to see view-dependent 3D content as if looking through a window; however, existing work on light field display interaction is limited. Yet, they have the potential to parallel 2D pen and touch screen systems which present a joint input and display surface for natural interaction. We propose a 4D display and interaction space using a dual-purpose lenslet array, which combines light field display and light field pen sensing, and allows us to estimate the 3D position and 2D orientation of the pen. This method is simple and fast (150 Hz), with position accuracy of 2–3 mm and precision of 0.2–0.6 mm from 0–350 mm away from the lenslet array, and orientation accuracy of 2° and precision of 0.2–0.3° within 50°. Further, we 3D print the lenslet array with embedded baffles to reduce out-of-bounds cross-talk, and use an optical relay to allow interaction behind the focal plane. We demonstrate our joint display/sensing system with interactive light field painting.

ACM Classification Keywords

I.3.1. Computer Graphics: Hardware Architecture — *Input devices*; H.5.2. HCI: User Interfaces — *Input devices and strategies*

Author Keywords

Light Fields; Joint IO; Through-the-lens Sensing; Pen Input.

UIST '15, November 08 - 11, 2015, Charlotte, NC, USA.

ACM 978-1-4503-3779-3/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2807442.2807477

1. INTRODUCTION

Light field displays — or auto-multiscopic, integral, or 4D displays — show glasses-free binocular stereo with motion parallax. This is often accomplished with 2D lenslet arrays — sometimes called fly's eye lenses — for horizontal and vertical effects. Light field displays allow 3D content to react naturally as the viewer moves around the display. The added angular views are traded off with spatial resolution, and so these displays are often seen as low resolution; however, with the recent push for high DPI displays, light field displays are approaching commercial feasibility [27].

However, relatively little work has addressed the interaction capabilities of light field displays. Many commercial and research techniques allow 3D input, e.g., multiple cameras or infrared projector/camera depth sensing, but these sensing approaches are usually decoupled from the display space and require additional hardware [35, 31, 37, 5, 15, 23, 38]. Existing light field interaction work is promising [17, 16], but has so far consisted either of low precision gesture detection, or of integrating real-world illumination with rendered content.

Many light field applications would be well served by a simple input device — the 4D equivalent of a 2D mouse or pen — yet, so far, it has not been presented. Unweighted, the human hand involuntarily tremors with an amplitude of 0.4 ± 0.2 mm [4], and so any solution must consider this its precision goal. Further, hand motions often reach instantaneous speeds of 5 m/s, and so any input device must be as fast as possible.

To meet this need, we propose a pen-based sensing system which uses the same lenslet array for both display and sensing. To prevent interference, we move sensing to IR wavelengths with a single IR camera and an IR pen. This *joint* mapping between input and output spaces is similar in spirit to other

joint display and input surfaces, like natural 2D touch and pen interfaces. However, to our knowledge, pen-based light field sensing through this joint method has not been demonstrated.

Further, we improve general joint light field display and sensing devices. First, to reduce 'out-of-bounds' cross-talk, we 3D print our lenslet array with in-built baffles. Second, to allow natural interaction behind the display screen, we use an optical relay. These two improvements complement: as the relay rotates the virtual image, the lenslet array is automatically darkened from view. We demonstrate our system with prototype light field painting and scene lighting applications.

1.1 — Integral Imaging and Display

Many display technologies enable some form of multi-view autostereoscopy or 4D display; we focus on integral types and refer to Lueder for the remainder [27]. Lippmann's [26] seminal paper used small lenticular lenses to image an incident light field onto a film-plane — a light field camera. The patterned film, displayed with the same lenses, can reproduce the captured light field. With computation, many modern advances have been proposed [1, 29, 12, 39, 18]. However, none addresses pen-based light field sensing.

1.2 — Combined Optical Sensing and Display

Adding sensing systems to displays *emancipates* pixels [34], and this is commonly accomplished with auxiliary sensing systems. Two video cameras placed behind a semi-transparent projection screen can simultaneously display and sense gestures [40], or a projection screen can be imaging based on frustrated total internal reflection to enable multi-touch 2D sensing [13]. An IR emitter and detector array can be placed behind a traditional LCD to detect gestures [19], or an electronically switchable diffuse/transparent projection screen can be toggled for display and sensing gestures [20]. Displays paired with depth cameras are also possible [2]. However, these works don't address light field-type sensing.

2D displays with LCD apertures can capture light fields and allow low-resolution gesture recognition [17]. A co-axial capture and projection setup enables a context-aware light to enhance scene features by projecting images [36]. Displays can also be lighting sensitive by reacting to surrounding illumination [28], and this extends to 6D passive reflectance field displays that depend on both view and incident illumination [10]. Volumetric displays created with parabolic mirrors and a spinning diffuser can also sense 3D gestures by observing the diffuser with an IR camera [5]. However, these approaches typically use more complex setups.

Through-the-lens or *joint* light field approaches are promising because they reduce complexity while retaining the power and flexibility of existing systems (within spatio-angular trade-off limits), and are a good fit for sensor-in-pixel (SIP) displays or displays overlaid with future thin-film cameras [25].

Cossairt et al. [7] and Hirsch et al. [16] are joint approaches. Cossairt et al. transfer 4D illumination between real and virtual objects for relighting. Hirsch et al. use the incoming illumination for limited interaction, e.g., to modulate transparency in medical volume rendering by the incoming light intensity. We refine these ideas and ground them in strong system evaluation.

Table 1. A broad-view comparison of common input/output systems that integrate sensing and display, either joint or with single camera setups. PS Move is as functional but more accurate than the Nintendo Wiimote. Accurate proprietary device details are difficult to find and are approximations. Sony PS Move numbers are from press interviews (URL), and may include IMU use, whereas we use purely optical sensing.

	2D pen tablet	Camera + wand + TV	Depth sensor + TV	Volumetric (spin mirror)	ProCam + lenslets	Our approach
Spatial resolution	Retina	1080p	1080p	200×160	274×154	198×160
Angular resolution	1 view	1 view	1 view	192 views	7×7	15×15
Glasses-free stereo	No	No	No	Yes	Yes	Yes
Field of view	40° pen tilt	85°	70°	360°	20°	45°
Joint I/O space	Yes	No	No	Yes	Yes	Yes
Range	0 mm	1.2–3 m	1.4–3 m	Undefined		0–350 mm
Response	75 Hz	30 Hz	30 Hz	30 Hz	Input via illumination transfer. Not explicit.	150 Hz
Accuracy	0.5 mm	3cm Z, 1mm XY?*	5–50 mm	Undefined		2 mm / 2 deg
Precision	0.6° tilt	3cm Z, 1mm XY?*	5–50 mm	Undefined		0.2–0.6 mm / 0.2–0.3°
Sense dimension	Pos + tilt	Pos + ori *	Pos	Pos		Pos + 2D ori
Cost	Medium	Low	Low	High	Medium	Medium
Example	Wacom Cintiq	Sony PS Move	MS Kinect	[5]	[16]	

While our design is related, it is also novel: an IR light pen can exploit relatively large depth ranges at high accuracy/precision as it is robust to blur from sensing depth of field limitations. To our knowledge, no system has exploited this to provide fast 5D pen sensing with equal precision to the human hand.

1.3 — 4D Interface Applications

Many applications might benefit from such a sensing device. Existing virtual painting and sketching work could be translated from the more traditional virtual reality settings [8, 22, 11, 21]. Direct 3D modeling would also be compelling [6, 24]. Finally, some works attempt painting 3D objects with 2D devices, and these would also extend to our scenario [30].

1.4 — Contributions

Our paper focuses on systemic advances. Every approach has trade-offs, and so we collate existing combined IO systems (Tab. 1), and discuss the consequences of our trade-offs (§7). Given this, we contribute:

- A light field display/sensing design with separate visible/IR wavelengths, a dual-purpose lenslet array, and one camera.
- Fast pen 3D position and 2D orientation from the camera/lenslets, with measured accuracy and precision.
- Extending the physical interaction space to behind the lenslet array with an optical relay, and reducing out-of-bounds cross-talk and views by 3D printing the lenslet screen with baffles. These complement to automatically darken the original display from relay space.

2. HARDWARE

We exploit a joint imaging and display surface behind a lenslet array to create an interaction and display volume (Fig. 1). In principle, our target platform is future SIP or thin-film camera displays; however, for now, we develop a prototype: An image is projected onto a diffuser on the back of a lenslet array. Each pixel refracts light through a lens in a different direction, creating an outgoing 4D light field display. In reverse, a pen with an LED illuminates the lenslet array, creating an incoming light field. This light is refracted onto the diffuser which is imaged by a camera. We analyze this image on a computer to determine the position and orientation of the pen.

2.1 — Illumination Invariance

To achieve sensing invariance to the displayed image and (some) environment illumination, we move the sensing to IR wavelengths with an IR pen and an IR camera. As most digital cameras are sensitive in near-IR, removing any existing IR shortpass filter and replacing it with an IR highpass or matched IR bandpass filter is sufficient. Further, as near-IR is commonly used for remote control, even bright IR LEDs are cheap. For button control only, we modify a Wiimote to house the IR LED (30° FOV) — none of the Wiimote sensing capabilities are used.

2.2 — Spatio-angular Resolution

All integral devices must reckon with the spatio-angular trade-off: for a given pixel budget, any increase in spatial resolution trades angular resolution. For display pixels, this means trading higher single view resolutions for more views; for sensing pixels, this means trading in-plane resolution for depth resolution. In our joint system, we aim first to achieve human-hand-ability sensing, then second to maximize image quality.

Our goal is to place as many pixels behind our lenses as is feasible: we use four Vivitek D952HD 1080p DLP projectors in a 2×2 grid with overlap, and a Basler Ace acA2040-180km 2k $\times$ 2k camera with an $f = 12.5$ mm lens. For our lenslet array, we wish to balance angular and spatial resolution as we have both display and sensing needs; however, a large lenslet array of a specific lens pitch and field of view is an expensive ($40k) custom engraved mold from most suppliers. Instead, we follow recent work [32] to optimize a lens shape given target lens parameters, and fabricate a sheet of lenses using an Objet 500 Connex 3D printer to create a cheap ($500) custom lenslet array. We adhere a Screen Solutions 'Definition' diffuser to the back of the lenslet array.

Our array contains 198×160 hexagonally-packed aspheric lenses with a 45° field of view and a 2.5 mm diameter, in a sheet 468×328 mm large. After projector overlap, this results in ≈15×15 display pixels behind each lenslet (and so 15×15 different angular views). Some camera resolution is lost due to physical positioning, so a 1700×1200 pixel camera image provides ≈8×8 views of the sensing volume.

2.3 — Out-of-bounds cross-talk

As the viewer or pen moves beyond the lenslet field of view, light begins to move through neighboring lenslets and causes ghosting. In contrast to standard lenslet arrays, we 3D print 0.2 mm black baffles between the lenslets. While not quite opaque due to being very thin, these still reduce light leakage (Fig. 2). This lessens visible cross-talk by causing the display to darken as the viewer moves beyond the field of view, and likewise culls sensing ability when then pen is out of bounds or when cross-talk will make the pen no longer accurate. However, there is still a trade-off, as now some projection and camera pixels will be occluded and unused.

2.4 — Depth of Field

To define the display depth of field, i.e., the region extending in front of and behind the lenslet array in which displayed content appears sharp, we follow Zwicker et al. [41]. We place the t plane at the focal point (back) of the lens sheet (where both projectors and camera are focused), and the v plane one unit distance in front. Objects greater than $\frac{\delta t}{\delta v}$ away from t will be blurred. With spatial sampling $\delta t = 2.7$ mm, and angular sampling $\delta v = 0.052$ mm, our display has a depth of field of 51.6 mm. As noted by Zwicker et al., psychophysically this allows depth effects up to ≈ 5 meters away.

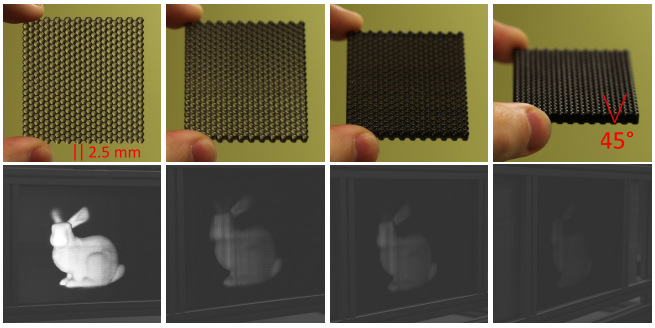

Figure 2. ***Top:*** **A small section of our custom 3D printed lenslet array, which introduces light-blocking baffles to reduce both display and sensing cross-talk.** ***Bottom:*** **Display photographed with identical camera settings and brightened for angle visibility. Vertical line artifacts exist as the printer fails to maintain a consistent 0.2 mm baffle thickness.**

By similar calculation, our sensing has $\delta v = 0.098$ mm and so a depth of field of 27.6mm. This is not the actual range limit, as our LED-based approach is designed to work with blurred light field images and so works accurately up to 350mm.

2.5 — Optical Relay Configuration

Given that the depth of field extends both in front of and behind the lenslet array, how can the user interact within the space behind the lenslet array when the light field pen cannot physically penetrate the lenslet array?

We introduce a second system configuration which uses an optical relay formed from a beamsplitter and a spherical mirror to create a virtual display rotated from the real display. This effectively repositions the light field to where there is no physical impediment to accessing 'negative' depths with the pen. Using the same lens optimization framework as before [32], we parameterize a spherical mirror to minimize ray comas in the virtual display (Fig. 3, and in simulation Fig. 8). A complementary benefit of the combination of baffled lenslets and optical relay is that, with the virtual relay image now at 90° to the lenslet array, the original output is automatically darkened. We imagine this configuration being used in permanent installations where cost is less sensitive, such as at amusement parks.

2.6 — Update Rate

We would like both pen sensing and display to be as fast as possible. For fast camera acquisition, we use a Teledyne Dalsa Xcelera-CL+ PX8 Full CameraLink framegrabber board at 150 Hz. For rendering convenience, we drive all four projectors from an ATI Radeon 6950 graphics board as a 4k×4k canvas. Both boards are daughtered to a PC with an Intel Core i7-2600 and 16 GB RAM. These components are now a little old, but our sensing is fast (§4); however, rendering light field scenes is always likely to be a bottleneck for complex scenes as many views must be rendered per frame (§5).

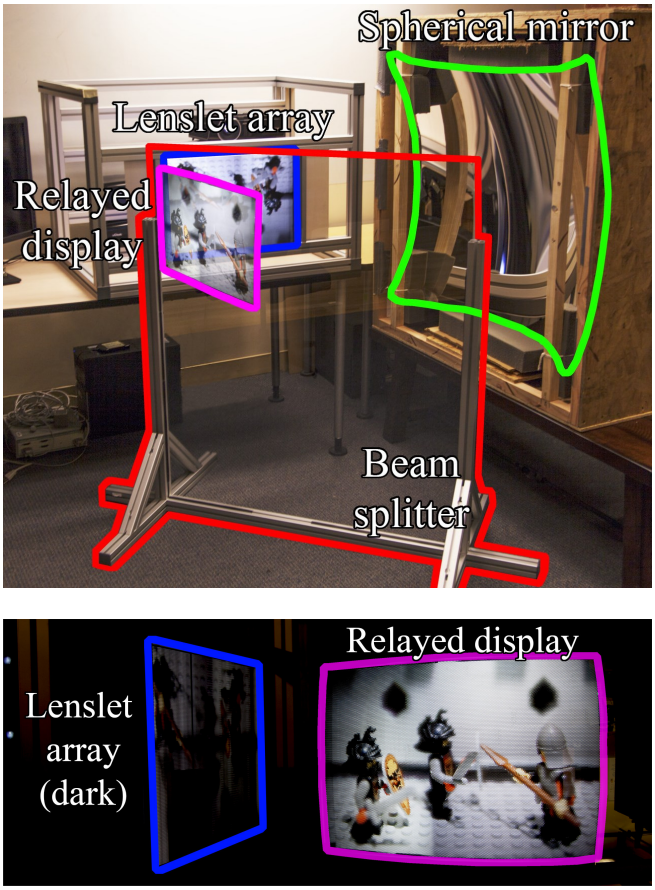

Figure 3. ***Top:*** **To expand sensing to behind the lenslet array (blue), we use a beamsplitter (red) and a spherical mirror (green) to relay the light field (magenta). Please see the appendix for a ray tracing simulation of the light paths.** ***Bottom:*** **The output from the lenslet array is automatically darkened by the 3D-printed baffles.**

2.7 — Costs

Given a lenslet-based light field display, our approach adds a camera and an IR pen ($10 / $40 with buttons). We use a high-speed machine vision camera ($2500), though cheaper alternatives exist as high-speed video is now a feature on smartphones. Light field displays cover a range of prices, with the simplest pico projector and lenslet array system costing ≈$500, ours at ≈$3000, and even, in principle, a SIP version with similar sensing performance and 2/3 display resolution built using a Samsung SUR40 for ≈ $5000. The major cost here is the display itself as our adaptations — the lens sheet in front and an IR remote — would cost ≈ $500 as built.

3. CALIBRATION

To display light field content, we must know the transformation between projector pixels and world rays. Similarly, to sense the pen position and orientation, we must know the transformation between camera pixels and world rays. The lenslet array refracts world rays onto points on the flat diffuser attached to the back plane of the array. As such, rays and pixels are principally related by two perspective transformations, or *homographies*: one between the projector pixel array plane and the lenslet array diffuser plane, and one between the

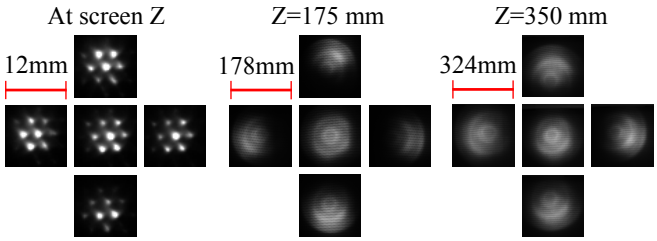

Figure 4. Through-the-lenslet-array camera images of pen rotations to ≈ ±25° in pitch and yaw, at three Z distances. ***PDF zoom recommended.***

camera pixel sensor array plane and the lenslet array diffuser plane. Additionally, the camera suffers radial distortion from the lens, and the projectors suffer both radial and tangential distortion from their lenses.

Calibration is critical to the working of the system, and so we develop an automatic method to accomplish this. However, as our approach follows mostly well-understood best practice, we include full details in the appendix (§A). For the reader, it is sufficient to understand that, following calibration, we know for each projector pixel and for each camera pixel the position and direction of a world ray.

4. PEN SENSING

Given a calibration, our task is to take a camera image and deduce the position XYZ and orientation θ, ϕ of our pen (Fig. 4). To find the pen, we must find where rays focus in space. One of our goals is to sense as fast as possible; as our camera can image at 150 Hz, we have a 6 ms time budget. Our approach is visualized in Figure 5, and follows:

1. The camera images the lenslet array diffuser.
2. Each lenslet sees the pen approximately as a point, which may be blurred to a blob. We compute fast blob contours.
3. For each blob, we compute a center as the mean pixel location weighted by individual pixel intensities, which copes with skewed blobs as the pen is viewed from different angles. This is sufficiently robust to cope with blurring beyond the depth of field boundary.
4. Each center is converted by the calibration to a world ray. Across all lenslets, this creates a bundle of rays which should approximately intersect at our pen light source.
5. A system of linear equations is constructed to solve for the position of the point which is closest to all rays in a least-squares sense [14], see Appendix B. This is our pen position XYZ.
6. Pen direction: This can be thought of as the vector difference between where the pen is and where the cone of light leaving the pen intersects the lenslet array. First, we approximate the center of this cone of light by computing the mean ray origin O_r from all rays intersecting the diffuser plane — the mean of all x,y positions from step 3 in lenslet space, with $z = 0$. Then, the vector direction is simply $XYZ - O_r$. From there, trigonometry will derive pitch and yaw orientations.

Sensing time is a function of the number of rays detected and is proportional to Z, i.e., how many lenslets see the pen. The

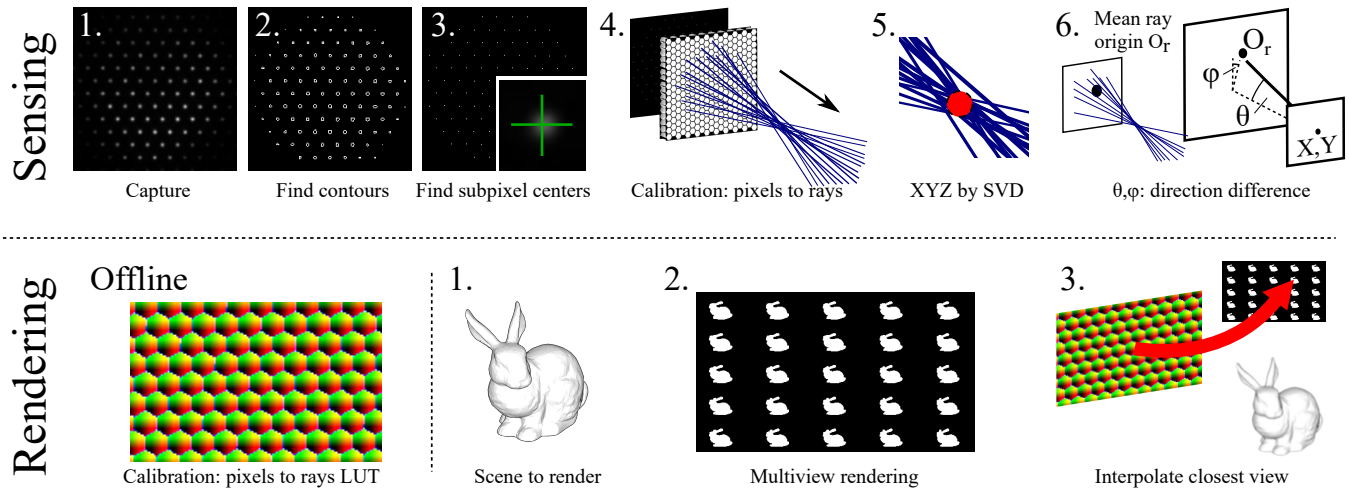

Figure 5. Light field sensing and display. ***Top:*** **The pen position is sensed by a linear least-squares solve of the closest point to our ray bundle.** ***Bottom:*** **Rendering is accomplished by corresponding a pixel to a direction LUT and a multi-view rendering.** ***Subpixel centers may not appear in print.***

time varies between 3–6ms (i.e., worst case ≈150 Hz). To increase the range of the pen, we adjust per frame the future exposure of the camera based on the maximum brightness of the current camera image. We target a brightness of 192/255, and adjust the camera exposure linearly towards or away from this target, up until our 6 ms maximum exposure time.

4.1 — Sensing Accuracy and Precision

We use a Polhemus Fastrak magnetic 6D tracking system as a ground truth to assess our system (≈$6000). The Fastrak is wired, runs at 120 Hz, is statically accurate to 0.8 mm/0.15° root mean square error (RMSE) within 760mm, and precise to 0.006 mm/0.003° within 305 mm (0.076 mm/0.015° within 610 mm). We affix the Fastrak receiver to our pen and measure the magnetic and optical center offsets. By default, the Fastrak filters to smooth the signal, but we present our method *without* filtering, and so we disable this feature. Henceforth, all measurements stated are relative to these errors.

We align the two coordinate systems with Procrustes analysis, and transform Fastrak coordinates to our world coordinates. We define accuracy as RMSE between the two sets of measurements, and precision as the standard deviation of our sensed values whenever the Fastrak sensor detects no movement — a measure of the smallest increment we can reliably detect.

To simulate other spatio-angular trade-offs, we artificially downsample by area mean our camera image by 2× and 4× (to 850×600 and to 425×300) and attempt to discover the pen. For position, we additionally compare to our previously-undocumented light field 3D pen sensing algorithm [33], which is now described in Appendix C.

Position

We measure Z accuracy by moving the pen three times from the front to the back of the volume in the XY center (10 seconds each). To give a representative estimate, we show results from a single trial, with no filtering or aggregation. Figure 9, top left, shows that our approach produces the smoothest and most accurate pen estimate across Z space. X (and Y) accuracy is ±3 mm, but as the pen reaches a screen edge and the number of lenslets which image the pen drops rapidly, then accuracy decreases sharply to ±10 mm. In principle, the difference in camera/projector opening angles will also affect precision, especially at screen edges. However, the drop in imaging lenslet number is a larger error concern.

Precision is measured by moving the pen at 5 mm increments for 10 seconds each. We achieve 0.2–0.4 mm across Z space (Fig. 9, top right), with X (and Y) precision similar at 0.2–0.6 mm and again dropping off at screen edges. This meets our goal of human hand tremble performance.

Orientation

The pen is rotated around the LED in the center of the volume in XY, and at 250 mm in Z, and angular differences are measured (Fig. 9, bottom). Accuracy is ±6°. For precision, we achieve 0.2–0.3° across angle space. Operating range is limited by lenslet and pen fields of view, and is usefully ≈ 50°. There is precision spiking at larger right angles: this is likely a lens manufacturing error as there is no principled reason why this should not also appear to the left.

Analysis

There is clearly a *characteristic* error in accuracy that relates to the space: over many trials, the same distinctive curve in Z appears. We investigated different causes — magnetic interference by moving the FasTrak transmitter and rerunning the experiment, pinhole model limitations regarding refraction and coma, manufacturing errors like deformed lenses or varying refractive index — but this remains future work. More positively, with the Fastrak, we can calibrate out this error by fitting a polynomial function to the aggregated error over many trials. This reliably improves position accuracy to ±2 mm (quadratic), and orientation accuracy to ±2° (cubic; Fig. 9). In a hypothetical commercial display, the repeatability of this correction, assuming a consistent spatio-angular trade-off, would comes down to manufacturing tolerances.

Post correction, the useful range spans from the lenslet surface at 18mm through to 350 mm (with the focal plane at 0mm). Down-sampling 2× still gives useful results, but at 4× the

Figure 6. ***Top left:*** **A desert scene is drawn with freehand sand, polyline tree, and Bezier curve cloud. To demonstrate that the people are drawn at different depths, the scene has been refocused.** ***Top right:*** **Digital makeup is added to a human head geometric model, with horizontal parallax shown.** ***Bottom:*** **Horizontal display parallax is shown through a meadow scene (no scene rotation). All captured without the optical relay.**

system is no longer useful. This suggests that still-acceptable sensing performance could be achieved for systems wishing to opt for smaller lenses and more spatial resolution.

Our raw orientation accuracy is less than theoretical limits. At a screen distance of 300 mm, with a position error of 5 mm in XYZ, it is theoretically possible to detect pitch and yaw to 1°. Errors compound: non-uniform pixel sampling over lenslets causes inaccuracy, as do lenslet manufacturing errors which direct light to incorrect pixels. This affects calibration because we assume homographic (plus radial) distortions, and these small perturbations require local non-linear correction warps. Our average calibration error of half a pixel corresponds to 2.5° error. Further, the IR LED itself is suboptimal: there is a slight mismatch between LED FOV (30°) and lens FOV (45°), which reduces potential angular samples, and LED light non-uniformity across the cone also affects sensing.

5. RENDERING

A light field scene is rendered into an auxiliary texture as a 20×20 grid of sheared views, each of 192×108 pixels, with shearing directions uniformly distributed over the field of view of the lenslets. Then, this texture is mapped to the screen using a precomputed lookup table which stores, for every pixel, the texture coordinate of the nearest ray among all sheared views. Finally, we correct for projector overlap by computing per pixel the number of intersecting projectors and weighting their output by the appropriate gamma curve. Given that we only have ≈15×15 ray directions per lens, oversampling to 20×20 allows us to accommodate imperfectly aligned lenses/pixels and so reduce angular quantization artifacts. Even at low spatial resolution, multi-view rendering is expensive for simple geometries: three 5k triangle Stanford bunnies is really 6 million triangles; on our setup, this runs at 25 Hz.

Finally, as the display is low spatial resolution, information like text is difficult to read. To compensate, we create a 100×328 mm 2D companion display with approximately 250×2000 of the pixels from our projectors. This allows high-resolution information display to the user, such as for application instructions and diagnostics.

6. APPLICATIONS AND INTERACTIONS

We demonstrate our system with interactive light field painting. This uses accurate free-form interactions, which previous works have demonstrated to be an expressive medium [8, 22, 24]. Certainly, this is not suitable for some tasks, e.g., technical drawing, where waving a pen in space is not an analog of precise 2D mousing; however, it would be a suitable tool for sketching directly in 3D during early design stages. While we include some photographs here (Fig. 6), the majority of the demonstration is via the supplemental video.

As the pen acts as a cursor, it is possible to extend simple 2D drawing tools, such as a paint brush, to work directly in 3D, with the 3D drawings immediately visible from many views in stereo. In 2D, changing brush color usually requires at least two interactions as color is a 3D space; here, we map hue, saturation, and brightness directly to the volume, with hue radially around the XY center, saturation as distance from the XY center, and brightness as the inverse distance from the lenslet array. This allows color to be picked by a single point in space. That sensing range is significantly extended over display depth of field is useful to provide finer (or greater) brightness control.

We aid building objects from 3D lines and curves with 3D proximity point snapping. When defining curves, the tangent direction also extends to 3D, which allows simple definition of elegant shapes. As the volume is calibrated, drawn objects can be measured in millimeters. With 2D interfaces, 3D primitive are often difficult to rotate, scale, and translate into position. In our system, position is given absolutely by the pen. Once the object is placed, we parameterize scale and rotation by the

drag distance away from the display and the drag direction in 3D (arcball) respectively. The extended range again provides finer (or greater) scaling and rotation.

Virtual camera motion is also often tricky in 2D; with our system, the camera can be moved and rotated directly in the volume as if it were physically in the hand. To overcome the limited display depth of field, moving the camera in and out in *Z* acts to refocus the scene so that desired parts are shown in detail, and this interaction benefits from the extended range. This can be used for navigation or for visual effects.

Analogous to cameras are lights, which in 2D have the same placement issues as cameras. For us, lights are directly parameterized from the pen to easily illuminate the scene, e.g., with a spot light. As the light is virtual, in comparison to existing real illumination light field transfer works, they can have any particular characteristic, e.g., narrow/wide illumination arcs.

Finally, some operations are harder to move to 3D, such as general 2D image operations like blur. However, in principle, the scope of interactions for a general purpose 4D sensing device with derived position and orientation is large, as we provide the 4D equivalent to a 2D pen.

6.1 — User Experience and Feedback

This paper documents an advanced version of our previously demonstrated system [33], which was used by hundreds of attendees over 5 days at the SIGGRAPH Emerging Technologies exhibition. They largely found it intuitive to view and paint directly in 3D with parallax (after a hands-on tutorial). One striking effect noticed with novices was observed within the first interactions, which often triggered a conceptual understanding: asked to draw a 3D object, most users begin to draw in 2D perspective, but then realize that our pen holds new possibilities as painting is more akin to sculpture in 3D.

In this environment, many simultaneous viewers had the freedom to move around the display and see in stereo+parallax what the primary user was drawing. As the display is large and the viewing space extends at least out to 5 m, occlusion is not a significant problem, especially as often the user will stands off-center so that their arm can span the display volume.

Arcball-syle rotation definition was easier than direct mapping between pen orientation and object rotation, because 1) wrist yaw is uncomfortable and slow (c.f. roll or pitch) and has narrow angular limits, 2) our lenslet FOV also has narrow angular limits, requiring many drag-drop subrotations to achieve a target orientation. Pen orientation is useful for relighting, providing natural spot light direction control, and for oriented strokes, such as drawing a 3D ribbon, which gives the appearance of calligraphy from fixed viewing angles.

Drawing in Z while looking straight down the Z-axis is more difficult. Users benefited from moving their heads (or the object) around to get better view. Precise in-plane (XY) drawing can be difficult in mid-air, so we add a toggle to lock the pen Z. Finally, often users wished to draw on the surface of objects, but placing the pen exactly requires practice. Hence, we allow strokes to be drawn on object surfaces by intersecting the pen direction vector with the scene and adding a surface reticle.

7. DISCUSSION

A consumer light field display with joint sensing is largely predicated on SIP-like displays existing at high density in the future (e.g., future SUR40, or thin-film camera based [25]). Costs for SIP-like displays are currently significantly higher than for non-sensing panels, but we posit that the usefulness of their intrinsic properties, i.e., constant sensing resolution over display space/volume, will increase over time with higher densities, and so their cost will decrease.

While the fundamental limits of light field displays do not disappear with higher density (e.g., limited depth of field [41]), and while diffraction ultimately limits pixel density, there is still significant headroom in large-size (SIP-like) displays vs. in camera sensors in general: By Abbe, green light has a minimum feature size of 0.25 micrometers. Modern smartphone cameras have pixels of ≈1 micrometer (Nokia 1020 - 1.12), whereas our sensing pixels are 0.25 *millimeters*. 600 dpi displays are common (Galaxy S6 - 576), with 0.04mm pixels. Ignoring fabrication issues, this would give a 11000 x 7500 SIP display of our size, for 10x10 angular views at ≈720p, or 6x6 views at ≈1080p.

Thus far, making the spatio-angular trade-off has been of niche appeal. As displays increase in pixel density to beyond eye capabilities for 2D images (e.g., 8k desktop monitors in 2016), these spare pixels can be used without sacrificing visible 2D resolution. As Hirsch et al. state, one goal of joint IO systems research is to inspire manufacturers to produce increased density SIP-like displays, and to this effort we show that relatively few sensing pixels per lens can provide fast, accurate, and precise 5D pen input, creating a responsive joint display and sensing system.

That said, the most visible limitation of the current system is spatial resolution as we must consider our sensing needs with relatively large lenses (though our results on down-sampled inputs show there is some leeway). Thus, there is a conflict between sensing range and best display appearance: As the image is low resolution, the natural response is to take a step back; however, the sensing begins to lose accuracy beyond 350 mm from the lenslet array (or relay virtual image). Ignoring future displays for now, we suggest application in operator-viewer situations, such as a performance or classroom, where a tutor demonstrates concepts in 4D to viewers farther back.

This conflict has roots in the limited depth of field, which affects the range of accurate sensing to near-display only. Auxiliary non-joint optical sensing systems, e.g., PS Move with IMU, sense far at 1.2–3 meters. These are difficult to get to work at near distances, e.g., both zero display distance and up to 350mm away, and in this way the high sensing resolution close to the display is an attribute of our joint approach. Considering far-distance sensing, while it would be compelling if our needed IR LED was the one inside a TV remote control, we cannot achieve 'from the couch' sensing.

Moving beyond standard TV parallels, our system could be seen as fish tank VR — a VR window in the real world rather than full immersion — but without the need for glasses or head tracking. One benefit to position-tracked head-mounted

VR systems is that they can provide high spatial resolution, e.g., PS Morpheus with PS Move. However, each user must wear a headset to view or interact. Our joint light field optical path draws display light *exactly* in the real world where the pen was placed, with stereo and parallax, for multiple people, and with no head-tracked glasses.

In principle, it should be possible to detect different ray focal points for multiple pen support, though multiple pens will interfere if they are both physically close. Far Hamming distance binary flash patterns would identify each pen. Further, in the future, we would like to detect roll orientation along with pitch and yaw by using a coded aperture over the IR LED. In principle this is a simple addition; yet, it might be a challenge to obtain precise roll measurements as the spatial sensing resolution is limited. Finally, in our rendering, we over-sample the view space, and it might be possible to cluster only used views given the calibration.

8. CONCLUSION

We present a simple joint design for light field sensing and display, suitable for future SIP-type displays. Our prototype senses an IR pen at 150 Hz through a lenslet array to yield pen 3D location and 2D direction. We use a 3D printed lenslet array to achieve 2–3 mm position accuracy and 0.2–0.6 mm precision, and 2° orientation accuracy and 0.2–0.3° precision. We reduce sensing cross-talk using baffles between lenslets and use an optical relay to allow a larger working volume. We demonstrate our system with interactive light field painting.

Acknowledgements

Thank you to Olivier Bau, Daniel Haehn, Xavier Snelgrove, and Stanislav Jakuscievski for their hard work and support; Maryam Pashkam and Ken Nakayama for the Polhemus Fastrak; and Lara Booth for her artistry. We thank the (New) Stanford Light Field Archive for the photographic light fields, the Stanford 3D Scanning Repository for the bunny model, and I-R Entertainment Ltd., Morgan McGuire, and Guedis Cardenas for the human head model. Marc Alexa thanks support from grant ERC-2010-StG 259550 (“XSHAPE”), and James Tompkin and Hanspeter Pfister thank NSF CGV-1110955.

REFERENCES

1. E.H. Adelson and J.Y.A. Wang. 1992. Single Lens Stereo with a Plenoptic Camera. *IEEE Trans. Pattern Anal. Mach. Intell.* 14 (1992). Issue 2.
2. H. Benko and A. Wilson. 2009. DepthTouch: Using depth-sensing camera to enable freehand interactions on and above the interactive surface. Technical Report MSR-TR-2009-23. (2009).
3. Filippo Bergamasco, Andrea Albarelli, Emanuele Rodola, and Andrea Torsello. 2013. Can a Fully Unconstrained Imaging Model Be Applied Effectively to Central Cameras? *IEEE CVPR* (2013).
4. P.R. Burkhard, J.W. Langston, and J.W. Tetrud. 2002. Voluntarily simulated tremor in normal subjects. *Clinical Neurophysiology* 32, 2 (2002), 119 – 126.
5. A. Butler, O. Hilliges, S. Izadi, S. Hodges, D. Molyneaux, D. Kim, and D. Kong. 2011. Vermeer: direct interaction with a 360 viewable 3D display. In *ACM UIST*. 569–576.
6. J. Butterworth, A. Davidson, S. Hench, and M.T. Olano. 1992. 3DM: A three dimensional modeler using a head-mounted display. In *ACM I3D*. 135–138.
7. O. Cossairt, S. Nayar, and R. Ramamoorthi. 2008. Light field transfer: Global illumination between real and synthetic objects. *ACM Trans. Graph.* 27, 3 (2008).
8. M.F. Deering. 1995. HoloSketch: a virtual reality sketching/animation tool. *ACM Trans. Comput.-Hum. Interact.* 2, 3 (1995), 220–238.
9. Anders Eikenes. 2012. Intersection Point of Lines in 3D Space. MATLAB Central File Exchange. (2012). Retrieved January 10, 2015.
10. M. Fuchs, R. Raskar, H.P. Seidel, and H.P.A. Lensch. 2008. Towards passive 6D reflectance field displays. *ACM Trans. Graph. (Proc. SIGGRAPH)* 27, 3 (2008).
11. H. Gardner, D. Lifeng, Q. Wang, and G. Zhou. 2006. Line Drawing in Virtual Reality using a Game Pad. In *7th Australasian User Interface Conference*, Vol. 50.
12. T. Georgiev, K.C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala. 2006. Spatio-Angular Resolution Tradeoff in Integral Photography. In *EGSR*.
13. J.Y. Han. 2005. Low-cost multi-touch sensing through frustrated total internal reflection. In *ACM UIST*.
14. R. I. Hartley and P. Sturm. 1997. Triangulation. *Computer Vision and Image Understanding* 68, 2 (1997), 146 – 157.
15. S. Heo, J. Han, S. Choi, S. Lee, G. Lee, H.E. Lee, S.H. Kim, W.C. Bang, D.K. Kim, and C.Y. Kim. 2011. IrCube tracker: an optical 6-DOF tracker based on LED directivity. In *ACM UIST*. 577–586.
16. M. Hirsch, S. Izadi, H. Holtzman, and R. Raskar. 2013. 8D: Interacting with a Relightable Glasses-free 3D Display. In *ACM SIGCHI*. 2209–2212.
17. M. Hirsch, D. Lanman, H. Holtzman, and R. Raskar. 2009. BiDi Screen: A Thin, Depth-Sensing LCD for 3D Interaction using Lights Fields. In *ACM Trans. Graph.*
18. M. Hirsch, G. Wetzstein, and R. Raskar. 2014. A compressive light field projection system. *ACM Trans. Graph.* 33, 4 (2014), 58.
19. S. Izadi, S. Hodges, A. Butler, A. Rrustemi, and B. Buxton. 2007. ThinSight: Integrated optical multi-touch sensing through thin form-factor displays. In *ACM UIST*.
20. S. Izadi, S. Hodges, S. Taylor, D. Rosenfeld, N. Villar, A. Butler, and J. Westhues. 2008. Going beyond the display: A surface technology with an electronically switchable diffuser. In *ACM UIST*.
21. M. Kavakli and D. Jayarathna. 2005. Virtual Hand: An Interface for Interactive Sketching in Virtual Reality. In *Proc. CIMCA*. 613–618.
22. D.F. Keefe, D. Acevedo, J. Miles, F. Drury, S.M. Swartz, and D.H. Laidlaw. 2008. Scientific Sketching for Collaborative VR Visualization Design. *IEEE TVCG* 14, 4 (2008), 835–847.
23. D. Kim, O. Hilliges, S. Izadi, A.D. Butler, J. Chen, I. Oikonomidis, and P. Olivier. 2012. Digits: Freehand 3D

Interactions Anywhere using a Wrist-worn Gloveless Sensor. In *ACM UIST*. 167–176.

24. M. Koike and M. Makino. 2009. CRAYON: A 3D Solid Modeling System on the CAVE. In *Proc. ICIG*. 634–639.
25. A. Koppelhuber and O. Bimber. 2014. LumiConSense: A Transparent, Flexible, and Scalable Thin-Film Sensor. *IEEE CG&A* 34, 5 (Sept 2014), 98–102.
26. G.M. Lippmann. 1908. La Photographie Integrale. *Comptes-Rendus* 146 (1908), 446–451.
27. E. Lueder. 2012. *3D Displays*. Wiley.
28. S.K. Nayar, P.N. Belhumeur, and T.E. Boult. 2004. Lighting Sensitive Display. *ACM Transactions on Graphics* 23, 4 (Oct 2004), 963–979.
29. R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. 2005. Light Field Photography with a Hand-Held Plenoptic Camera. Stanford University CSTR 2005-02. (April 2005).
30. J. Schmid, M.S. Senn, M. Gross, and R.W. Sumner. 2011. OverCoat: An implicit canvas for 3D painting. *ACM Trans. Graph.* 30, 4, Article 28 (2011), 10 pages.
31. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. In *IEEE CVPR*. 1297–1304.
32. J. Tompkin, S. Heinzle, J. Kautz, and W. Matusik. 2013. Content-adaptive Lenticular Prints. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, Vol. 32.
33. J. Tompkin, S. Muff, S. Jakuschevskij, J. McCann, J. Kautz, M. Alexa, and W. Matusik. 2012. Interactive Light Field Painting. In *ACM SIGGRAPH Emerging Technologies*. Article 12, 1 pages.
34. J. Underkoffler, B. Ullmer, and H. Ishii. 1999. Emancipated Pixels: Real-world Graphics in the Luminous Room. In *ACM SIGGRAPH*. 385–392.
35. A. Vorozcovs, A. Hogue, and W. Stuerzlinger. 2005. The Hedgehog: a novel optical tracking method for spatially immersive displays. In *IEEE VR*.
36. O. Wang, M. Fuchs, C. Fuchs, J. Davis, H.-P. Seidel, and H.P.A. Lensch. 2010. A Context-Aware Light Source. In *IEEE ICCP*. 1–8.
37. R. Wang, S. Paris, and J. Popović. 2011. 6D hands: Markerless hand-tracking for computer aided design. In *ACM UIST*. 549–558.
38. F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler. 2013. Analysis of the Accuracy and Robustness of the Leap Motion Controller. *Sensors* 13, 5 (2013).
39. G. Wetzstein, I. Ihrke, D. Lanman, and W. Heidrich. 2011. Computational Plenoptic Imaging. *CGF* 30, 8 (2011).
40. A.D. Wilson. 2004. TouchLight: An imaging touch screen and display for gesture-based interaction. In *Proc. ICMI*. 69–76.
41. M. Zwicker, W. Matusik, F. Durand, and H.P. Pfister. 2006. Antialiasing for Automultiscopic 3D Displays. In *Eurographics Symposium on Rendering*.

APPENDIX

A. CALIBRATION

We solve calibration classically as an optimization: to estimate homography and distortion model parameters which minimize the error between points of correspondence. For the camera, these are between light imaged through the lenslet array and the known hexagonal structure parameters of the lenslet array; for the projectors, these are between a checkerboard pattern projected onto the diffuser and imaged by the camera, and the known pixel checkerboard pattern locations. Then, all pixels and lenslet positions are related by chaining transformations (and their inverses).

We use the Brown-Conrady lens distortion model for both cameras and projectors, which considers a distortion center (2 dims.), with radial and tangential terms (2 dims. each), plus a homography (8 dims.) for the planar geometric alignment. In total, there are 14 dimensions to optimize per device.

This approach relies on the camera to image projector pattern corners which, if the optical path were truly separated into visible and IR wavelengths, shouldn't be possible. With a long-enough exposure, it is still possible to image the projector patterns through the IR filter covering the camera. We proceed:

1. An IR pen illuminates the center of the lenslet array, from a known distance (typically as large as possible), in a dark room. Then, the automatic calibration routine begins:
2. We capture images:
   (a) To find lens centers, one long exposure to image the pen from afar, with all black projection.
   (b) For each projector, one full-black and one full-white projection capture, to find white/black levels as seen by the camera (no pen).
   (c) For each projector, one checkerboard pattern (no pen).
3. We find camera correspondences from image a). Given local image maxima, we cluster via k-means using a neighbor distribution model prior on feature points, based on the lenslet hexagonal grid. In polar coordinates, we expect the six closest neighbors at $(0, \pi/3, 2\pi/3, \pi, -\pi/3, -2\pi/3)$. To assign lenslet indices to pixel positions, we pick one $(0,0)$ point, and propagate coordinates outwards to maxima using a maximum likelihood estimator. For instance, the maxima closest to the 0° angle cluster in the neighbor distribution model will likely have the coordinates of the current point plus $(1,0)$. This effectively eliminates outliers because they are unlikely to correspond to any coordinate.
4. We find projector correspondences from images b) and c). This time, from the checkerboard pattern, we expect neighbors at $(0, \pi/2, \pi, -\pi/2)$ to correspond to our known projector pixel checkerboard corner locations.
5. For each device, to fit the geometric transformation parameters, we minimize in an unconstrained non-linear optimization the error between the known model coordinates and the sub-pixel image locations.

Calibration typically has a mean error of half pixel for both cameras and projectors, which is respectively approximately 2.5° and 1.5° as angular error, though this is not systematic.

Algorithm 1 Compute the closest point to a set of skew lines in a least-squares sense. $\circ$ is Hadamard or element-wise product.

Require: **P** - $n \times 3$ matrix of line starting points.
Require: **N** - $n \times 3$ matrix of normalized line vectors.

function CLOSESTPOINT(**P**, **N**)
$S_{XX} = N_{*,1} \circ N_{*,1} - 1$
$S_{YY} = N_{*,2} \circ N_{*,2} - 1$
$S_{ZZ} = N_{*,3} \circ N_{*,3} - 1$
$S_{XY} = N_{*,1} \circ N_{*,2}$
$S_{XZ} = N_{*,1} \circ N_{*,3}$
$S_{YZ} = N_{*,2} \circ N_{*,3}$
$A = \begin{bmatrix} \sum S_{XX} & \sum S_{XY} & \sum S_{XZ} \\ \sum S_{XY} & \sum S_{YY} & \sum S_{YZ} \\ \sum S_{XZ} & \sum S_{YZ} & \sum S_{ZZ} \end{bmatrix}$
$B_X = \sum(P_{*,1} \circ S_{XX} + P_{*,2} \circ S_{XY} + P_{*,3} \circ S_{XZ})$
$B_Y = \sum(P_{*,1} \circ S_{XY} + P_{*,2} \circ S_{YY} + P_{*,3} \circ S_{YZ})$
$B_Z = \sum(P_{*,1} \circ S_{XZ} + P_{*,2} \circ S_{YZ} + P_{*,3} \circ S_{ZZ})$
$B = [C_X, C_Y, C_Z]^\top$
Solve $Ax = B$
end function

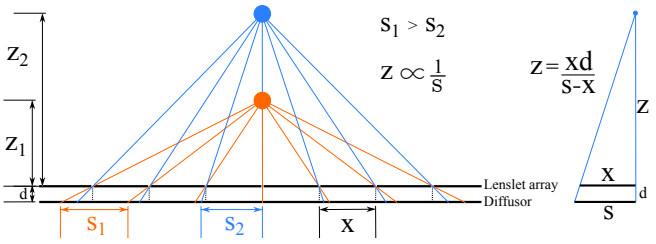

Figure 7. Computing z by similar triangles from a pin-hole lenslet model. s is the distance between the projections of the light pen in the lenslets. s decreases as z increases, until $s = x$ when $z = \inf$. x is the distance between lenslet centres and d is the focal length of the lenslet array.

B. PEN POSITION SOLVE

Given a bundle of rays, we wish to minimize the sum of distances of the point to each ray. We follow Eikenes [9] as per Algorithm 1. We solve this system with a Jacobi SVD.

C. ALTERNATIVE 3D PEN SENSING

In our analysis (§4), we compare against an alternative 3D pen sensing method [33]. This method does not sense pen direction. It exploits the predictable pattern that a point light creates through a lenslet array (Fig. 7), and begins as in the 5D sensing algorithm (§4), but diverges after Step 1c):

1. Find local maxima in camera image (Steps 1a-c in §4).
2. Convert local maxima positions into world space millimetres via calibration transformations.
3. *Compute s:* We measure the mean distance between all neighborhood local maxima. This operation is $O(n^2)$, but can be accelerated with a spatial binning data structure as we know the hexagonal lenslet pattern to search within. This is a robust sub-millimeter estimate as it is averaged over all neighboring lenslets.
4. *Compute z:* This can be recovered by similar triangles (Figure 7) given the lenslet pitch x (2.7mm) and focal length d (3.02mm), as $z = xd/(s-x)$.

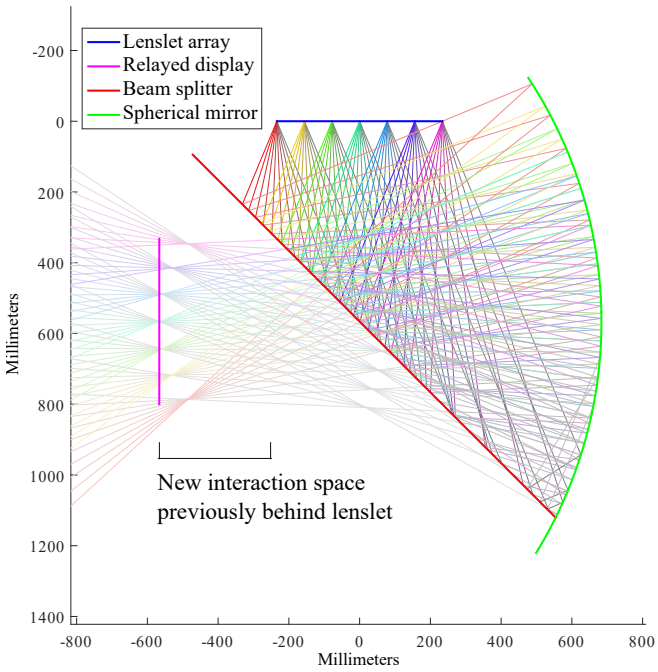

Figure 8. Lenslet array, beamsplitter, and curved mirror work together to rotate the view volume 90°. This effectively allows the pen to move behind the lenslet array without obstruction. This simulation shows only the rays of light that form the relayed display. Ray brightness is reduced by a half each time they interact with the beamsplitter.

5. *Compute x, y:* The mean of all world space maxima rays (step 2) intersected with the z plane.

D. OPTICAL RELAY

Figure 8 simulates light rays leaving the lenslet array, reflecting against the beamsplitter into a spherical mirror, being focused, then reflected again to the beamsplitter, transmitted through, and converging to form a relayed image. For sensing, the direction of the rays is reversed. At each beamsplitter interaction, half the light is equivalently transmitted or reflected (we exclude these other rays from Fig. 8), which reduces brightness and so Z range. One corrective option is to increase camera gain; another is to use a higher-output light source as shown in our supplemental video. Some geometric distortion is visible, which requires further calibration. Unconstrained ray model calibration methods may ease this process [3], though we leave this for future work. Further, the relay flips the display horizontally and vertically. This is not a problem for pen interaction as the optical paths are shared, but our 2D secondary display must be flipped to appear correct.

As the display now free floats in space, it is harder to judge the interaction volume size without the lenslet array and frame reference. One way to overcome this is to build a physical box within which interaction is guaranteed. Another issue is sensing very near to $z = 0$, typically ± 20 mm, as the pen light illuminates very few lenslets. Here, the ray convergence approach for position and orientation sensing is unreliable. Instead, we switch to a simple XY position based on peak brightness, and forfeit the orientation sensing. This clamps Z to 0, and creates the effect of the focal plane being sticky.

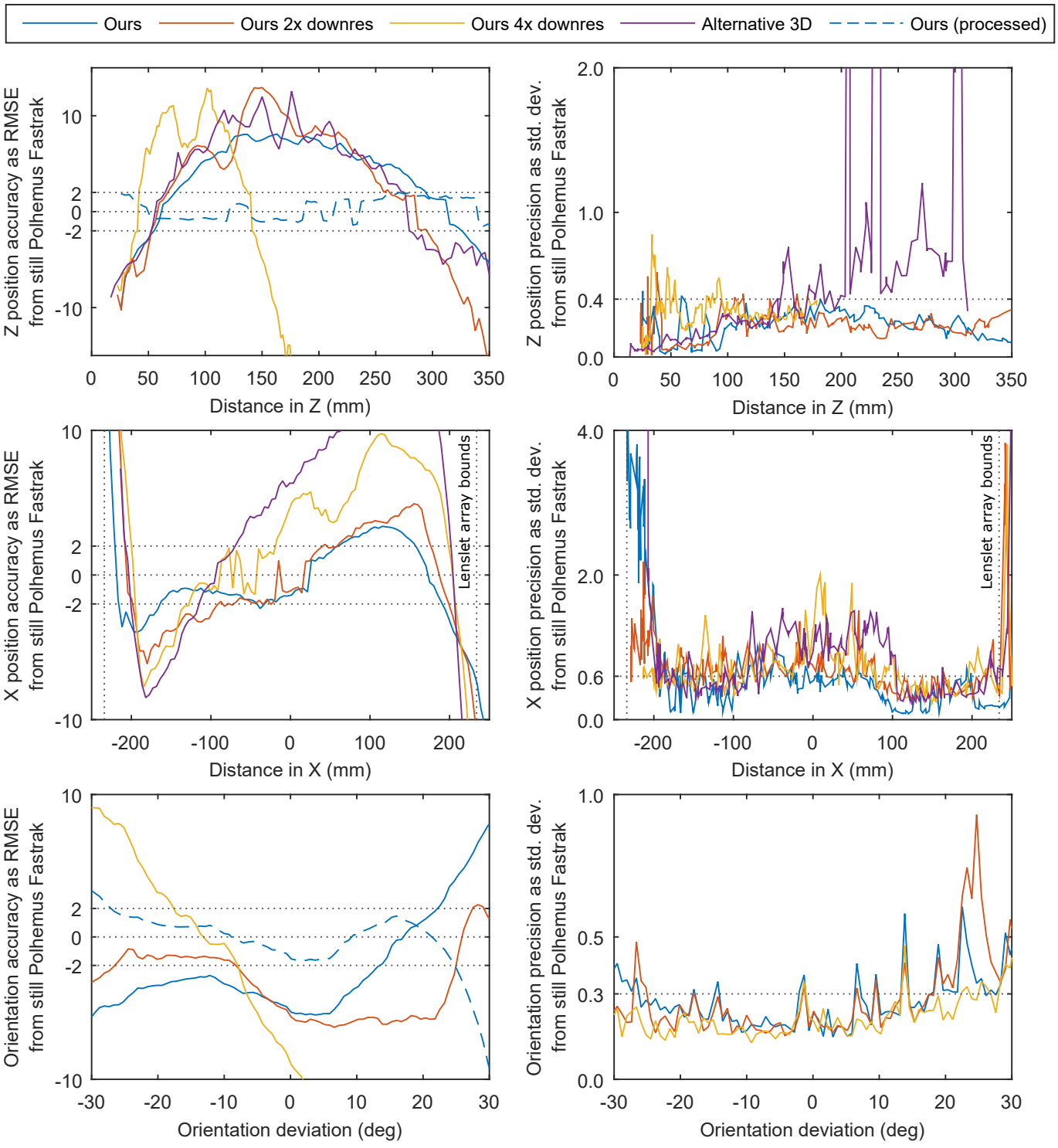

Figure 9. ***Top:*** **Position accuracy/precision against Z.** ***Middle:*** **X accuracy/precision with the pen at 175mm in Z.** ***Bottom:*** **Orientation accuracy/precision against pitch/yaw angle deviation from Z. All graphs show raw data, except for "Ours (processed)", which improves accuracy by calibrating for characteristic error. Even with 2× downsampled input data, it is possible to achieve reasonable accuracy and precision; however, at 4× there is too little information. The 2× downsampling can even be seen to be partially more accurate in orientation sensing, though this may be due to the reduction in noise from the synthetic downsampling. The 'Alternative 3D' position sensing is from our previous undocumented system demonstration [33], as explained in Appendix C.**