Time-Lapse Light Field Photography With a 7 DoF Arm

John Oberlin and Stefanie Tellex

***Abstract*—A photograph taken by a conventional camera captures the average intensity of light at each pixel, discarding information about the angle from which that light approached. Light field cameras retain angular information about the rays they collect, allowing re-integration of rays during post processing. Consumer light field cameras have small apertures, and laboratory camera arrays with large baselines are expensive and not portable. In this paper we demonstrate time-lapse light field photography with the eye-in-hand camera of the Baxter robot. Using the eye-in-hand we can collect light densely and precisely over large distances. The collected rays can be refocused in software with varying lens and aperture properties to form conventional 2D photographs. This refocusing allows us to perform 3D reconstruction and segmentation and suggests approaches for enabling existing computer vision algorithms to robustly handle optically active surfaces. The techniques in this paper can contribute to robotic visual systems for object manipulation as well as media collection systems for virtual reality devices.**

I. Introduction

Robots move themselves in the world and to be supremely useful to us they should be able to move other objects around in the world as well. In order to move those objects, the robot must be able to perceive them. IR based depth cameras are impressively capable but suffer from calibration difficulties which make precise alignment of RGB and D fields difficult. Additionally, IR based depth cameras are not well suited for large scale or outdoor use due to interference from each other and the sun. In many settings we would like robots to work under the same constraints and assumptions that we do. Therefore it would be convenient if a robot could perform all of its duties with an optically passive RGB sensor.

Computer vision has become more accessible in recent years. Nonetheless, a fixed camera can be fooled and if it is fooled it cannot recover. If a camera can move and collect additional views at inference time, many more options are available for the solution of any given problem. The more degrees of freedom a camera has, the more views and the more options. But what is the right way to make use of all of these views? We suggest that light field photography, or plenoptic photography, provides natural and powerful avenues of inference for object classification, segmentation, localization, and manipulation using 3D reconstruction and 2D computational photography.

Robots with 7 DoF arms are becoming standardized and less expensive. Eye-in-hand layouts with a camera next to the end effector can facilitate visual servoing and other activities. Baxter has two 7 DoF arms each with an eye in hand. Furthermore, the encoders in Baxter provide pose annotation in position and orientation for the end effector that is accurate enough to enable metrically calibrated light field photography

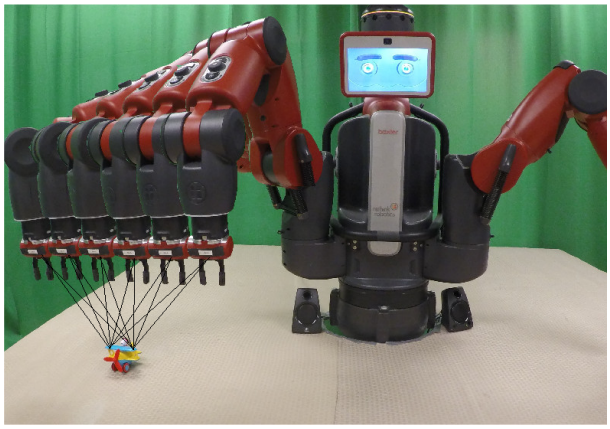

Fig. 1: When collecting images at many locations parallel to a target plane, each pixel in a single image describes light approaching the camera from a unique angle. The light emanating from a single point in space is captured in different pixels across different images.

if images of a stationary target can be collected over time. Time lapse light field photography has precedent [9] , but the movement is typically constrained to a few dimensions. Fixed camera [8] and microlens [3] arrays are stable once calibrated and can capture angular information from many directions simultaneously, but camera arrays are not very portable and microlens arrays do not have a very large baseline. Baxter's arm allows us to densely collect images (in sub millimeter proximity to each other) across large scales (about a meter) over 6 DoF of pose in a 3D volume 5 4. This enables the study of light fields in a diverse variety of modes on an widely available piece of equipment (Baxter), and to our knowledge may be the most flexible and accessible apparatus for light field research despite the limits ultimately imposed by joint encoder quantization and a relatively inexpensive camera.

II. Our Software Camera

An everyday photograph describes the mean intensity of the light hitting each of its pixels. Light field photography retains not only intensity information but also information about the angle from which light approaches in an image. A light field captures phenomena such as parallax, specular reflections, and refraction by scene elements to a much better degree than a single photograph. There are many ways to record a light field. We start by collecting images while moving the camera in a plane and recording the camera location for each image.

When collecting images at many locations parallel to the target image plane, each pixel in a single image describes

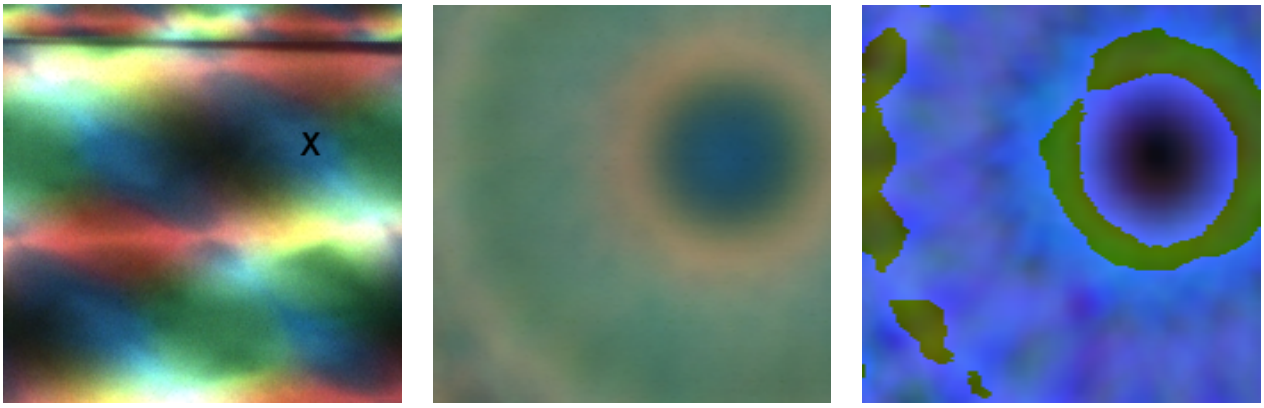

Fig. 2: In order to accurately focus rays in software, we must be able to determine the global (x, y, z) coordinate of the origin point of a ray given the camera pose, an origin depth, and the pixel coordinate (i, j) of that ray. Our transformation from pixel coordinates to world coordinates depends upon knowing the projection of the end effector into the image plane at four heights. We can obtain these values automatically by aiming the camera at a textured surface and spinning the end effector. The point under the gripper remains fixed while the other points move in a circle. Left: Our calibration target, a superposition of three plane waves, as viewed from the wrist camera. The black "X" is where the end effector would touch down on the paper at its axis if it moved toward the image plane. Middle: A time lapse averaged image of the calibration target viewed through the wrist camera as the end effector spins about its axis. This "smearing" average is useful for finding fixed points under camera motions. The projection of the end effector into the image plane is clearly visible as a blue dot. Right: The variance of the colors which contributed to the time lapse average. Darker is lower variance. The projection of the end effector is fixed in the image during the spin, so it has the lowest variance.

light approaching the camera from a unique direction. The light emanating from a single point in space is captured in different pixels across different images. Whereas a pinhole camera would assign precise angular information to each pixel, real cameras have apertures of nonzero diameter. This means that the light hitting a pixel is collected over a range of angles. The camera we use has a very large depth of field, so most of the image is in focus and we do not deviate too much from the pinhole model. We can use ray casting to reproject the rays from all the images as if they has emanated from a common depth, thereby mimicking the action of a lens, and forming a new *refocused* image with a depth of field controlled by the area of the pixels over which we integrate. Rays which truly originated from the same point in space at the target depth will then be projected to the same pixel in the refocused image and thus will form a sharp image at that point. Images can be refocused in order to perform object segmentation, detection, localization, manipulation, and 3D reconstruction.

The full light field in a volume describes the direction of all light traveling through all (x, y, z) points in the volume, for six dimensions in all. It is typical to instead consider only one z value per (x, y) pair to form a two dimensional manifold, and assume that rays only emanate from one side of the manifold. This is called the 4D light field, lumigraph, or photic field. The *light slab* is a common visualization of the 4D light field 3 and is similar in structure to light field photographs formed by microlens arrays.

To understand the refocusing process, recall that images are collected in a plane and recast down to a parallel plane at a specified focal depth or distance or height. Imagine that there is a projector array in space projecting each of the collected wrist images from the camera pose down onto a screen placed at the focal depth. The image on that screen is the refocused image. Another way of putting it is that the refocused image is seen as if the rays had all originated a that depth.

There is an analogy between this software camera and a physical camera. Choosing the depth at which we render is like controlling the focus of the lens, the resolution we render at is analogous to the zoom, and the area of pixels in the wrist camera image over which we integrate to form the refocused image is like the aperture. We say *angular aperture* to emphasize that it controls the angle of the rays that are collected, as viewing an object from the edge of the wrist image elicits a side view formed by oblique rays, while viewing the same object from the center of the wrist image elicits a top down view whose rays are normal to the refocused image plane. The wider the angular aperture is set, the faster objects go out of focus as they depart from the focal plane.

Rendering nearer to the camera causes oblique rays to come into focus, which naturally tilts the perspective out of the image plane, showing the sides of objects whose tops face the camera. Rendering far from the camera causes more direct or more normal rays to come into focus, making renders of objects more invariant to perspective. But we can choose the ray angles to consider at any height by setting the angular aperture accordingly. In one limiting case, we can form an image with totally normal rays which shows a wide scene as if it were viewed from infinity, or as if the camera had been directly above all of the objects at once, analogous to having a very wide lens. This dramatically simplifies object detection and pose estimation. We can approximate such an image with the marginal and max likelihood renders we describe next and

illustrate in 4. In the other direction, we can consider rays over a very wide range to dramatically narrow the depth of field, using defocus to eliminate objects from the image and simplify many tasks 5.

A. Refocused Image Model and Calibration

In order to accurately focus rays in software, we must be able to determine the global (x, y, z) coordinate of the origin point of a ray given the camera pose, an origin depth, and the pixel coordinate (i, j) of that ray. Our transformation from pixel coordinates to world coordinates depends upon knowing the projection of the end effector into the image plane at four known heights as well as having an accurate measurement of the physical distance between the camera and the end effector in the image plane 2.

Using the pixel-to-world transform we can create a sharp refocused image whose pixels have physical dimensions of less than a millimeter when rendered at Baxter scale depths. While creating the refocused image, we record not only the mean intensity of the pixel values that contribute each refocused pixel, but also the independent variance of the each of the three color channels. The variance of a time lapse image captures information about object boundaries, motion, and asymmetric lighting, and as such is a nice one-dimensional description of the angular content of the light field at a point. It is also a good measure of how well focused a refocused pixel is.

During end effector projection calibration, we fix a false, constant camera pose for reprojection to "smear" the image over time instead of correctively aligning it. This allows us to find fixed points in the camera image while the end effector undergoes motions, such as spinning about its axis to find its projection in the plane, or zooming towards and away from the plane to find the vanishing point for that motion in the camera image. We can estimate a gripper mask similarly by smearing the camera over the calibration target and masking points with low variance.

Our calibration is extremely straightforward and involves printing some copies of our calibration pattern, placing them haphazardly on a table in front of Baxter, and running a single program. It is repeatable, precise, and yields a mapping which, given a pixel in the camera and a target distance from the camera, produces the global (x,y,z) coordinate from which rays contributing to that pixel would originate. Accuracy is maintained even centimeters from the camera.

We start by estimating the vanishing point and gripper projections as described above. We then use bundle adjustment to iteratively refine a depth dependent magnification correction, which accounts for radial distortion. We iteratively render a scene with images taken at each height and optimize the camera model with gradient descent on an objective which measures the software camera's ability to focus rays consistently across space: that is, we minimize the variance of the refocused image. We do not use the vanishing point in the calibration but we want rays normal to the camera plane to be in the center of our aperture and the normal rays arrive at the pixel that casts perpendicularly into the image plane, i.e. the vanishing point.

B. Jointly Estimating RGB and Depth

Our depth reconstruction algorithm is only a local method but is at its heart similar to that in [4] in that it relies on defocus measurement.

Recall that we model the color distribution in each pixel or cell of a refocused image with an independent Gaussian on each color channel. A cell is more in focus when the sum of variances across its channels is small. By sweeping focus over a series of depths, we can assign to each cell the depth of the render in which its variance is the smallest. This is a maximum likelihood estimate of the depth at a point. We can induce a maximum likelihood refocused RGB image by assigning each pixel the color value it has in the image focused at its maximum likelihood height.

Similarly, for a given height we can use the Gaussian at a cell to evaluate the probability of the mean of that Gaussian. That value is the likelihood of making the ray observations at that cell under that Gaussian. Using this likelihood at each height to weight an average over depths, we can evaluate the expected value of the depth over a range to yield a marginal estimate of the depth at a point. Likewise we can weight the RGB estimates at each depth to form a marginal estimate of the RGB values at each refocused pixel, yielding a marginal refocused image.

Consider the depth maps in 4, the images of which were taken with the camera 38 cm from the table. The top of the mustard is 18 cm from the camera and very shiny, so this degree of local estimation is non-trivial. The maximum likelihood estimate was pooled over a 14×14 cell area to add a little bit of global information. The RGB maps are metrically calibrated images that give a top down view that appears in focus at every depth. Such a map greatly facilitates object detection, segmentation, and other image operations.

III. Future Work

Time lapse light field photography has exciting applications in motion analysis and light source estimation. Learning to tackle such problems in this new medium is bound to reveal some amusing results. We also want to explore geometric operations like object removal [5].

Our 3D reconstruction method is a nearly local method at the moment. Global priors on depth maps and 3D occupancy will improve the structural estimates [7]. The structure can in turn further improve the photographic techniques by modeling ray occlusion and reflectance. Processing light fields involves large data volumes and computation times, so sparse representations are needed to improve storage and computational efficiencies [2].

Modern methods in computer vision are data driven. As such there is a lot of motivation to reuse data. A standard form for data would facilitate sharing and reuse. The standard form should be as close to the original data, i.e. labeled camera images, as possible, while being immediately useful and accessible for as many applications as possible. What form should data take on to satisfy these conditions? Converting among image formats and managing camera profiles is difficult. Therefore we suggest that each robot be responsible for converting its images into metrically calibrated rays. These

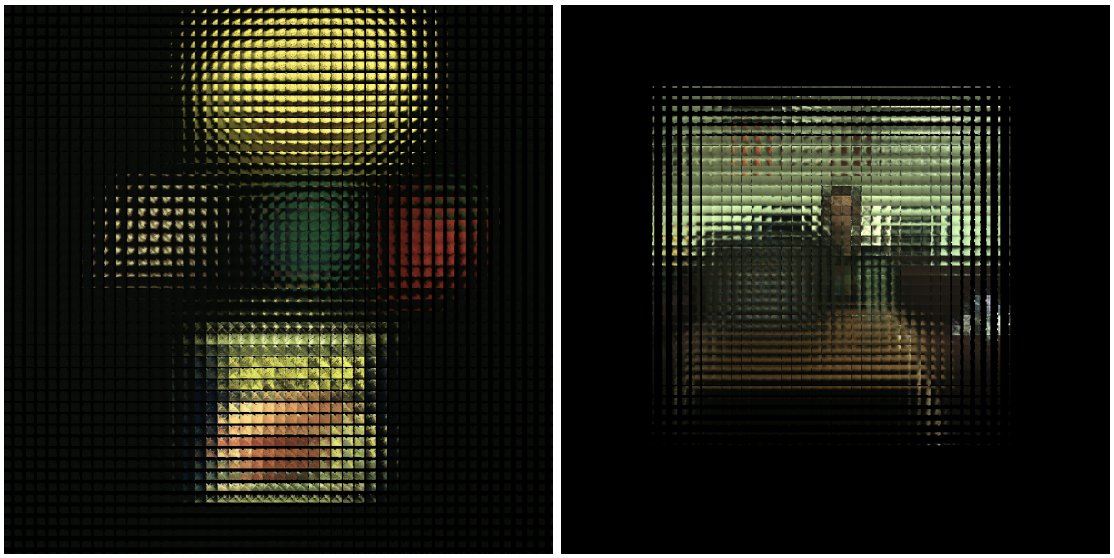

Fig. 3: Light slabs for the tabletop and room scenes. Each image is an array of sub images. Each sub images sorts the rays incident to the sub image according to angle of emission. The main lens is focused on the surface of the SPAM container and on the human subject, respectively. The reflections of the overhead lights are visible near the top of the SPAM label.

rays can be stored in a totally non-parametric format, shared, and re-rendered in 2D and 3D formats.

We have preliminary results with a graphical model over light fields which we can use to perform object detection, localization, segmentation, and grasping. It would be valuable to explore superresolution [1] techniques and see whether we can exceed the wrist camera resolution in refocused images.

The ability to calculate light slabs makes available to us the algorithms which use them as input. Employing Fourier optics should yield faster refocusing operations. We have used a one-dimensional lenticular array to view compatible light slabs in stereo, complete with multiple types of depth cue. We look forward to viewing light fields with other displays.

Our calibration model allows inference of camera parameters from pose annotated image frames. A system equipped to segment objects such as glass containers, windows, mirrors, floors, counters, and other reflective surfaces can infer camera properties given rays which interact with a target surface. Comparing the camera values estimated on the target surface against known free-space camera values reveals the optical properties of the target surface and allows them to be compensated for and exploited. It has already been observed that light field cameras can help robots perceive glossy surfaces [6].

IV. Conclusion

In this paper we have contributed a demonstration of a light field camera which can be implemented on a 7 DoF robotic arm with an eye in hand 2D RGB camera. We briefly described the algorithms necessary to calibrate the camera and demonstrated the use of the camera and the effects of the key parameters.

To our knowledge, before our work, Baxter and other 7 DoF arms were unable to collect and render light field data. Furthermore, the light field capturing abilities of Baxter in this paradigm are unique in scale, flexibility, and precision when compared to other modalities of light field collection. We hope that this work helps robots see better through light fields and helps researchers learn more about light fields through the use of robots. What makes this possible and accessible is our automatic and theoretically intuitive calibration process which is accurate, precise, and repeatable. Once the camera is calibrated over depth, straightforward algorithms based on ray casting produce consistent results. Without calibration, developed images will be out of focus (if recognizable) and not metrically interpretable.

The depth estimates and various rendering techniques we demonstrated are encouraging and suggest that passive light field sensing can make powerful contributions to object classification, localization, and manipulation. The system we implemented will be available in our next software release and demonstrates our capabilities in this domain.

References

[1] Tom E Bishop, Sara Zanetti, and Paolo Favaro. Light field superresolution. In *Computational Photography (ICCP), 2009 IEEE International Conference on*, pages 1–9. IEEE, 2009.

[2] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus H Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73–1, 2013.

[3] Ren Ng. *Digital light field photography*. PhD thesis, stanford university, 2006.

[4] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[5] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C Lawrence Zitnick, and Sing Bing Kang. Reconstructing occluded

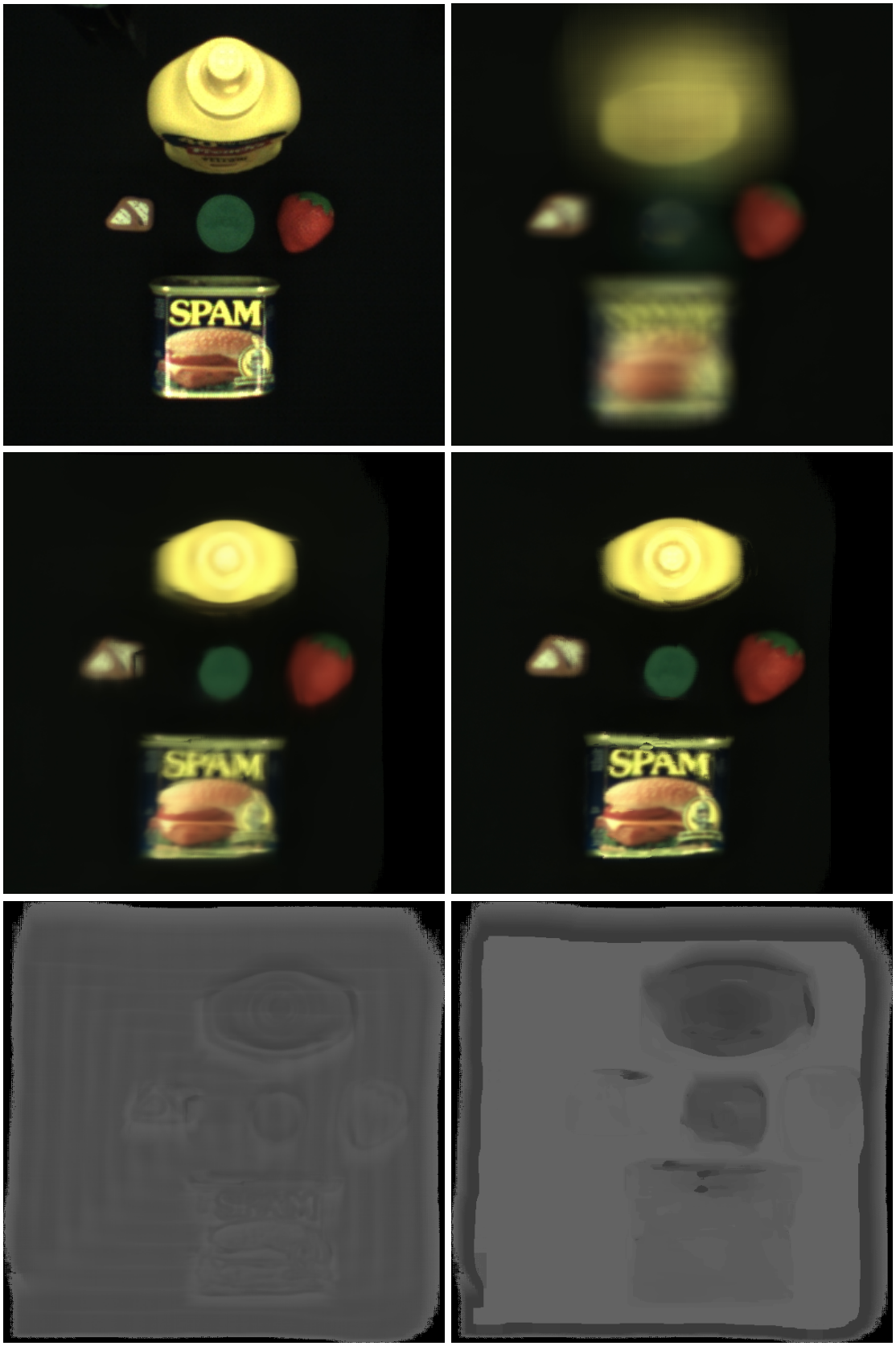

Fig. 4: A tabletop scene. Top Left: A single image from the wrist camera, showing perspective. Top Right: Refocused image converged at table height, showing defocus on tall objects. Middle Left and Right: Marginal and maximum likelihood RGB images, showing all objects in focus, specular reflection reduction, and perspective rectification. Bottom Left and Right: Depth estimates for marginal and maximum likelihood images.

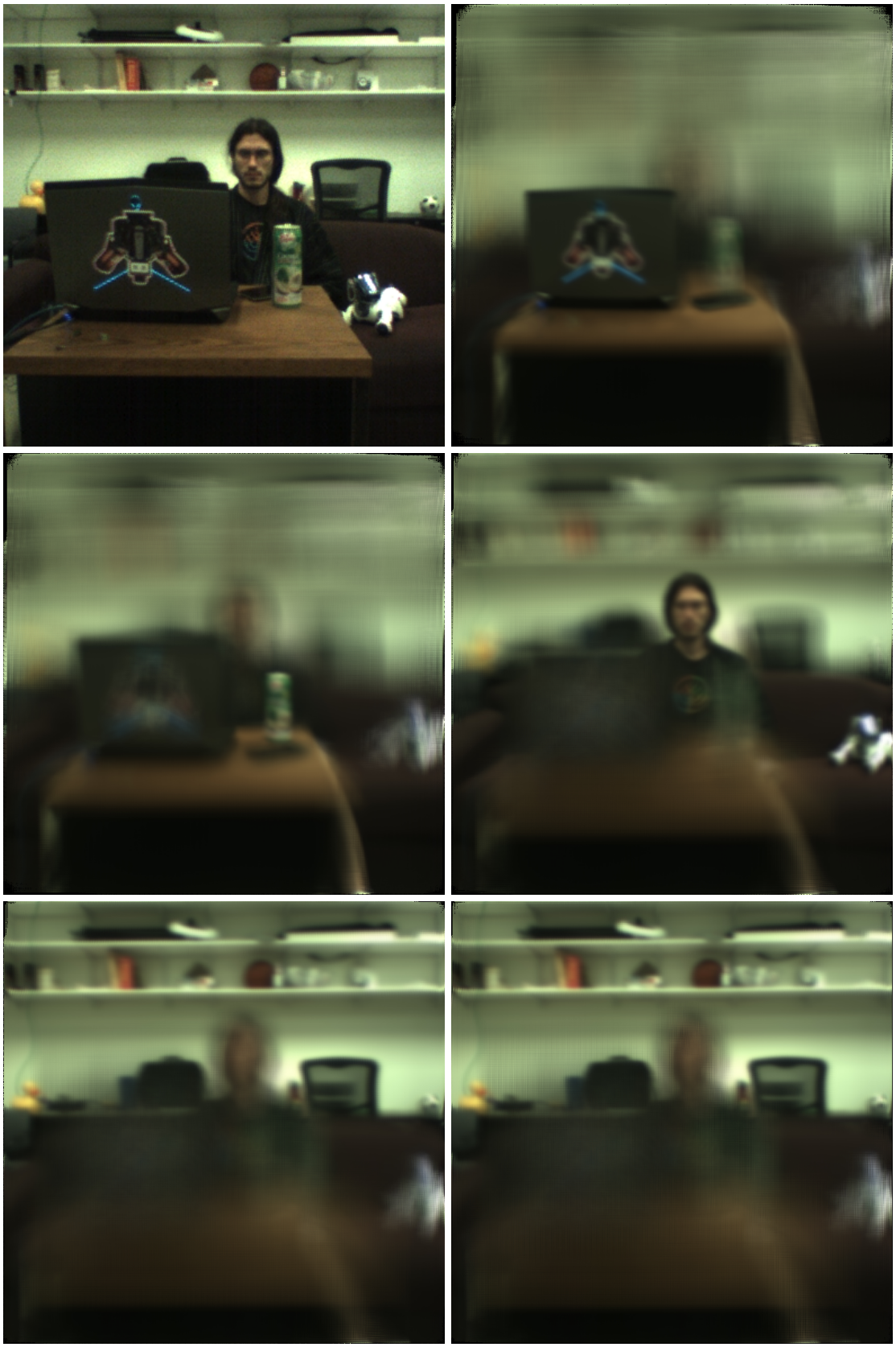

Fig. 5: A room scene. Top Left: A single image from the wrist camera. Remaining: Refocused photographs computed with approximately 4000 wrist images and focused at 0.91, 1.11, 1.86, 3.16, and 3.36 meters.

surfaces using synthetic apertures: Stereo, focus and robust measures. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2331–2338. IEEE, 2006.

[6] Christoph Walter, Felix Penzlin, Erik Schulenburg, and Norbert Elkmann. Enabling multi-purpose mobile manipulators: Localization of glossy objects using a light-field camera. In *Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on*, pages 1–8. IEEE, 2015.

[7] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):606–619, 2014.

[8] Jason C Yang, Matthew Everett, Chris Buehler, and Leonard McMillan. A real-time distributed light field camera. *Rendering Techniques*, 2002:77–86, 2002.

[9] Matthias Zobel. Object tracking and pose estimation using light-field object models. 2002.