External mask based depth and light field camera

Dikpal Reddy
NVIDIA Research
Santa Clara, CA
dikpalr@nvidia.com

Jiamin Bai
University of California, Berkeley
Berkeley, CA
bjiamin@eecs.berkeley.edu

Ravi Ramamoorthi
University of California, Berkeley
Berkeley, CA
ravir@eecs.berkeley.edu

Abstract

We present a method to convert a digital single-lens-reflex (DSLR) camera into a high resolution consumer depth and light field camera by affixing an external aperture mask to the main lens. Compared to the existing consumer depth and light field cameras, our camera is easy to construct with minimal additional costs and our design is camera and lens agnostic. The main advantage of our design is the ease of switching between an SLR camera and a native resolution depth/light field camera. Using an external mask is an important advantage over current light field camera designs since we do not need to modify the internals of the camera or the lens. Our camera sequentially acquires the angular components of the light field of a static scene by changing the location of the aperture in the mask. A consequence of our design is that the external aperture causes heavy vignetting in the acquired images. We calibrate the mask parameters and estimate multi-view scene depth under vignetting. In addition to depth, we show light field applications such as refocusing and defocus blur at the sensor resolution.

1. Introduction

Consumer depth cameras using coded light [21] and time-of-flight [10] have become extremely popular and have lead to an improved performance in computer vision applications. These active depth acquisition techniques, compared to passive techniques like stereo, provide robustness in low light and are accurate but are relatively expensive, consume significant power and have limited range. As an alternative, light field cameras are an emerging technology for capturing scene depth. The increasing importance of this passive depth acquisition technology is illustrated by the emergence of light field camera companies like Lytro [1], Raytrix [2] and Pelican Imaging [3]. Light field cameras are a generalization of stereo cameras and sample the angular variation in the incident light fields in addition to the usual spatial variation. It has been shown recently in [9] that light fields captured at wide baseline by SLR cameras can provide high quality depth information. Though light field cameras have so far been used primarily for consumer photography and scientific imaging, their potential as an enabling technology for computer vision is immense [5]. In addition to depth acquisition, the angular information captured by light field cameras could improve many computer vision problems such as segmentation, stabilization and material classification. However, the current light field cameras have poor spatial resolution [1] due to spatio-angular tradeoff or are significantly expensive [2]. In this paper we propose a method to convert a high spatial resolution DSLR camera, into a native resolution depth and light field camera.

Our new light field and depth camera is built with a DSLR camera and an external aperture mask cut from black paper as shown in Figure 1. The external mask affixed to the main camera lens acts as a modulator, allowing only light rays within a small solid angle. We capture angular information in the incident light field by changing the mask sequentially, allowing a different set of solid angles at each instance. Our light field camera is very easy to construct (making paper aperture masks takes less than an hour) with minimal marginal cost, provides the option of switching between a regular and light field camera and provides a high resolution depth. Furthermore, our design altogether avoids accessing the internals of the lens, redesigning the optics and redesigning the basic camera processing such as demosaicing and color processing [20].

Our design is motivated by the ideas of programmable aperture [12], external modulation [7] and mask-based modulation [16]. However, our design is significantly easier to implement, uses no additional optical elements and does not tinker with the lens system. This makes our design particularly attractive as a consumer depth camera since any existing DSLR camera can be converted into a light field camera with high resolution depth with just an additional aperture mask. The key insight of our paper is that it is not necessary to place the mask in the aperture plane of the lens to capture angular information of the light field. A similar mask affixed external to the lens can capture the angular variation in the light field as well. Our assumption is that the scene is significantly farther from the aperture plane than the mask from the aperture plane which is often true except in macro photography.

The placement of the mask in front of the lens, removed from the aperture plane, causes the captured images to be heavily vignetted as shown in Figure 1. We explain the vignetting mathematically in Section 3 and show that each

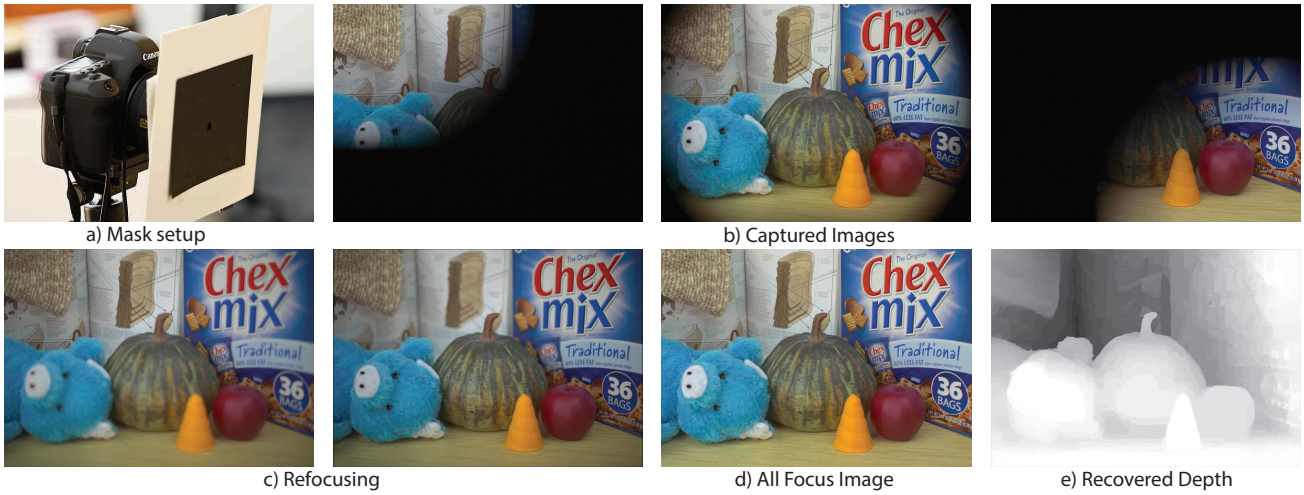

Figure 1. a) Our setup: DSLR camera and external paper mask b) Acquired vignetted images corresponding to different 5×5 mask sub-apertures c) Examples of synthetic refocusing (zoom into the PDF to see the out-of-focus areas in each image) d) An all-in-focus image at the central sub-aperture. Notice the absence of any vignetting. e) Estimated scene depth at the central sub-aperture.

image captures a sloped slice of the light field as shown in Figure 2. A programmable aperture camera [12] on the other hand captures a horizontal slice of the same light field.

We use a mask with a 5×5 sub-aperture array to acquire the light-field of a static scene. From the 5×5 views we estimate multi-view scene depth with occlusion reasoning. The depth estimation is particularly hard since each captured image is vignetted and has only a limited field-of-view (f.o.v) of the lens. Our problem is akin to the multi-view depth estimation problem described by Kang et al.[8] with the constraint that we need to rely on robust photoconsistency measures due to vignetting. The lack of a single image with full f.o.v. necessitates the depths estimated at each image to be fused together to create single depth image for the scene.

If the scene is nearly Lambertian, using the estimated depth and occlusions we can interpolate intermediate views between a captured 5×5 array of images to generate finer sampling of the light field as shown in Section 5. The captured vignetted images can then be transformed into non-vignetted, all-focus images through a simple resampling of the light field space. The finely sampled light field also allows us to achieve alias free digital refocusing.

In this paper we have presented traditional light field applications such as depth estimation, refocusing and all-focus images but the information provided by light fields is much richer and goes beyond imaging applications. We foresee light field data improving many computer vision applications such as segmentation, stabilization, material classification and recognition.

We note that currently our design is applicable only for static scenes since we cycle through 5×5 array of apertures. Further, our reconstruction technique is computationally expensive since the multi-view passive stereo requires significant disparity search and regularization. Nevertheless, we believe the advantages offered by our simple design, flexibility and little marginal costs make this approach exciting as a consumer depth camera. In summary, our contributions are:

1. A consumer depth and light field camera design which can be built easily and flexibly from a DSLR camera and an external paper mask with little marginal costs.

2. Estimating high resolution depth under vignetting and limited field-of-view enabling view interpolation and light field reparameterization.

2. Related Work

Depth cameras and reconstruction: Since light field cameras are a generalization of stereo cameras, they inherit the advantages and disadvantages of the passive depth estimation techniques compared to the active methods such as coded light imaging and ToF imaging. i.e. they involve little marginal cost, can work in bright scenes, are not power hungry and have large range but are computationally expensive, perform poorly in low light and have low depth sensitivity. But since our light field camera is built upon a high resolution DSLR it achieves the high native spatial resolution unlike most consumer depth cameras. Recently [9] showed that high quality depth can be recovered from finely sampled light fields but the datasets were acquired over a significantly wide baseline and required external camera calibration. Since our capture uses a single center-of-projection no external camera calibration is needed.

Our depth estimation procedure relies on constructing a disparity space image (DSI) by reparameterizing the light field (refocusing) and searching for the best focus or photoconsistency of pixels in the array images. The stereo correspondence chapter in Szeliski's book [15] provides an excellent overview of different techniques available for depth. During multi-view depth estimation, for each image view we warp the other views to this image and estimate the depth by using a varying spatio-temporal window as described in Kang et al. [8]. The occlusions are reasoned from

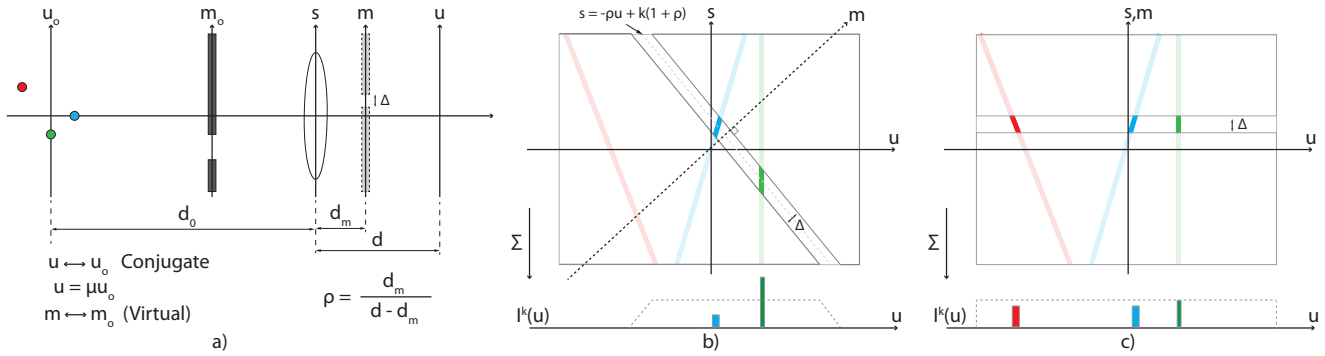

Figure 2. a) The scene with axis u_o and $u = \mu u_o$ in the conjugate plane where μ is the spatial magnification, the axis s in the aperture plane, the axis m_o in the external mask plane and the corresponding axis m in the virtual mask plane. The light field is parameterized as $L(u, s)$ and the colored scene points have corresponding colored support lines in b) and c). b) The modulation by mask with a sub-aperture k is shown by a sloped band with the slope $-\rho$ given by $\rho = \frac{d_m}{d-d_m}$. Since the mask sub-aperture band has a limited spatial extent, the image $I^k(u)$ at the sensor is vignetted and has a limited f.o.v. c) When the mask is in the aperture plane, $m = s$, the band is horizontal and there is no vignetting.

the depth estimates by checking for photoconsistency across views. For spatial consistency of depth and to fill the holes in textureless regions we use graph cut techniques based on MRFs [6]. Since the angular dimension is not finely sampled with our 5×5 sub-aperture mask, the epipolar plane image (EPI) based techniques for depth estimation [4],[17] cannot be applied to our data.

Comparison with existing light field designs: Our light field camera design is novel and builds on the theory and design principles laid out in the previous light field cameras. A good overview of the sampling of the plenoptic function can be found in the survey work by Wetzstein et al. [18] and Zhou et al. [22]. We use the terminology in Zhou et al. [22] to classify the previous light field capture methods into three classes: *sensor side modulation, aperture modulation and object side modulation*.

Sensor side modulation: The basic idea of sensor-side modulation is to project the angular information of the light field onto the spatial dimension. This is accomplished by using either lenslet arrays [4, 14, 13] or masks [16] in front of the sensor. These techniques usually leave high-frequency patterns making demoisaicing and color processing hard [20]. Furthermore, lenslets introduce optical aberrations. More importantly, these techniques require significant modifications to the hardware that offers no flexibility to switch between capturing light fields or regular images. *Our design fundamentally avoids any internal access and is easy to construct, flexible and requires no reinvention of camera processing.*

Aperture modulation: Modulation in the aperture plane [12] does not project the angular dimension of the light field on the sensor, preserving the full resolution of the spatial dimension. Instead angular resolution is gained by sacrificing temporal resolution. Levin et al. [11] proposed a coded aperture technique for depth estimation but it was not used for light field capture. These techniques require the lens body to be accessed to place the mask in the optical pathway, since the aperture plane in a regular lens system is inside the lens. Our design shows that the angular resolution can be sampled even by placing the mask external to the lens in front of the camera instead of the aperture plane [12]. We note that the vignetting encountered in Liang et al. [12] is primarily due to cosine falloff whereas the vignetting in our camera is a consequence of our design.

Object side modulation: External modulation offers flexibility and avoids reinventing camera processing. An example which avoids temporal tradeoff is by Georgiev et al. [7]. They use an external concave lens array with prisms to achieve spatio-angular tradeoff by packing the angular information contiguously. Nevertheless their system requires careful engineering of the external lens system and the additional relay lens can be bulky. The additional optical elements also change the effective focal length of the system and also introduce aberrations.

The closest design to our camera is the lensless two plane mask based camera by Zomet and Nayar [23]. Their design allows wider applications than light field capture but suffers from image quality due to lack of a lens [12]. On the other hand, our design introduces no additional optical elements and can be built from simple opaque paper with easy post capture calibration. Our external mask also allows the use of different aperture sizes and configurations without modifying the effective focal length.

Multiple cameras in an array can be used to capture light fields with a wide baseline (in addition to other applications) and was demonstrated by Stanford's camera array [19]. However, this system is expensive, not portable and requires careful synchronization and calibration.

Parameterization and calibration: We parameterize the light field with two planes at aperture and the sensor like in the previous mask based design [16], [12]. We show that an external mask is mathematically equivalent to a scaled and inverted internal mask close to the aperture plane. Each image is a sloped slice of the light field in the two plane parameterization and the angle is given by the ratio of the distance between the mask to aperture and sensor [16]. We calibrate the mask to determine the mask offset from the principal point and the aperture plane axes.

3. Mask Modulation and Calibration

We first present the basics of light fields and external mask modulation with a 2D light field. Consider the scene shown in Figure 2(a) where the camera is imaging the scene and a mask with sub-aperture is placed in front of the main lens. We parameterize the light field external to the camera as $L_o(u_o, s)$ and the light field internal to the camera as simply $L(u, s)$ where s is the axis at the aperture plane and u_o and u are axes in the conjugate object and sensor planes respectively. The distances d_0 and d of the planes u_0 and u from the aperture plane s are related by the thin lens equation $\frac{1}{d_0} + \frac{1}{d} = \frac{1}{f}$ where f is the effective focal length of the lens system. Since the sensor captures a magnified (and inverted) image of the light field, we have $u = \mu u_o$ where $\mu = -\frac{d}{d_0}$ is the spatial magnification. This means that the internal light field is a scaled and flipped version of the external light field. The image captured at the sensor is an integration of the light field over the aperture plane i.e. $I(u) = \int_s L(u, s)$.

The external mask is a 5×5 grid of sub-apertures attached to the lens. We sequentially acquire 25 images by opening each of these sub-apertures. Note that a coded aperture [12] acquisition would provide better noise properties but that is not the focus of this paper. Like the light field, the external mask axis m_o in the mask plane has a virtual flipped and scaled mask axis m in the virtual mask plane inside the camera. Hence, we consider only the light field inside the camera and investigate the effect of the internal mask sub-aperture on the light field. Let the distance from the aperture axis s to the mask axis m be d_m. We define the ratio $\rho = \frac{d_m}{d - d_m}$ relating the axis s, m and u as

$$s = -\rho u + m(1 + \rho). \tag{1}$$

The mask sub-aperture modulates the light field and the modulation is shown as a sloped band in the light field space with slope $-\rho$ in Figure 2(b). The modulated light field is given by

$$L^k(u, s) = \int_{m=(k-0.5)\Delta}^{(k+0.5)\Delta} L(u, s)\delta(s + \rho u - (1 + \rho)m)\, ds. \tag{2}$$

When the mask is in the aperture plane [12] as shown in Figure 2(c), $\rho = 0$ and $s = m$.

The image captured with the kth sub aperture open is $I^k(u) = \int_s L^k(u, s)$. Since the modulation band does not span the entire sensor range, the image $I^k(u)$ is vignetted and has a limited f.o.v. as illustrated in Figure 2. Note that as the f-number of the camera increases, the f.o.v. decreases since the modulation band has smaller range in s, thus restricting the range in u. As the sub-aperture k changes, the modulation band shifts in both u and s resulting in a shifted f.o.v. and a parallax shift.

Consider the three colored points in the scene at different depths in Figure 2(a). The support of the three points in the light field space is given by the sloped lines. The point in focus has support parallel to the s axis. This implies that the integration of the light field at the sensor induces no blur of the point. The point farther from the camera has a negative slope and the point closer to the camera has a positive slope resulting in defocus blur in the image. The slope of the lines give the distance of the point from the focal plane. Determining the depth of the points is nothing but estimating the slope of the lines in the light field space from the vignetted images $I^k(u)$ and is discussed in Section 4.

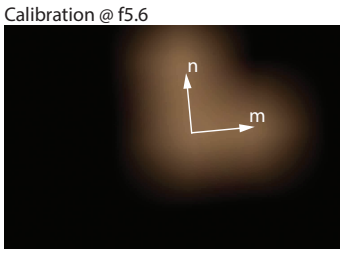

Figure 3. Overlayed images captured at $f5.6$ with the central sub-aperture of the 5×5 mask and sub-apertures above and next to it. We calibrate the offset, orientation and the shift in f.o.v. in pixels by localizing the blob centers.

In 4D, the light field is represented as $L(u, v, s, t)$ with the mask plane defined by axis m and n. For simplicity we choose the aperture axes s, t to align with the mask axes m, n which may not be aligned with the sensor axes u, v.

Mask calibration: We need to calibrate the masks so that we can determine the axis of the sub-apertures. In practice, a 2D mask center can be offset from the principal point and the mask axes m and n may not align with the image axes u and v. We capture calibration images of a diffuse white board to estimate the offset of the mask center by measuring the shift in f.o.v in pixels and estimate the mask axis rotation. We pick $f5.6$ and photograph the white screen with mask locations at the center and one each along the axes as shown in Figure 3 to localize the vignetted image centers accurately. The center image gives the offset from the principal point. The images along different axes gives the direction of the mask axis with respect to the sensor axis and the shift in pixels along the axis gives the f.o.v. shift in pixels. The offset, f.o.v. shift and axis rotation allow us to achieve physically accurate refocusing and depth estimation.

4. Depth Estimation

Depth estimation of the scene underlies many of the applications of light fields such as view interpolation, alias-free refocusing and multi-view image fusion. But estimating the depth from vignetted images acquired by an external mask is challenging. In this section we first briefly describe the previous approaches to light field depth estimation. We pose the problem of estimating scene depth as that of aligning the sensor axis with the light field gradient. Based on this formulation we present our approach to multi-view depth estimation and depth fusion under vignetting.

Previous light field depth estimation: Depth of scene points from 4D light field $L(u, v, s, t)$ can be determined by estimating the gradient of their support in the epipolar plane image (EPI) $L(u, s)$ and $L(v, t)$ as described in [4]. This approach was improved by Wanner et al. [17] by further reasoning about occlusions in the EPI using a structure ten-

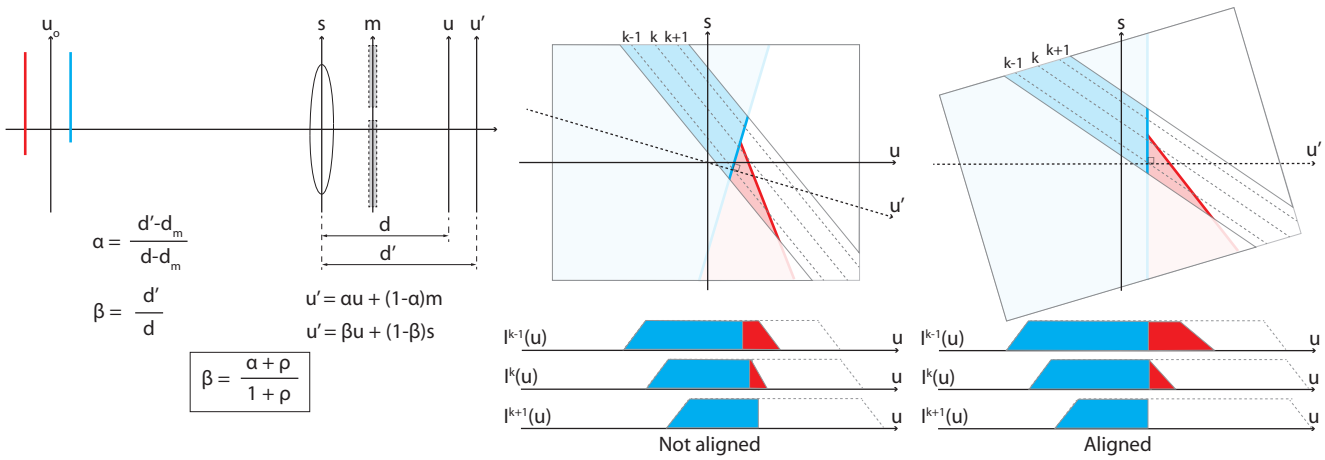

Figure 4. a) The scene with blue and red occluding lines. To estimate depth we reparameterize the sensor plane by a factor α to u'. The light field of the lines is shown as shaded overlapping blue and red regions in b) and c). b) The images at sub-apertures $k-1$, k and $k+1$ show the right edges of blue regions not aligned and progressively occluding the red region. c) After reparameterization by factor $\alpha > 1$, the new sensor plane u' is perpendicular to the blue border in $L(u', s)$. The right edges of the blue region are now aligned in the warped image. Depth estimation is simply searching for α which aligns the image intensities in the warped images.

sor framework. Both the methods were designed for finely sampled angular dimensions (s, t). When the angular samples are limited, the depth of the scene points is estimated through traditional stereo matching techniques. Liang et al. [12] perform multi-view depth estimation at each of the sub-aperture images $I^k(u)$ and occlusion is reasoned between every pair of neighboring views.

In this paper, we have access to $L(u, m)$ and not $L(u, s)$. Hence we estimate depth with $L(u, m)$ and explain our method in terms of $L(u, m)$ as well.

Depth estimation as light field alignment: In multi-view stereo, the depth of the scene at a reference view is estimated by warping the other views to that view and checking for photoconsistency of the scene points. The warping is a homography transformation corresponding to a virtual scene depth [8]. In light fields, the homography transformation corresponding to a virtual depth is simply a reparameterization to a virtual sensor plane u' as shown in Figure 4(b) and 4(c).

$$L_\alpha(u', m) = L(\alpha u + (1-\alpha)m, m). \tag{3}$$

The reparameterization factor α indicates the scene depth. $\alpha > 1$ corresponds to moving u' away from the aperture plane bringing the virtual depth closer and $\alpha < 1$ corresponds to moving the virtual depth farther. When the light field is reparameterized, the images $I^k(u)$ are warped to $I^k_\alpha(u')$ as illustrated in Figure 4. In Figure 4(b), the red and blue lines in the scene correspond to regions which intersect in the light field $L(u, s)$. As the sub-aperture k is changed, the blue and red regions in the image $I^k(u)$ move closer to each other with the blue region finally occluding the red region. The right edge of the blue region in image $I^k(u)$ is also shifting right. When the light field is reparameterized as shown in Figure 4(c), the right edges of the blue region across different views align. Since blue is closer to the camera, the factor $\alpha > 1$. This corresponds to moving the sensor axis u' away and rotating the light field anti-clockwise. In other words we search for α which makes the axis u' perpendicular to the blue line in $L_\alpha(u', m)$. Likewise, we search for α which makes the red line perpendicular to u'.

Multi-view stereo under vignetting: In our camera, the sub-aperture images $I^k(u)$ are vignetted and have limited f.o.v. of the scene. We describe the multi-view depth estimation under vignetting in Figure 5. Since the vignetted pixels are unreliable, we do not use those pixels for depth estimation and the depth is estimated at only the non-vignetted pixels. All masked images $I^k(u)$ are warped (an affine transformation) to $I^k_\alpha(u')$ by a factor α corresponding to a virtual depth. Note that the pixel u' in $I^{k_r}_\alpha(u')$ changes with changing α. To estimate the scene depth at a reference k_r, we apply another affine transformation to all views to ensure that $I^{k_r}_\alpha(u') = I^k(u)$ for every α. This additional warping is only for practical purposes and helps avoid the problem of tracking pixels u of I^{k_r} across different α.

Next, we construct a disparity space image (DSI) $D(u, \alpha, k_r)$ [15] to estimate the depth at the view k_r. DSI at pixel u of sub-aperture image k_r quantifies the photoconsistency of other views at that pixel when the views are warped by factor α. The DSI is constructed as

$$D(u, \alpha, k_r) = \sum_k f(I^{k_r}_\alpha(u), I^k_\alpha(u)) \tag{4}$$

The function $f()$ is a robust measure of photoconsistency which measures the difference in color as well as image gradients at the pixel across views and is given by

$$f(I_1(u), I_2(u)) = (1-\lambda)|I_1(u) - I_2(u)| + \lambda|\nabla_u I_1(u) - \nabla_u I_2(u)|. \tag{5}$$

In Equation (4), temporal selection [8] is done at each pixel to weed out poor view matches (such as vignetted regions of some views) and remove their contribution to DSI.

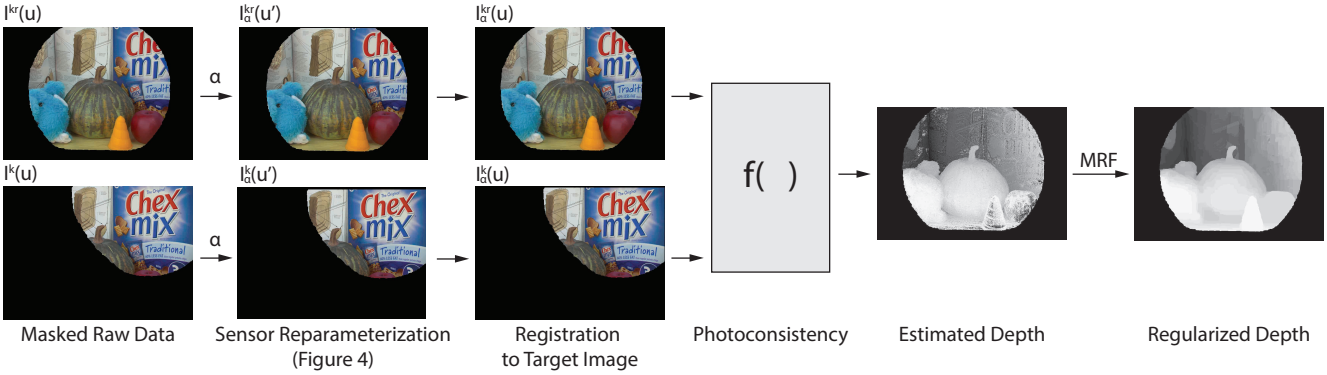

Figure 5. We estimate the scene depth at a reference view k_r by checking with other warped views for photoconsistency. In each image we only consider the non-vignetted regions of the image for depth estimation. The masked raw images are first warped by a factor α corresponding to virtual sensor position u'. These images are further affine transformed to ensure that the pixel $u' = u$ in I^{k_r} for every α. Then the image view k is compared with k_r for photoconsistency. The resulting depth $d^{k_r}(u)$ is estimated only at the non-vignetted pixels and is regularized with an MRF to fill holes.

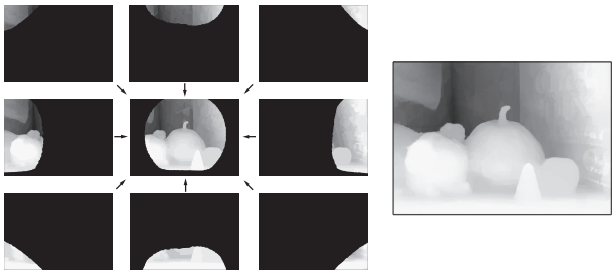

Figure 6. To generate the full depth at a reference view we warp the depth at other views to the reference view and perform visibility reasoning to handle occlusions.

The depth $d^{k_r}(u)$ at the reference view k_r is then given by

$$d^{k_r}(u) = \underset{\alpha}{\operatorname{argmin}}\, D(u, \alpha, k_r). \tag{6}$$

We then employ a standard MRF based depth estimation [15] to ensure spatial consistency and to fill holes in the smooth regions. The unary potential at pixel u is given by

$$E(u) = D(u, \alpha_u). \tag{7}$$

and the smoothness constraints between neighboring pixels u_1 and u_2 is

$$E(u_1, u_2) = \frac{1}{|I(u_1) - I(u_2)|} \max(|\alpha_{u_1} - \alpha_{u_2}|, \sigma) \tag{8}$$

Multi-view depth fusion: The depth $d^{k_r}(u)$ estimated at reference k_r is limited to non-vignetted pixels and corresponds to only a fraction of the f.o.v. of the scene. But different views k have different f.o.v. regions. Hence we use the depth of the scene points from other views to complete the depth information at k_r. The procedure shown in Figure 6 warps the depth $d^k(u)$ at pixel u to the view k_r by an affine transformation corresponding to the virtual scene depth $d^k(u)$. Note that some scene points occluded in view k_r will be visible in other views. This causes conflict in the depth estimates in the occluding regions when other views are warped to k_r. We resolve the conflict in such regions by performing visibility reasoning i.e. we simply take a minimum of all warped depths. The combined depths from different views at the central view is shown in Figure 6.

5. Applications

High spatial resolution depth and light fields are a rich source of information about the plenoptic function and potentially useful for many computer vision applications such as segmentation, stabilization and recognition [5]. In this paper we restrict our focus to light field imaging applications [1] and hope the emergence of light field cameras will spur research in their use in computer vision applications as well. The estimation of dense scene depth at the sensor resolution allows us to implement the standard light field applications [1] despite the lower angular resolution of our captured light field. We use the depth of the scene points to fuse images from multiple views to achieve an all-in-focus image. Multi-view depth information also makes occlusion reasoning easy, enabling view interpolation between the sub-aperture views. The interpolated views allow us to overcome the aliasing in the angular dimension enabling alias-free refocusing.

All-in-focus images: Examples of all-in-focus images are shown in Figure 1 and Figure 9. Since each view is vignetted, the all-in-focus image is created by borrowing pixels from different views. Using the depth information at the source view, we determine the amount of warp needed to transform the source image to the reference view. But a naive warp of the images to the reference view will cause tearing artifacts. To prevent that, we reason about the pixels which will be occluded in the reference view. The depth information at both these views allows us to determine the occlusions and disocclusions through visibility reasoning. We use the estimated occlusion map along with the required warp to propagate pixel values to the reference view creat-

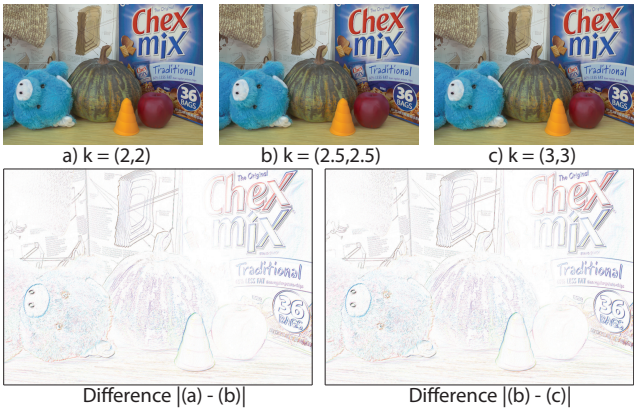

Figure 7. The view $k = (2.5, 2.5)$ has been interpolated from the views $k = (2, 2)$ and $k = (3, 3)$ and visualized through the difference images. White areas denote the least difference and colored areas have high difference. Notice that the difference is larger only at the farther and nearer ends of the scene where the motion is largest.

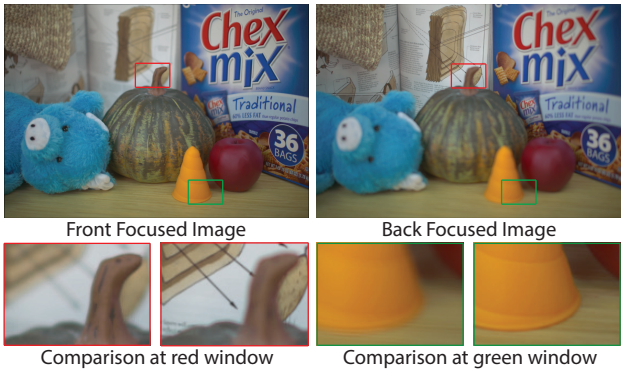

Figure 8. Digital refocusing done at two different scene depths. Notice that the refocused images are not aliased despite the low angular resolution since we integrate over the interpolated views.

ing a seamless all-in-focus image.

View Interpolation: Given the all-in-focus images between two neighboring views we interpolate the intermediate views. Figure 7 shows the interpolated view $k = (2.5, 2.5)$ between the sub-aperture views $k = (2, 2)$ and $k = (3, 3)$. The interpolated image has been visualized through difference images. Notice that the difference is larger at the farther and closer parts of the scene where the motion is largest. We note that the knowledge of depth allows alias-free interpolation compared to ghosting seen in a simple alpha blending. We discuss this more in the supplementary material and also provide videos of smooth transition in viewpoints along interpolated views.

Refocusing: We refocus at different depths of the scene by warping the light field to the virtual sensor position u' and then integrate over the synthetic aperture window W.

$$I_\alpha(u') = \sum_{m \in W} L_\alpha(u', m). \tag{9}$$

In Figure 8, we show the refocusing at two different scene depths. Notice that the scene has no aliasing despite limited angular resolution of our camera since we integrate over the interpolated views.

6. Conclusions

We presented a novel consumer depth and light field camera built with a DSLR camera and an external mask. The key feature of our design is the ability to convert any camera into a light field camera at will and extract high resolution depth with minimal marginal costs. We hope that this design will be a starting point for further investigation into simple, easy-to-build consumer depth and light field capture devices which can be used for solving a wide range of computer vision problems. The sampling of the angular resolution of the light field, in addition to high resolution depth, also provides additional information to improve the quality of vision applications such as segmentation, tracking and classification and we hope to explore this in future work. In this paper we demonstrated that our design allows acquisition of quality depth information even with a small baseline of the lens aperture, enabling imaging applications such as refocusing and multi-view all-in-focus images.

Our method currently captures the light field of a static scene at full sensor resolution. Since we capture the different sub-apertures sequentially, we tradeoff temporal resolution to gain angular resolution. We adopted a multi-view stereo reconstruction framework for depth estimation due to limited f.o.v and small angular resolution of the acquired light field. This makes our depth estimation computationally expensive and hence our reconstruction is not real-time. The design of a sequence of external mask patterns which makes the acquisition fast and exploration of fast multi-view depth algorithms are an avenue for future work.

Acknowledgement: D. Reddy and R. Ramamoorthi were supported by ONR PECASE grant N00014-09-1-0741. J. Bai was supported by A*STAR NSS PhD fellowship. We also acknowledge support and funding from Samsung and Nokia.

References

[1] `https://www.lytro.com/`.
[2] `http://www.raytrix.de/`.
[3] `http://www.pelicanimaging.com/`.
[4] E. Adelson and J. Wang. Single lens stereo with a plenoptic camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1992.
[5] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 91(1):3–20, 1991.
[6] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2001.
[7] T. Georgiev, C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala. Spatio-angular resolution tradeoffs in integral photography. In *Eurographics Symposium on Rendering*, 2006.
[8] S. B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *International Journal of Computer Vision*, 2004.

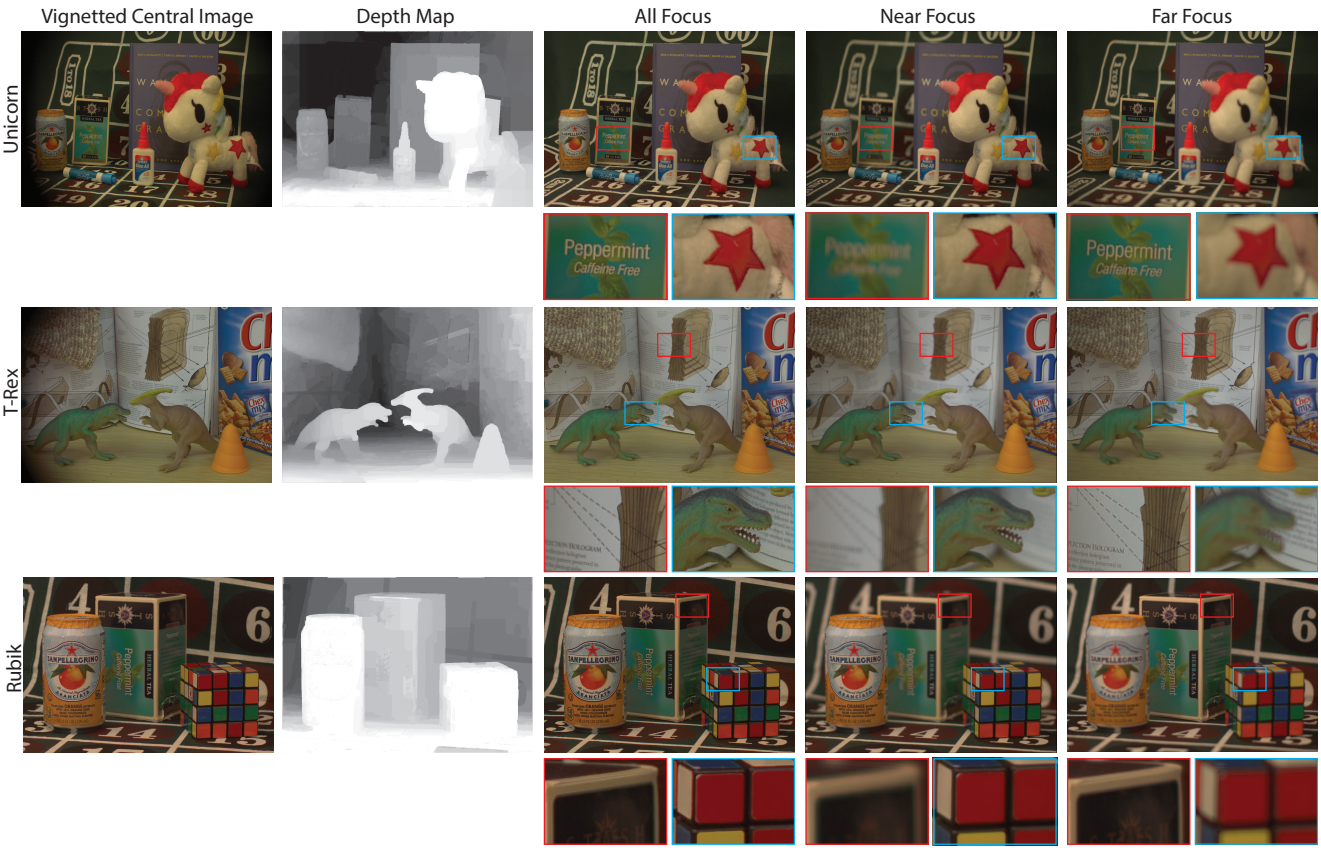

Figure 9. In the left column we show the images captured by the central sub-aperture of our mask. In the second column we show the depth estimated at the central sub-aperture through depth fusion. Note that the depth edges are sharp and the depth in the curved regions is gracefully changing. In the next column we show the all-in-focus image. The image has no vignetting and the non-vignetted regions are as sharp. In the next two columns we show digital refocusing at the front and back of the scene. Note that despite the limited angular resolution, the refocusing is alias-free.

[9] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 32(4):73:1–73:12, 2013.

[10] S. Lee, O. Choi, and R. Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer, 2013.

[11] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.*, 2007.

[12] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen. Programmable aperture photography: multiplexed light field acquisition. In *ACM Trans. Graph.*, 2008.

[13] A. Lumsdaine and T. Georgiev. The focused plenoptic camera. In *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009.

[14] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2005.

[15] R. Szeliski. *Computer vision: algorithms and applications*. Springer, 2010.

[16] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 2007.

[17] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D lightfields. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[18] G. Wetzstein, I. Ihrke, D. Lanman, and W. Heidrich. Computational plenoptic imaging. In *Computer Graphics Forum*, 2011.

[19] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*

[20] Z. Yu, J. Yu, A. Lumsdaine, and T. Georgiev. An analysis of color demosaicing in plenoptic cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[21] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012.

[22] C. Zhou and S. Nayar. Computational cameras: Convergence of optics and processing. *Image Processing, IEEE Transactions on*, 2011.

[23] A. Zomet and S. K. Nayar. Lensless imaging with a controllable aperture. In *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on*, volume 1, pages 339–346. IEEE, 2006.