Depth from Combining Defocus and Correspondence Using Light-Field Cameras

Michael W. Tao[1], Sunil Hadap[2], Jitendra Malik[1], and Ravi Ramamoorthi[1]

[1]University of California, Berkeley [2]Adobe

Abstract

Light-field cameras have recently become available to the consumer market. An array of micro-lenses captures enough information that one can refocus images after acquisition, as well as shift one's viewpoint within the sub-apertures of the main lens, effectively obtaining multiple views. Thus, depth cues from ***both*** *defocus and correspondence are available simultaneously in a single capture. Previously, defocus could be achieved only through multiple image exposures focused at different depths, while correspondence cues needed multiple exposures at different viewpoints or multiple cameras; moreover, both cues could not easily be obtained together.*

In this paper, we present a novel simple and principled algorithm that computes dense depth estimation by combining both defocus and correspondence depth cues. We analyze the x-u 2D epipolar image (EPI), where by convention we assume the spatial x coordinate is horizontal and the angular u coordinate is vertical (our final algorithm uses the full 4D EPI). We show that defocus depth cues are obtained by computing the ***horizontal*** *(spatial) variance after vertical (angular) integration, and correspondence depth cues by computing the* ***vertical*** *(angular) variance. We then show how to combine the two cues into a high quality depth map, suitable for computer vision applications such as matting, full control of depth-of-field, and surface reconstruction.*

1. Introduction

Light-fields [6, 15] can be used to refocus images [21]. Light-field cameras also hold great promise for passive and general depth estimation and 3D reconstruction in computer vision. As noted by Adelson and Wang [1], a single exposure provides multiple viewpoints (sub-apertures on the lens). The recent commercial light-field cameras introduced by RayTrix [23] and Lytro [9] have led to renewed interest; both companies have demonstrated depth estimation and parallax in 3D. However, a light-field contains more information about depth than simply correspondence; since we can refocus *and* change our viewpoint locally, *both* defocus and correspondence cues are present in a single exposure. Previous works have not exploited both cues together.

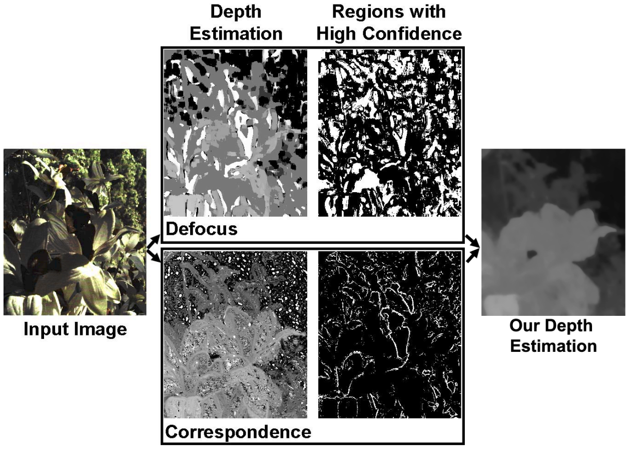

Figure 1. *Real World Result. With a Lytro camera light-field image input, defocus cues produce consistent but blurry depth estimates throughout the image. Correspondence cues produce sharp results but are inconsistent at noisy regions of the flower and repeating patterns from the background. By using regions from each cue with higher confidences (shown in the binary mask form), our algorithm produces high quality depth estimates by combining the two cues. Lighter pixels are registered as closer to the camera and darker as farther. This convention is used throughout this paper.*

We analyze the combined use of defocus and correspondence cues from light-fields to estimate depth (Fig. 1), and develop a simple algorithm as shown in Fig. 2. Defocus cues perform better in repeating textures and noise; correspondence is robust in bright points and features (Fig. 3). Our algorithm acquires, analyzes, and combines both cues to better estimate depth.

We exploit the epipolar image (EPI) extracted from the light-field data [3, 4]. The illustrations in the paper use a 2D slice of the EPI labeled as (x, u), where x is the spatial dimension (image scan-line) and u is the angular dimension (location on the lens aperture). Our final algorithm uses the full 4D EPI. We shear to perform refocusing as proposed by Ng et al. [21]. As shown in Fig. 2, for each shear value, our algorithm computes the *defocus cue response* by considering the spatial x (*horizontal*) variance, after integrating over the angular u (vertical) dimension. In contrast, we compute the *correspondence cue response* by considering the

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.89

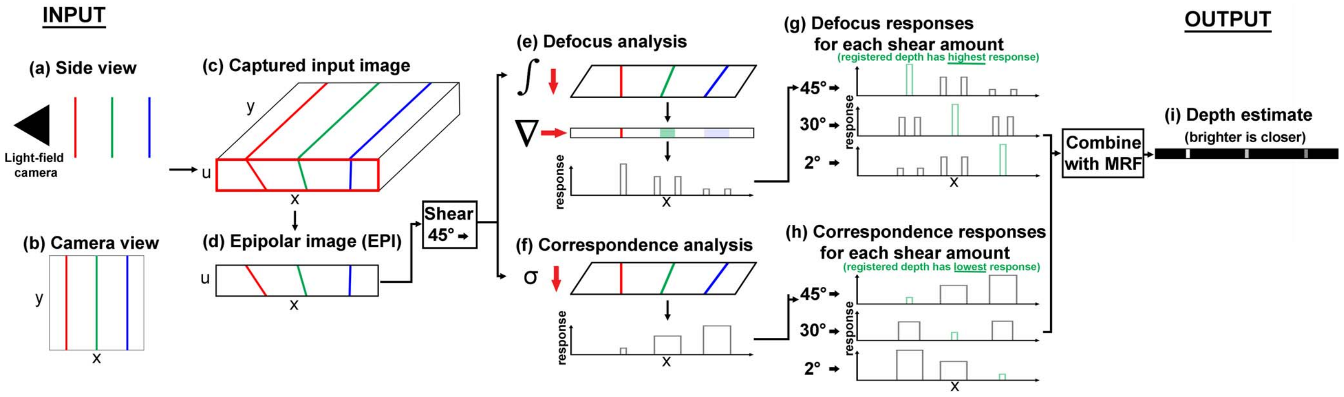

Figure 2. *Framework. This setup shows three different poles at different depths with a side view of (a) and camera view of (b). The light-field camera captures an image (c) with its epipolar image (EPI). By processing each row's EPI (d), we shear the EPI to perform refocusing. Our contribution lies in computing both defocus analysis (e), which integrates along angle u (vertically) and computes the spatial x (horizontal) gradient, and correspondence (f), which computes the angular u (vertical) variance. The response to each shear value is shown in (g) and (h). By combining the two cues using Markov random fields, the algorithm produces high quality depth estimation (i).*

	Implementation	Occlusions	Repeating Patterns	Bright/Dark Features	Noise
Defocus	+ no calibration needed - aperture-size dependent - patch size dependent	+ easily affected - more stable	+ contrast detection distinguishes	- contrast detection ambiguous	+ 2D blur kernel provides better support with noise
Correspondence	+ not dependent on DOF - noise from using pinhole - correspondence problem	+ less affected - unstable if affected	- correspondence ambiguity	+ correspondence not affected as much	- matching prone to noise - pinhole image noise

Figure 3. *Defocus and Correspondence Strengths and Weaknesses. Each cue has its benefits and limitations. Most previous works use one cue or another, as it is hard to acquire and combine both in the same framework. In our paper, we exploit the strengths of both cues.*

angular u (*vertical*) variance. The defocus response is computed through the Laplacian operator, where high response means the point is in focus. The correspondence response is the vertical standard deviation operator, where low response means the point has its optimal correspondence. With both local estimation cues, we compute a global depth estimate using MRFs [10] to produce our final result (Figs. 1, 7, 8, and 9).

We show that our algorithm works for multiple different light-field images captured with a Lytro consumer camera (Figs. 1, 8, and supplement). We also evaluated our data by comparing our results against user marked occlusion boundaries (Fig. 7). The high quality depth-maps provide essential information to enable vision applications such as masking and selection [5], modifying depth-of-field [13], and 3D reconstruction of surfaces [27] (Fig. 9).

Image datasets and code are available on our webpage[1]. To our knowledge, ours is the first publicly available method for estimating depth from Lytro light-field images, and will enable other researchers and the general public to quickly and easily acquire depth maps from real scenes. The images in this paper were captured from a single passive shot of the $400 consumer Lytro camera in different scenarios, such as high ISO, outdoors and indoors. Most other methods for depth acquisition are not as versatile or too expensive and difficult for ordinary users; even the Kinect [26] is an active sensor that does not work outdoors. Thus, we believe our paper takes a step towards democratizing creation of depth maps and 3D content for a range of real-world scenes.

2. Background

Estimating depth from defocus and correspondence has been studied extensively. Stereo algorithms usually use correspondence cues, but large baselines and limited angular resolutions prevent these algorithms from exploiting defocus cues. Schechner and Kiryati [25] and Vaish et al. [32] extensively discuss the advantages and disadvantages of each cue (Figure 3).

Depth from Defocus. Depth from defocus has been achieved either through using multiple image exposures or a complicated apparatus to capture the data in one exposure [34]. Defocus measures the optimal contrast within a patch, where occlusions may easily affect the outcome of the measure, but the patch-based variance measurements improve stability over these occlusion regions. However, out-of-focus regions, such as certain high frequency regions and bright lights, may yield higher contrast. The size of the analyzed patch determines the largest sensible defocus size. In many images, the defocus blur can exceed the patch size, causing ambiguities in defocus measurements. Our work not only can detect occlusion boundaries, we can provide dense stereo.

[1] Dataset and Source Code:
http://graphics.berkeley.edu/papers/Tao-DFC-2013-12/index.html

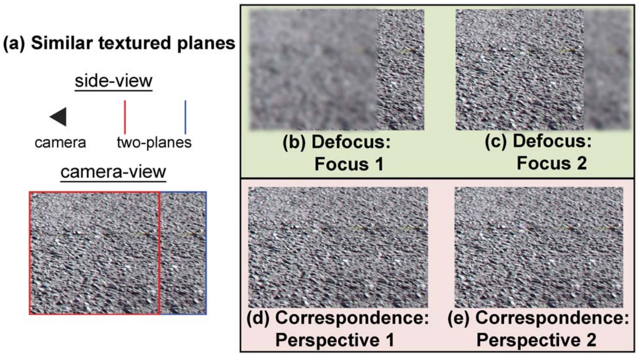
(a) Similar textured planes
side-view
camera
two-planes
camera-view
(b) Defocus: Focus 1
(c) Defocus: Focus 2
(d) Correspondence: Perspective 1
(e) Correspondence: Perspective 2

Figure 4. *Defocus Advantages at Repeating Patterns. In this scene with two planes (a), defocus cues, visually, give less depth ambiguity for the two planes at different depths (b) and (c). Correspondence cues from two different perspective pinhole images are hard to distinguish (d) and (e).*

Depth from Correspondences. Extensive work has been done in estimating depth using stereo correspondence, as the cue alleviates some of the limitations of defocus [20, 24]. Large stereo displacements cause correspondence errors because of limited patch search space. Matching ambiguity also occurs at repeating patterns (Fig. 4) and noisy regions. Occlusions can cause impossible correspondence. Optical flow can also be used for stereo to alleviate occlusion problems as the search space is both horizontal and vertical [8, 18], but the larger search space dimension may lead to more matching ambiguities and less accurate results. Multi-view stereo [16, 22] also alleviates the occlusion issues, but requires large baselines and multiple views to produce good results.

Combining Defocus and Correspondence. Combining both depth from defocus and correspondence has been shown to provide benefits of both image search reduction, yielding faster computation, and more accurate results [12, 29]. However, complicated algorithms and camera modifications or multiple image exposures are required. In our work, using light-field data allows us to reduce the image acquisition requirements. Vaish et al. [32] also propose using both stereo and defocus to compute a disparity map designed to reconstruct occluders, specifically for camera arrays. Our paper shows how we can exploit light-field data to not only estimate occlusion boundaries but also estimate depth by exploiting the two cues in a simple and principled algorithm.

Depth from Modified Cameras. To achieve high quality depth and reduce algorithmic complexity, modifying conventional camera systems such as adding a mask to the aperture has been effective [14, 17]. The methods require a single or multiple masks to achieve depth estimation. The general limitation of these methods is that they require modification of the lens system of the camera, and masks reduce incoming light to the sensor.

Depth from Light-field Cameras. There has not been much published work on depth estimation from light-field cameras. Perwass and Wietzke [23] propose correspondence techniques to estimate depth, while others [1, 15] have proposed using contrast measurements. Kim et al. and Wanner et al. [11, 33] propose using global label consistency and slope analysis to estimate depth. Their local estimation of depth uses only a 2D EPI to compute local depth estimates, while ours uses the full 4D EPI. Because the confidence and depth measure rely on ratios of tensor structure components, their result is vulnerable to noise and fails at very dark and bright image features. Our work considers both correspondence and defocus cues from the complete 4D information, achieving better results in natural images (Fig. 7, 8).

3. Theory and Algorithm

Our algorithm (shown in Fig. 2) comprises of three stages as shown in Algorithm 1. The first stage (lines 3-7) is to shear the EPI and compute both defocus and correspondence depth cue responses (Fig. 2e,f). The second stage (lines 8-10) is to find the optimal depth and confidence of the responses (Fig. 2g,h). The third stage (line 11) is to combine both cues in a MRF global optimization process (Fig. 2i). α represents the shear value.

For easier conceptual understanding, we use the 2D EPI in this section, considering a scan-line in the image, and angular variation u, i.e. an (x-u) EPI where x represents the spatial domain and u represents the angular domain as shown in Fig. 2. Ng et al. [21] explain how shearing the EPI can achieve refocusing. For a 2D EPI, we remap the EPI input as follows,

$$L_\alpha(x,u) = L_0(x + u(1 - \frac{1}{\alpha}), u) \tag{1}$$

Algorithm 1 Depth from Defocus and Correspondence

```
 1: procedure DEPTH($L_0$)
 2:     initialize $D_\alpha, C_\alpha$
                        ▷ For each shear, compute depth response
 3:     for ($\alpha = \alpha_{min}; \alpha <= \alpha_{max}; \alpha += \alpha_{step}$) do
 4:         $L_\alpha$ = shear($L_0, \alpha$)
 5:         $D_\alpha$ = defo($L_\alpha$)                ▷ Defocus response
 6:         $C_\alpha$ = corr($L_\alpha$)      ▷ Correspondence response
 7:     end for
                        ▷ For each pixel, compute response optimum
 8:     $\alpha^\star_D$ = argmax($D_\alpha$)
 9:     $\alpha^\star_C$ = argmin($C_\alpha$)
10:     $\{D_{\text{conf}}, C_{\text{conf}}\}$ = conf($\{D_\alpha, C_\alpha\}$)
                        ▷ Global operation to combine cues
11:     Depth = MRF($\alpha^\star_D, \alpha^\star_C, D_{\text{conf}}, C_{\text{conf}}$)
12:     return Depth
13: end procedure
```

L_0 denotes the input EPI and L_α denotes the sheared EPI by a value of α. The extended 4D form is in Eqn. 8.

3.1. Defocus

Light-field cameras capture enough angular resolution to perform refocusing, allowing us to exploit the defocus cue for depth estimation. We will use a contrast-based measure to find the optimal α with the highest contrast at each pixel. The first step is to take the sheared EPI and integrate across the angular u dimension (vertical columns),

$$\bar{L}_\alpha(x) = \frac{1}{N_u} \sum_{u'} L_\alpha(x, u') \tag{2}$$

where N_u denotes the number of angular pixels (u). $\bar{L}_\alpha(x)$ is simply the refocused image for the shear value alpha. Finally, we compute the defocus response by using a measure:

$$D_\alpha(x) = \frac{1}{|W_D|} \sum_{x' \in W_D} |\Delta_x \bar{L}_\alpha(x')| \tag{3}$$

where W_D is the window size around the current pixel (to improve robustness) and the Δ_x is the horizontal (spatial) Laplacian operator, using the full patch. For each pixel in the image, we now have a measured defocus contrast response for each α.

3.2. Correspondence

Light-field cameras capture enough angular information to render multiple pinhole images from different perspectives in one exposure. Because of the small-baseline, we can construct an EPI, which can be used for the correspondence measure [19]. Consider an EPI as shown in Fig. 2d. For a given shear α (Fig. 2f), we consider the angular (vertical) variance for a given spatial pixel.

$$\sigma_\alpha(x)^2 = \frac{1}{N_u} \sum_{u'} (L_\alpha(x, u') - \bar{L}_\alpha(x))^2 \tag{4}$$

For each pixel in x, instead of just computing the pixel variance, we need to compute the patch difference. We average the variances in a small patch for greater robustness,

$$\mathrm{C}_\alpha(x) = \frac{1}{|W_C|} \sum_{x' \in W_C} \sigma_\alpha(x') \tag{5}$$

where W_C is the window size around the current pixel to improve robustness. For each pixel in the image, we now have a measured correspondence response for each α.

3.3. Depth and Confidence Estimation

We seek to maximize spatial (horizontal) contrast for defocus and minimize angular (vertical) variances for correspondence across shears. We find the α value that maximizes the defocus measure and the α value that minimizes the correspondence measure.

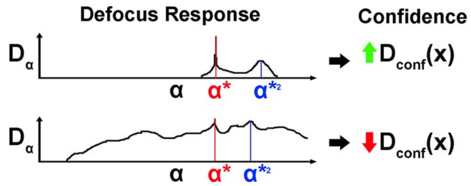

Figure 5. *Confidence Measure. From defocus Eqn. 3, we extract a response curve. Using the Peak Ratio confidence measure from Eqn. 7, the top curve has a higher confidence because the response ratio of $D_{\alpha^\star_D}(x)$ to $D_{\alpha^{\star 2}_D}(x)$ is higher than the bottom response curve. $\alpha^\star_D(x)$ represents the highest local maximum and $\alpha^{\star 2}_D(x)$ represents the second highest local maximum.*

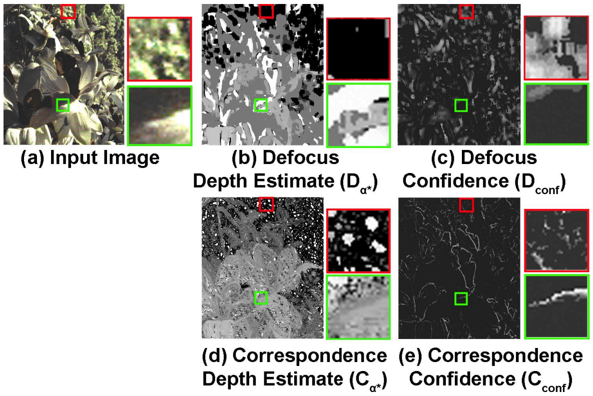
(a) Input Image
(b) Defocus Depth Estimate (D_{α^*})
(c) Defocus Confidence (D_{conf})
(d) Correspondence Depth Estimate (C_{α^*})
(e) Correspondence Confidence (C_{conf})

Figure 6. *Verifying Depth Estimation and Confidence. The red patch refers to a region with repeating patterns. Defocus performs better in showing the region is farther away from the camera (b) with higher confidence (c). Correspondence shows unstable results (d) with lower confidence (e). The green patch refers to a region with bright and dark regions. Defocus gives incorrect depth values (b) with lower confidence (c). Correspondence gives better results (d) with higher confidence at feature edges (e).*

$$\begin{aligned} \alpha^\star_D(x) &= \underset{\alpha}{\operatorname{argmax}}\ D_\alpha(x) \\ \alpha^\star_C(x) &= \underset{\alpha}{\operatorname{argmin}}\ C_\alpha(x) \end{aligned} \tag{6}$$

Defocus and correspondence cues might not agree on the optimal shear; we address this using our confidence measure and global step. To measure the confidence of $\alpha^\star_D(x)$ and $\alpha^\star_C(x)$, we use Peak Ratio as introduced by Hirschmüller et al. [7],

$$\begin{aligned} D_{\text{conf}}(x) &= D_{\alpha^\star_D}(x) / D_{\alpha^{\star 2}_D}(x) \\ C_{\text{conf}}(x) &= C_{\alpha^\star_C}(x) / C_{\alpha^{\star 2}_C}(x) \end{aligned} \tag{7}$$

where $\alpha^{\star 2}$ is the next local optimal value or the next largest peak or dip. The confidence is proportional to the ratio of the response estimate of $\alpha^\star$ to $\alpha^{\star 2}$. The measure produces higher confidence values when the maxima is higher than other values as shown in Fig. 5.

Discussion In Fig. 6, we observe two patches from the image input, depth estimate, and confidence. The patch shown in red represents a patch with repeating patterns and the patch shown in green represents bright features.

In the red patch, the depth estimation from correspondence is inconsistent, as we see noisy depth estimates. Our correspondence confidence measure in these regions is also low. This matches our observation in Fig. 4. In the green patch, the depth estimation from defocus is inconsistent with the image geometry. Our confidence measure also shows low confidence in the region.

Although we do not handle occlusions explicitly, given the confidence levels from both cues, our computation benefits from the defocus cues in handling occlusions better than correspondence cues (see occlusion boundaries in Fig. 7).

4. Implementation

In this section, we extend the 2D EPI theory to the complete 4D light-field data and use Markov Random Fields (MRF) to propagate our local measures globally. The input, L_0, is now replaced with the full 4D Light-field input data instead of the 2D EPI.

In our implementation, $\alpha_{min} = 0.2$, $\alpha_{max} = 2$, and $\alpha_{step} = 0.007$. Both W_D and W_C are local 9×9 windows.

Shear To perform shearing on the full 4D Data, we use the following equation from Ng et al. [21], which is analogous to Eqn. 1.

$$L_\alpha(x,y,u,v) = L_0(x+u(1-\frac{1}{\alpha}), y+v(1-\frac{1}{\alpha}), u, v) \quad (8)$$

MRF Propagation Since both defocus and correspondence require image structure to have non-ambiguous depth values, propagation of the depth estimation is needed. We used MRF propagation similar to the one proposed by Janoch et al. [10]. We concatenate the two estimations and confidences as follows,

$$\begin{aligned} \{Z_1^{\text{source}}, Z_2^{\text{source}}\} &= \{\alpha_C^\star, \alpha_D^\star\} \\ \{W_1^{\text{source}}, W_2^{\text{source}}\} &= \{C_{\text{conf}}, D_{\text{conf}}\} \end{aligned} \quad (9)$$

Source is used to denote the initial data term. We then use the following optimization to propagate the depth estimations.

$$\begin{aligned} \underset{Z}{\text{minimize}} \sum_{\text{source}} \lambda_{\text{source}} \sum_i W_i^{\text{source}} |Z_i - Z_i^{\text{source}}| \\ + \lambda_{\text{flat}} \sum_{(x,y)} \left(\left|\frac{\partial Z_i}{\partial x}\right|_{(x,y)} + \left|\frac{\partial Z_i}{\partial y}\right|_{(x,y)} \right) \\ + \lambda_{\text{smooth}} \sum_{(x,y)} |(\Delta Z_i)|_{(x,y)} \end{aligned} \quad (10)$$

λ_{source} controls the weight between defocus and correspondence. λ_{flat} controls the Laplacian constraint for flatness of the output depth estimation map. λ_{smooth} controls the second derivative kernel, which enforces overall smoothness.

Minimizing Eqn. 10 will give us $Z^\star$. $Z^\star$ may deviate from all source, flatness, and smoothness constraints. To improve the results, we find the error, δ, between $Z^\star$ and the constraints. We use an error weight matrix, E, which is constructed as follows,

$$\begin{aligned} \text{error}^2 &= \delta^2 + \epsilon_{\text{softness}}^2 \\ E &= 1/\text{error} \end{aligned} \quad (11)$$

where $\epsilon_{\text{softness}}$ provides a softening of the next iteration. We then solve the minimization function above with the weight, E. The iteration stops when the RMSE of the new $Z^\star$ compared to the previous $Z^\star$ is below the threshold (convergence fraction).

In our implementation, $\lambda_{\text{source}} = 1$ for both defocus and correspondence, $\lambda_{\text{flat}} = 2$, $\lambda_{\text{smooth}} = 2$, $\epsilon_{\text{softness}} = 1$, and convergence fraction $= 1$.

5. Results and Evaluation

We compare our work (defocus only, correspondence only, and global depth) against Sun et al. [30] and Wanner et al. [33]. Sun et al. is one of the top performers on the Middlebury's dataset [2]. Although it is not a light-field method, we use it to benchmark the best competing correspondence-only stereo algorithms, allowing us to evaluate the benefits of using both correspondence and defocus. We chose Sun et al. since it supports stereo without rectification, which is important for light-field data. Our supplementary material showcases more didactic comparisons and results.

Experiment For all images in the paper, we used the Lytro camera. While most visual effects are processed by Lytro's software, they do not make the light-field data accessible to users. We wrote our own light-field processing engine to take the RAW image from the sensor, and create a properly parameterized light-field, independent of the Lytro software. We use the acquired data to compute our epipolar and sub-aperture images to run our and competing algorithms. We tested the algorithms across images with multiple camera parameters, such as exposure, ISO, and focal length (Figs. 1, 7, 8, 9, and supplement).

Parameters. For Sun et al., we generated two sub-aperture images, spanning 66% horizontally of the main lens aperture. We use the authors' default settings to generate the stereo displacement maps. For Wanner et al., the local tensor structure default parameters are inner scale radius of 6 and σ of 0.8 and outer scale radius of 6 and ρ of

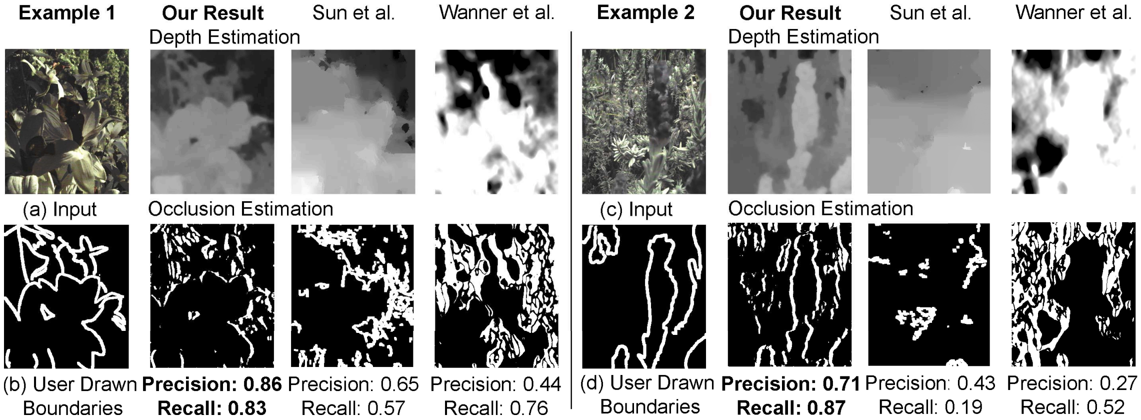

Figure 7. *Finding Occlusion Boundaries. With our dataset images (a,c), we manually marked the regions where occlusion boundaries occur (b,d). Our result performs better with a recall rate of occlusion boundaries with high accuracy, compared to Sun et al. [30] and Wanner et al. [33]. The left example (a) shows a difficult case where occlusion boundaries occur at multiple depths. The right (b) shoes another example where some occlusions are obvious to the users and some are not. Our occlusion boundaries are more accurate than other methods, with significantly higher precision as well as recall.*

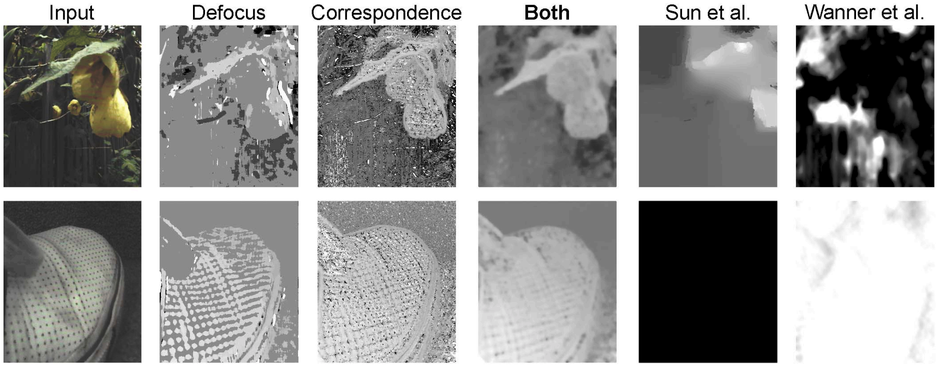

Figure 8. *Lytro Results Comparison. Defocus consistently shows better results at noisy regions and repeating patterns, while correspondence provides sharper results. By combining both cues, our method provides more consistent results in real world examples; whereas, Sun et al. show inconsistent edges and high frequency regions throw off Wanner et al. results. The flower (top) shows how we recover complicated shapes and scenes. The shoe (bottom) was captured at a high ISO with prominent color noise and banding. By combining both cues, our algorithm still produces reasonable results, while Sun et al. was not able to register correspondence and Wanner et al. fail in these high noise situations.*

0.8. For the global step, because code was not provided, we used our MRF to propagate the local depth measures.

Error metric. We first consider occlusion boundary detection, as shown in Fig. 7. We have a user mark the ground truth occlusion boundaries, an approach similar to one proposed by Sundberg et al. [31] and Stein and Hebert [28]. For each algorithm, we run a simple average of the absolute horizontal and vertical gradient values of the depth map. We mark the pixels as occlusions if the gradient value is greater than 0.008.

Results. As observed from Fig. 6, we see that defocus and correspondence have their advantages and disadvantages. Our final global result exploits the advantages and provides results that outperform Sun et al. and Wanner et al. visually and numerically. In Fig. 7, although the occlusion boundary recall rate of Sun et al. is high, the precision is low because of its over estimation of edges. Wanner et al. do not work well with the natural images generated by the Lytro camera because noise throws off both their depth and confidence measures. In Fig. 8, defocus is less affected by noise and repeating patterns while correspondence provides more edge information. Our combined results consistently perform better than Sun et al. and Wanner et al., providing better shape recovery as shown in the flower example (Fig. 8(top)) and high ISO example (Fig. 8(bottom)).

6. Applications

We show that our algorithm produces high quality depth maps that can be used for depth-of-field manipulation, matting and selection, and surface reconstruction.

Depth-of-Field Modifying depth-of-field has been a topic of significant interest with light-field data and cannot be achieved with current commercial software, which can only perform refocusing. Using our depth estimation, we simulate both lens aperture and refocusing (Fig. 9 Top). We use

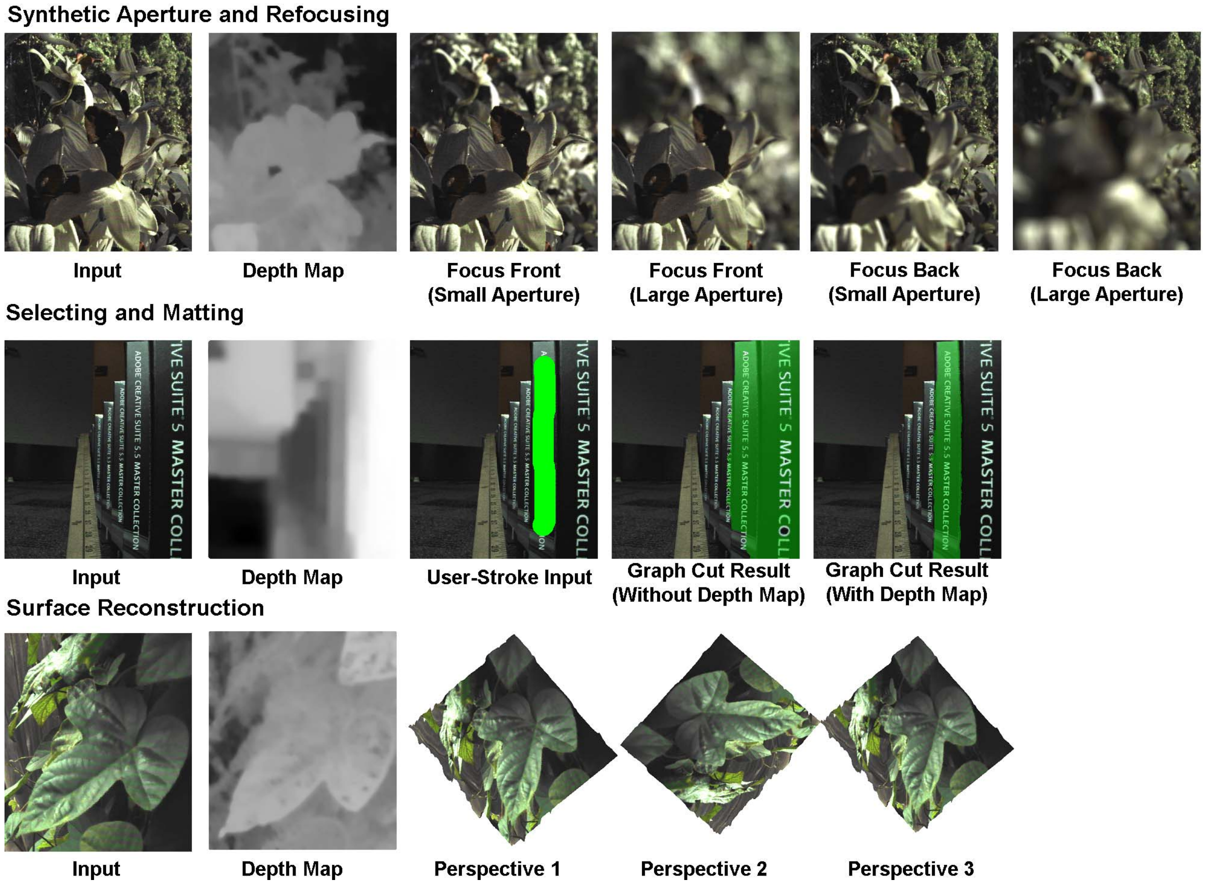
Synthetic Aperture and Refocusing
Input
Depth Map
Focus Front (Small Aperture)
Focus Front (Large Aperture)
Focus Back (Small Aperture)
Focus Back (Large Aperture)
Selecting and Matting
Input
Depth Map
User-Stroke Input
Graph Cut Result (Without Depth Map)
Graph Cut Result (With Depth Map)
Surface Reconstruction
Input
Depth Map
Perspective 1
Perspective 2
Perspective 3

Figure 9. *Applications. With our extracted depth maps, synthetic adjustment of both depth of field and refocusing is possible (top). For selection and matting, objects with similar color but different depths can be selected with depth information (middle). By using the depth map as the z-buffer, we can change perspective of the image, producing a 3D look (bottom).*

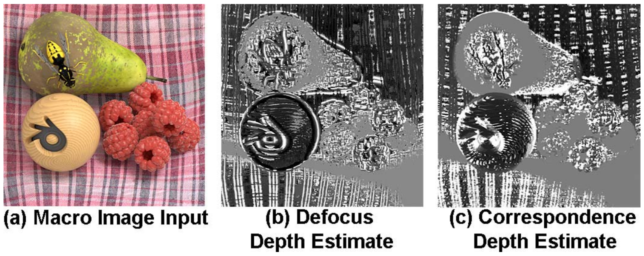

Figure 10. *Failure Case: Large Displacements. Macro images exhibit large displacements and defocusing (a). Both defocus (b) and correspondence (c) estimates fail. More sophisticated defocus or correspondence techniques are part of our future work.*

the depth map and a user input desired focus plane depth value. Regions with depth values farther from the input depth will have larger blurs. In the figure, we can see that the flowers and background foliage are blurred naturally.

Selection Current matting and selection graph-cut methods use only color information. Instead of using RGB, we use RGBD, where D is our depth estimation. With just a simple stroke, we can select out objects of similar colors, where previous color techniques fail (Fig. 9 Middle).

Surface Reconstruction One common use of depth-maps is to reconstruct surfaces, which goes beyond the limited parallax shift in Lytro's software. We remap the pixels with respect to our depth-map Z buffer into 3D space with mesh interpolation (Fig. 9 Bottom). This enables the users to explore surface shapes and bumps. Our results show that the perspective can be changed drastically and realistically.

7. Limitations and Discussion

Because our pipeline relies on shearing, objects that are too far from the main lens's focus plane will have incorrect depth estimations. For defocus, the out-of-focus blur becomes too large, creating ambiguity in the contrast measure. For correspondence, these areas show large stereo displacement. Since our method uses a fixed window size to compute these depth cues, ambiguities occur in our depth measurements (see Fig. 10). This paper has focused on the fundamentals of combining cues, using simple defocus and correspondence algorithms. In the future, more advanced defocus and correspondence algorithms may be used.

8. Conclusion

In this paper, we presented an algorithm that extracts, analyzes, and combines both defocus and correspondence depth cues. Using principled approaches, we show that defocus depth cues are obtained by computing the ***horizontal*** (spatial) variance after vertical (angular) integration of the epipolar image, and correspondence depth cues by computing the ***vertical*** (angular) variance. By exploiting the advantages of both cues, users can easily acquire high quality depth maps in a single shot capture. By releasing our code upon publication[1] (Page 2), we will enable researchers and lay users to easily acquire depth maps of real scenes, effectively making a point-and-click 3D acquisition system publicly available to anyone who can afford a consumer light-field camera. This in turn will democratize 3D content creation and motivate new 3D-enabled applications.

Acknowledgements This work was funded by ONR PECASE grant N00014-09-1-0741, and an NSF fellowship to M. Tao. We are grateful for the support from Nokia, Samsung, and Adobe.

References

[1] E. Adelson and J. Wang. Single lens stereo with a plenoptic camera. *PAMI*, 1992. 1, 3

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *ICCV*, 2007. 5

[3] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: an approach to determining structure from motion. *IJCV*, 1997. 1

[4] A. Criminisi, S. Kang, R. Swaminathan, R. Szeliski, and P. Anandan. Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *CVIU*, 2005. 1

[5] A. Criminisi, T. Sharp, and C. Rother. Geodesic image and video editing. *ACM Transactions on Graphics*, 2010. 2

[6] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *ACM SIGGRAPH*, 1996. 1

[7] H. Hirschmuller, P. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *IJCV*, 2002. 4

[8] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 1981. 3

[9] http://www.lytro.com. Lytro redefines photography with light field cameras. Press Release, June 2011. 1

[10] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell. A catergory-level 3D object dataset: putting the kinect to work. In *ICCV*, 2011. 2, 5

[11] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. In *SIGGRAPH*, 2013. 3

[12] W. Klarquist, W. Geisler, and A. Brovic. Maximum-likelihood depth-from-defocus for active vision. In *Inter. Conf. Intell. Robots and Systems*, 1995. 3

[13] T. J. Kosloff, M. W. Tao, and B. A. Barsky. Depth of field postprocessing for layered scenes using constant-time rectangle spreading. In *Graphics Interface*, 2009. 2

[14] A. Levin. Analyzing depth form coded aperture sets. In *ECCV*, 2010. 3

[15] M. Levoy and P. Hanrahan. Light field rendering. In *ACM SIGGRAPH*, 1996. 1, 3

[16] J. Li, E. Li, Y. Chen, L. Xu, and Y. Zhang. Bundled depth-map merging for multi-view stereo. In *CVPR*, 2010. 3

[17] C. Liang, T. Lin, B. Wong, C. Liu, and H. Chen. Programmable aperture photography: multiplexed light field acquisition. In *ACM SIGGRAPH*, 2008. 3

[18] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Imaging Understanding Workshop*, 1981. 3

[19] M. Matousek, T. Werner, and V. Hlavac. Accurate correspondences from epipolar plane images. In *Computer Vision Winter Workshop*, 2001. 4

[20] D. Min, J. Lu, and M. Do. Joint histogram based cost aggregation for stereo matching. *PAMI*, 2013. 3

[21] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photographhy with a hand-held plenoptic camera. *CSTR 2005-02*, 2005. 1, 3, 5

[22] M. Okutomi and T. Kanade. A multiple-baseline stereo. *PAMI*, 1993. 3

[23] C. Perwass and P. Wietzke. Single lens 3D-camera with extended depth-of-field. In *SPIE Elect. Imaging*, 2012. 1, 3

[24] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 3

[25] Y. Schechner and N. Kiryati. Depth from defocus vs. stereo: how different really are they? *IJCV*, 2000. 2

[26] J. Shotton, R. Girshick, F. A., T. Sharp, M. Cook, M. Finocchio, M. Richard, P. Kohli, A. Criminsi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *PAMI*, 2012. 2

[27] S. Sinha, D. Steedly, R. Szeliski, M. Agrawala, and M. Pollefeys. Interactive 3D architectural modeling from unordered photo collections. In *ACM SIGGRAPH Asia*, 2008. 2

[28] A. Stein and M. Hebert. Occlusion boundaries from motion: low-level detection and mid-level reasoning. *IJCV*, 2009. 6

[29] M. Subbarao, T. Yuan, and J. Tyan. Integration of defocus and focus analysis with stereo for 3D shape recovery. *SPIE Three Dimensional Imaging and Laser-Based Systems for Metrology and Inspection III*, 1998. 3

[30] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010. 5, 6

[31] P. Sundberg, J. Malik, M. Maire, P. Arbelaez, and T. Brox. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, 2011. 6

[32] V. Vaish, R. Szeliski, C. Zitnick, S. Kang, and M. Levoy. Reconstructing occluded surfaces using synthetic apertures: stereo, focus and robust measures. In *CVPR*, 2006. 2, 3

[33] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D light fields. In *CVPR*, 2012. 3, 5, 6

[34] M. Wantanabe and S. Nayar. Rational filters for passive depth from defocus. *IJCV*, 1998. 2