Depth Recovery from Light Field Using Focal Stack Symmetry

Haiting Lin[1] Can Chen[1] Sing Bing Kang[2] Jingyi Yu[1,3]
[1]University of Delaware [2]Microsoft Research [3]ShanghaiTech University
{haiting, canchen}@udel.edu sbkang@microsoft.com yu@eecis.udel.edu

Abstract

We describe a technique to recover depth from a light field (LF) using two proposed features of the LF focal stack. One feature is the property that non-occluding pixels exhibit symmetry along the focal depth dimension centered at the in-focus slice. The other is a data consistency measure based on analysis-by-synthesis, i.e., the difference between the synthesized focal stack given the hypothesized depth map and that from the LF. These terms are used in an iterative optimization framework to extract scene depth. Experimental results on real Lytro and Raytrix data demonstrate that our technique outperforms state-of-the-art solutions and is significantly more robust to noise and undersampling.

1. Introduction

Given the commercial availability of light field (LF) cameras such as the Lytro [1] and Raytrix [4], the use of LFs for scene capture and analysis is becoming more attactive. It has been shown that given the simultaneous multiple views, LFs enable improved image analysis, e.g., stereo reconstruction [20], refocusing [29], saliency detection [23], and scene classification [39].

In our work, we use commercially available LF cameras, namely Lytro and Raytrix. Note that these cameras have significantly lower sampling density (380×380) than most previous LF-based approaches (e.g., Stanford camera array, camera gantry [3]). Using the Lytro and Raytrix cameras presents challenges: while they provide high angular sampling, they are still spatially undersampled (causing aliasing in refocusing, as shown in Fig. 2), and SNR is low due to ultra small aperture ($14\mu m$ in Lytro, $20\mu m$ in Lytro Illum, and $50\mu m$ in Raytrix) and limited view extracting toolbox [2]. As shown in Figs. 1 and 2, previous approaches have issues with noise and refocusing.

In this paper, we propose a new depth from light field (DfLF) technique by exploring two new features of the focal stack. Our contributions are:

- Symmetry analysis on the focal stack. We show that the profile is symmetrically centered at the in-focus slice if the pixel corresponds to a non-occluding 3D point, even under noise and undersampling.
- New data consistency measure based on analysis-by-synthesis. Given a depth map hypothesis, we synthesize the focal stack. This is compared with that computed directly from the LF.
- Iterative optimization framework that incorporates the two features.

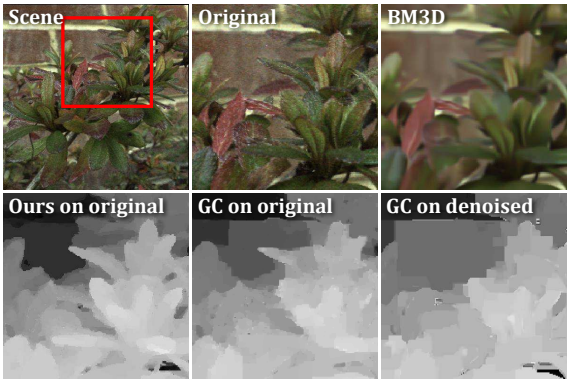


Figure 1. Noise handling (Lytro raw data extracted by [2]). Traditional stereo matching approaches either use a large smoothness term or first denoise the input [8, 41]. Both have the effect of blurring boundaries. Our technique is able to recover fine details without oversmoothing using the original noisy input.

Experimental results on real Lytro and Raytrix images demonstrate that our technique outperforms the state-of-the-art solutions and is significantly more robust to noise and undersampling.

2. Related Work

Our work is related to multi-view reconstruction and Depth-from-Focus; more detailed surveys can be found in [10, 11, 27, 28]. Here we only briefly discuss the most relevant ones to our approach.

Using LF data as input, Wanner and Goldlücke [37, 36, 39] optimize the direction field in 2D Epipolar Image (EPI)

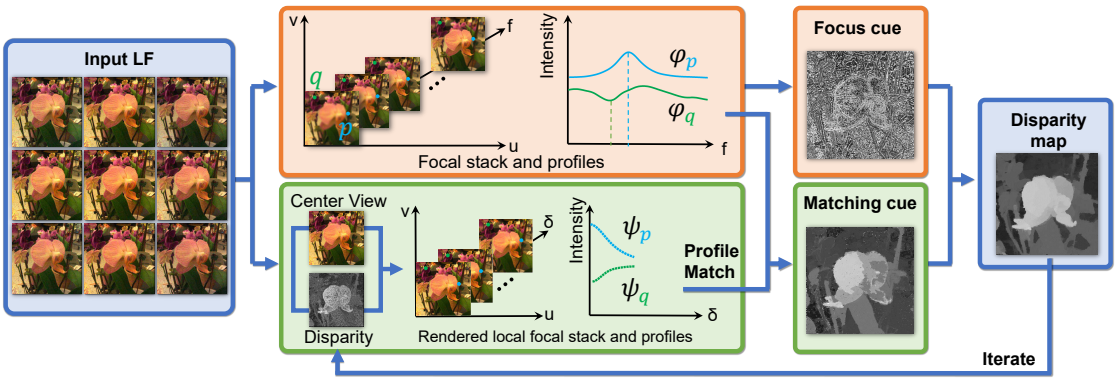


Figure 3. Pipeline of our method.

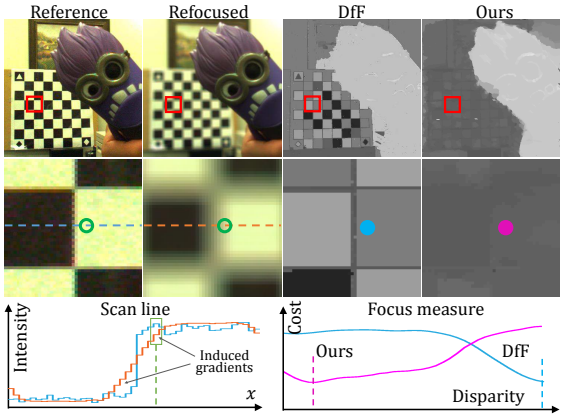


Figure 2. Aliasing in Synthesized LF Focal Stack. The scan lines show that refocusing possibly induces gradients (aliasing) to out-of-focus flat regions. Brute-force sharpness measure treats aliasing as edges and produces incorrect estimation (blue cost profile). Our method is able to obtain the correct background disparity (magenta cost profile). We add smoothness prior to both focus cues.

and directly map the directions to depths for stereo matching and object segmentation. However, this technique does not work well under heavy occlusion and significant image noise. Chen et al. [5] propose a bilateral consistency metric that separately handles occluding and non-occluding pixels. They combine both distance and color similarity to determine their likelihood of being occluded. The approach works robustly in the presence of heavy occlusion but requires low noise inputs. Heber *et al*. [18, 17] model depth from LF as a rank minimization problem and propose a global matching term to measure the warping cost of all other views to the center view. While their method is robust to reflections and specularities, it tends to produce smooth transition around edges. Kamal *et al*. [19] adopt similar rank minimization idea in local patches across views. They assume clean data of Lamebrain surfaces and use sparse error term in the modeling (accounts for mismatching due to occlusion, but is ineffective in resolving Gaussian noises). Kim *et al*. [21] are able to generate high quality results but need dense spatio-angular sampling.

Our DfLF is based on the principles of depth-from-defocus/focus (DfD/DfF). In DfD, several images focusing at different depths are captured at the fixed viewpoint [10, 11, 42] and focus variations are analyzed to infer scene depth. In a similar vein, DfF [35, 6, 24, 26, 27] estimates depth by the sharpness of a series of focus changing images. Depth is computed based on the most in-focus slice. To avoid issues caused by textureless regions, active illumination methods [25] are used. Hasinoff et al. [16] analyzed a series of images with varying focus and aperture to form a focus-aperture image (AFI), then apply AFI model fitting to recover scene depth.

There are studies [31, 34] that explore the strengths and weaknesses of DfD/DfF and stereo. They show that DfD/DfF is robust to noise while stereo can more reliably handle over/under saturated features. There are also techniques that combine stereo with DfD/DfF. In [22], a disparity defocus constraint is computed to form a defocus kernel map as a guidance for segmenting the in-focus regions. [22] models defocus and correspondence measure as data cost in the energy minimization frame work. Rajagopalan et al. [30] measure the consistency of the point spread function (PSF) ratios estimated from DfD for disparity computation under MRF framework. However, their solution is less effective with larger blur kernels.

Camera array systems where each camera has a different focus have been constructed, e.g., [13, 14, 15]. Here, edges of each view are estimated using DfD, with an edge consistence term being used in the multi-view stereo pipeline. Their system requires complex hardware settings. Tao et al. [33] combine DfD and depth from correspondence (DfC) by first estimating disparity on the EPI then applying MRF propagation. However, objects that are too far from the main lens focus plane and pixels near occlusion boundaries may result in large errors.

Our approach also combines focus analysis and stereo. Our work has two unique characteristics: (1) our focus measure is robust to image noise and aliasing due to un-

dersampling, and (2) we propose a novel data consistency measure based on analysis by synthesis. Fig. 3 shows the processing pipeline of our approach. In contrast to traditional DfD/DfF methods where sharpness is estimated on single focal stack images, we perform symmetry analysis on the entire focal stack. The matching cost is the difference between the hypothesized local focal stack (of each pixel) based on the hypothesized depth map and the LF version. These measures, together with a data consistency term, are optimized using MRF.

3. Color Symmetry in Focal Stack

A focal stack of a scene is a sequence of images captured with different focus settings; an LF can be used to produce a synthetic focal stack. We first describe our notations. The input LF is parameterized in two-plane parametrization (2PP), where camera plane st is at $z = 0$ and the image plane uv is at $z = 1$. In 2PP, a ray is represented as a vector (s, t, u, v) and we denote its radiance as $r_1(s, t, u, v)$, with the subscript indicating the depth of uv plane. The disparity output o is with respect to the center reference view I, where the ground truth disparity is denoted as d. We use $\varphi_p(f)$ to denote the color profile of pixel p in o at focal slice f in the focal stack (f is defined in disparity). We first analyze the local symmetry/asymmetry property of $\varphi_p(f)$ with respect to f and set out to derive a new focusness metric.

In our LF refocusing, a focal slice at disparity f is generated by integrating all the recorded rays corresponding to disparity f in sub-aperture views. Without loss of generality, we simplify our analysis by using 2D LF consisted of only s and u dimensions and only three sub-aperture views. Fig. 4 illustrates the ray integration process for three focal slices at disparities $(d - \delta)$, d, and $(d + \delta)$, where δ is a small disparity shift. In the analysis, the scene is planar and parallel to the image plane with ground truth disparity d. By similarity rule, we can compute its depth as $\frac{B}{d}$, where B is the baseline between two neighboring sub-aperture views.

Fig. 4(a) shows an example of a texture boundary pixel p with its coordinate p_u. We show that the color profile $\varphi_p(f)$ is locally symmetric around the ground truth disparity d. For the focal slice at $d + \delta$, the radiance from the left view is $r_1(-B, -B+p_u+(d+\delta))$. By reparameterizing it using the u-plane at $z = \frac{B}{d}$, we get the radiance as:

$$\begin{aligned} & r_1(-B, -B + p_u + (d + \delta)) \\ &= r_{\frac{B}{d}}(-B, -B + \frac{B}{d}(p_u + (d + \delta))), \\ &= r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta)). \end{aligned} \tag{1}$$

Similarly, the radiance from the right view is:

$$r_1(B, B + p_u - (d + \delta)) = r_{\frac{B}{d}}(B, \frac{B}{d}(p_u - \delta)). \tag{2}$$

The pixel value at p in the rendered focal slice is the result of integrating the radiance set $A_\delta = \{r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta)), r_{\frac{B}{d}}(0, \frac{B}{d}p_u), r_{\frac{B}{d}}(B, \frac{B}{d}(p_u - \delta)\}$.

We conduct a similar analysis for the focal slice at $d - \delta$. The radiance set will be $A_{-\delta} = \{r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u - \delta)), r_{\frac{B}{d}}(0, \frac{B}{d}p_u), r_{\frac{B}{d}}(B, \frac{B}{d}(p_u + \delta)\}$. Since the surface is exactly at depth $\frac{B}{d}$, according to Lambertion surface assumption, we have

$$\begin{aligned} r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta)) &= r_{\frac{B}{d}}(B, \frac{B}{d}(p_u + \delta)), \\ r_{\frac{B}{d}}(B, \frac{B}{d}(p_u - \delta)) &= r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u - \delta)), \end{aligned} \tag{3}$$

i.e. $A_\delta = A_{-\delta}$, which means $\varphi_p(d + \delta) = \varphi_p(d - \delta)$. The color profile $\varphi_p(f)$ is locally symmetric around the true surface depth d.

Fig. 4(b) and (c) show examples of occlusion boundary pixels. The ray integrations for pixels on the occluder (Fig. 4(b)) and on the occluded surface (Fig. 4(c)) are different. Unlike the texture boundary pixels, their color profiles do not have exact local symmetry property[1]. However, with more assumptions about the color variations on surfaces, we can show that the color profiles for those occlusion boundary pixels approximately exhibit local symmetry/asymmetry properties.

Notice that for occlusion boundary pixel on the occluder (Fig. 4(b)), the only different radiances between the integration set A_δ and $A_{-\delta}$ are those rays marked as green *i.e.* $r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta))$ and $r_{\frac{B}{d}}(B, \frac{B}{d}(p_u + \delta))$. Assuming that the surface color is smooth, which indicates that $r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta)) \approx r_{\frac{B}{d}}(B, \frac{B}{d}(p_u + \delta))$, we will have $\varphi_p(d + \delta) \approx \varphi_p(d - \delta)$. In other words, the color profile $\varphi_p(f)$ for boundary pixels on the occluder is approximately symmetric around the true surface depth d.

For occlusion boundary pixel on the occluded surface (Fig. 4 (c)), except the center ray, none of the other rays originate from the same surface[2]. When the disparity varies from $d - \delta_{max}$ to $d + \delta_{max}$, the integrated rays sweep across the surfaces in the directions indicated by arrows in the figure. Assuming the radiances vary linearly during the sweep, *i.e.* $r_{\frac{B}{d}}(-B, \frac{B}{d}(p_u + \delta)) = k_p^1\delta + b_p^1$ and $r_{\frac{B}{d}}(B, \frac{B}{d}(p_u - \delta)) = k_p^2\delta + b_p^2$ where k_p^1, b_p^1, k_p^2, and b_p^2 are the coefficients of the linear model for each surface[3] and δ varies in range $[-\delta_{max}, \delta_{max}]$, we have the $\varphi_p(d+\delta)$ computed as:

$$\varphi_p(d + \delta) = \frac{1}{3}(k_p^1 + k_p^2)\delta + \frac{1}{3}(b_p^1 + b_p^2 + b_p), \tag{4}$$

where $\delta \in [-\delta_{max}, \delta_{max}]$, b_p is the constant radiance from the center view. This shows that $\varphi_p(f)$ is locally linear

[1] This difference directs to a probability estimation of the occlusion map in section 5.

[2] Since δ is small, if the point is blocked from one view (in this figure, the right view) when $\delta = 0$, the blocking status will not change.

[3] Note that $k_p^* = 0$ indicates constant color surface.

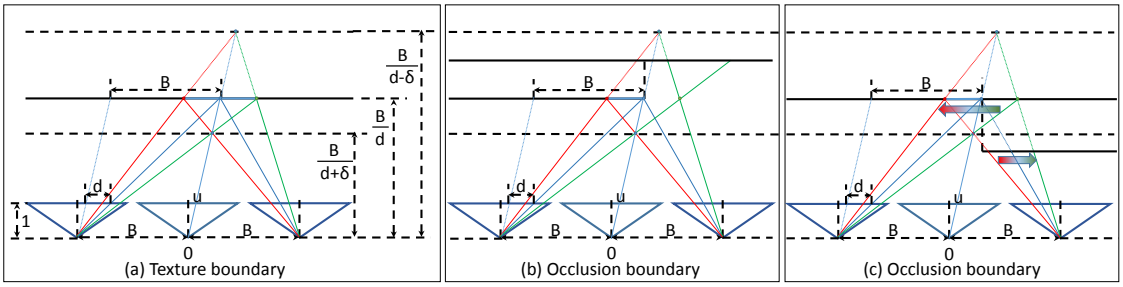


Figure 4. Local symmetry/asymmetry property analysis in LF focal stack.

around the true depth d under the linear surface color assumption. The modified function $\varphi'_p(f) = \varphi_p(f) - \varphi_p(d)$ is thus locally asymmetric around the true depth d.

Based on the analysis, for each pixel p at focal plane f, we can define the following in-focus score $s_p^{in}(f)$ according to the location of pixel p[4]:

$$s_p^{in}(f) = \begin{cases} s_p^{\varphi}(f) & \text{if } p \text{ is a non-occluded pixel} \\ s_p^{\varphi'}(f) & \text{if } p \text{ is an occluded pixel} \end{cases} \quad (5)$$

where

$$s_p^{\varphi}(f) = \int_0^{\delta_{max}} \rho\left(\varphi_p(f+\delta) - \varphi_p(f-\delta)\right) d\delta, \quad (6)$$

$$s_p^{\varphi'}(f) = \int_0^{\delta_{max}} \rho\left(\varphi'_p(f+\delta) + \varphi'_p(f-\delta)\right) d\delta, \quad (7)$$

and the function $\rho(v) = 1 - e^{-|v|_2/(2\sigma^2)}$ is a robust distance function with σ controlling its sensitiveness to noises. This distance function will be reused in other equations in this paper but probably with different σ values.

In order to exactly evaluate Eq. 5, we need to distinguish between occluded boundaries and non-occluded boundaries. However, the information about occlusion is unknown without the depth/disparity map. We resolve the chicken-and-egg problem by probabilistic reasoning. Given an occlusion probability map β (described in section 5), our final in-focus score is defined as:

$$s_p^{in}(f) = \beta_p \cdot \min(s_p^{\varphi}(f), s_p^{\varphi'}(f)) + (1 - \beta_p) \cdot s_p^{\varphi}(f). \quad (8)$$

We expect that $s_p^{in}(d(x))$ will be locally minimum (if not a global one).

Aliasing and Noise. Our analysis shows that the symmetry/asymmetry property of a pixel in the focal stack is independent of image noise or sampling rate: the focal stack synthesis process blends the same set of pixels. For example, local symmetry holds in the color profile for a texture boundary pixel is because the sets of radiances A_δ and $A_{-\delta}$ are the same. Changing the angular sampling rate or lowering spatial resolution only changes the size of A_δ and $A_{-\delta}$, and does not affect the relationship between A_δ and $A_{-\delta}$. While noise affects the individual values of the elements in the radiance set, the integrating process averages out the noise for the focal slice output. (We assume that the noise has zero mean gaussian distribution.)

[4]Non-occluded boundaries include texture boundaries and occlusion boundaries on the side of the occluder.

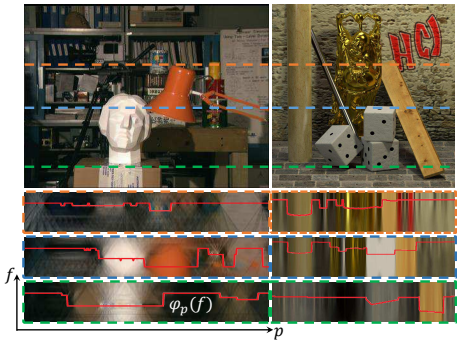


Figure 5. Examples of $\varphi_p(f)$ for scanlines of the reference view. The ground truth disparities are marked as red lines on $\varphi_p(f)$.

Fig. 5 shows examples of the $\varphi_p(f)$ for scanlines of the reference view. The left is obtained from LF data "ohta" with 5×5 sub aperture views and its disparity (reciprocal to the depth) varies within $[3, 16]$ in pixel unit, and the right one is from LF data "buddha2" with 9×9 sub aperture views and disparity range $[-0.9, 1.4]$. We can see that the local symcenter indicates its true disparity for non-occluded pixels.

4. Data Consistency Measure

In conventional stereo matching, the data consistency metric of a hypothesized disparity is based on the color difference between corresponding pixels across all input views (e.g., the data term in the graph-cut framework). If the light field captured by an LF camera has significant noise due to low exposure (small aperture), this metric becomes less reliable.

Our data consistency metric is instead based on focal stack synthesis/rendering. More specifically, given a hypothesized disparity map and an all-focus central image, we render a local focal stack around each pixel. By local: (1) we only use a small patch around the pixel to produce the focal stack, and (2) we only render a section of the focal stack of

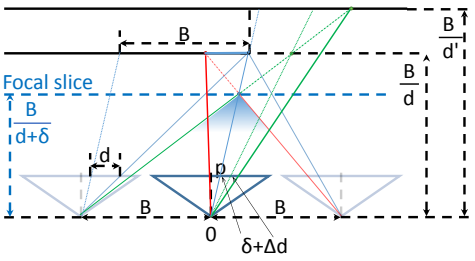


Figure 6. Retrieving the radiances from the center view that correspond to the radiances in other views. In rendering the focal slice shifted from the true disparity d by δ, the green (red) ray in the center view corresponds to the green (red) ray in the left (right) view. See text for corresponding ray identification.

range $[d(p), d(p)+\delta_{max}]$ for each pixel around its true surface disparity $d(p)$. In other words, the rendered focal stack section of each pixel starts from in-focus to out-of-focus by focal shift δ_{max}. Although the true $d(p)$ is unknown, we show that the section can be rendered given the center view and a rough disparity estimation. This leads to a focal profile for each pixel $\psi_p(\delta)$, where $\delta \in [0, \delta_{max}]$ is the focal shift deviated from the true surface disparity. We match the section $\psi_p(\delta)$ to the LF synthesized one $\varphi_p(f)$ and compute the difference as the data consistency measure.

Our key observation is that the reference center view records the majority of the scene radiances except those blocked from the center view. Fig. 6 shows an example of retrieving the radiances in the center view corresponding to the radiances from the other views for rendering the focal slice at disparity $d+\delta$ for pixel p. Although the true surface disparity d of pixel p is used in the illustration, we show that the rendered result is independent of d. We derive the corresponding radiance position in the center view through reparametrization.

When focusing at disparity $d+\delta$, the radiance from the left view is $r_1(-B, -B+p_u+d+\delta) = r_{\frac{B}{d'}}(-B, -B+\frac{B}{d'}(p_u+d+\delta))$, where d' is the disparity of the surface point where the radiance originates from. Using the Lambertian surface assumption, we have $r_{\frac{B}{d'}}(-B, -B+\frac{B}{d'}(p_u+d+\delta)) = r_{\frac{B}{d'}}(0, -B+\frac{B}{d'}(p_u+d+\delta))$. Ray $(0, -B+\frac{B}{d'}(p_u+d+\delta))$ will intersect with $z=1$ plane at $(-B+\frac{B}{d'}(p_u+d+\delta))/\frac{B}{d'} = p_u+(d-d'+\delta)$. This means that if the corresponding surface point is not occluded wrt the center view, the corresponding radiance in the center view is at distance $(\Delta d+\delta)$, where $\Delta d = d-d'$, to the current rendering pixel p. For the right view, a similar derivation shows that the corresponding radiance in the center view is at distance $-(\Delta d+\delta)$ to p. So, instead of depending on the true surface disparity, the locations of the corresponding radiances in the center view only depend on the relative disparity differences and the amount of focal shift.

Using the above analysis, we replace the radiances from other views with those from the center view to render the section of the focal slice $f \in [d(p), d(p)+\delta_{max}]$, *i.e.* $\delta \in [0, \delta_{max}]$ for each pixel p. We define a mask $k_p^\delta(q)$ to indicate the locations of the radiances in the center view when focusing at $d(p)+\delta$. This sampling kernel $k_p^\delta(q)=1$ if and only if $q = p \pm (d(p)-d(q)+\delta)$. The rendered focal slice section $\psi_p(\delta)$ can be represented as:

$$\psi_p(\delta) = \int_q k_p^\delta(q) I_q dq, \tag{9}$$

where I_q is the color of pixel q in the reference view I.

With the rendered $\psi_p(\delta)$ for each pixel, our focal stack matching score is computed as:

$$s_p^m(f) = \int_0^{\delta_{max}} \rho\left(\psi_p(\delta) - \varphi_p(f+\delta)\right) d\delta \tag{10}$$

As for initialization, we start with $\forall q, d(p)-d(q)=0$, and the sampling kernel $k_p^\delta(q)$ reduces to that of uniform sampling. We denote the correspondence matching score at initialization as $s_{p,0}^m(f)$, which will be used in Section 5 for occlusion map estimation.

Our focal stack matching measure averages over the angular samples (and hence reduces noise), making it robust for comparison against the ground truth focal stack. However, it does not account for angular color variations. In contrast, the traditional measures (color/gradient consistencies across all sub-aperture views) serve this purpose and thus we add these two traditional metrics to further improve the robustness of our estimation:

$$s_p^c(d) = \frac{1}{N}\sum_{i=1}^{N} \lambda\rho(I_{i_{q_i(d)}} - I_p) + (1-\lambda)\rho(G_{i_{q_i(d)}} - G_p), \tag{11}$$

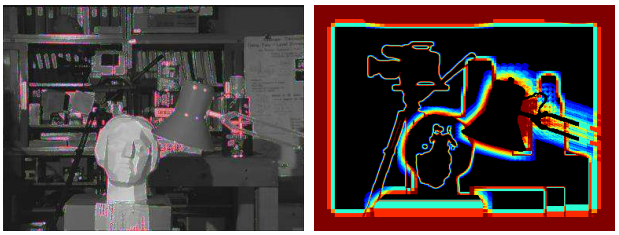
where N is the number of sub-aperture views, I_i, G_i are the i-th sub-aperture view and its gradient field, I, G are the reference image and its gradient field. Function $q_i(d)$ corresponds to the pixel in view i that corresponds to pixel p in the reference view with the hypothesis depth d. In all our experiments, we use $\lambda(=0.5)$.

5. Depth Estimation via Energy Minimization

Finally, we show how to integrate our symmetry-based focusness measure with our new data consistency measure.

Occlusion Map. For more reliable estimation, we seek to approximate an occlusion map β. In our analysis (section 3), we have shown that $\varphi_p(f)$ exhibits local symmetry for texture boundary pixels. This local symmetry gets weaker for pixels on the occluder and disappears for pixels at the true depth on the occluded surface.[5] From this

[5] We do not consider pixels on smooth region (non-boundary) since for smooth region, it is well known that the focus cue is theoretically ambiguous. However, for those smooth region, $\varphi_p(f)$ is locally constant, i.e., technically is still symmetric.



(a) Prob. map of occ. boundary (b) Ground truth occ. map

Figure 7. (a) Our probability map of occlusion boundary. (b) Ground truth occlusion map (black: no occlusion; blue to red: occluded from 1 to more than 12 views).

observation, occluded pixels will result in higher minimum in-focus score $s^{\varphi}_{p,min} = \min_f s^{\varphi}_p(f)$. They also have higher correspondence matching cost $s^m_{p,0}(f)$ since the initialization assumption is invalid at occlusion boundary pixels. Boundary pixels have relatively high variance in $\varphi_p(f)$, and hence high variance in the in-focus score $s^{\varphi}_p(f)$. By combining the above three factors, we use the following equation to compute the probability β_p:

$$\beta_p = \rho_1(s^{\varphi}_{p,min}) \cdot \rho_2(s^m_{p,0}) \cdot \rho_3(var(s^{\varphi}_p)), \qquad (12)$$

where $\rho_i(\upsilon) = 1 - e^{-\upsilon^2/(2\sigma_i^2)}$, $i \in \{1,2,3\}$, which maps υ to $[0,1]$ with σ_i set as 90% upper quartiles of the corresponding quantities over the entire image, and $var(\cdot)$ computes the variance. Fig. 7 shows a probability map of the occlusion boundary.

Algorithm. We model depth estimation as an energy minimization problem. The energy function is a typical MRF formulation:

$$E(o) = \sum_p E^{data}(o_p) + \lambda_{\mathcal{R}} \sum_{q \in \Omega_p} E^{smooth}(o_p, o_q), \qquad (13)$$
$$E^{data}(o_p) = s^{in}_p(o_p) + \lambda_m s^m_p(o_p) + \lambda_c s^c_p(o_p),$$
$$E^{smooth}(o_p, o_q) = \rho(I_p - I_q) \cdot (o_p - o_q)^2,$$

where Ω_p is the four neighborhood of pixel p, λ_m, λ_c and $\lambda_{\mathcal{R}}$ are weighting factors, and $\rho(\upsilon) = 1 - e^{-|\upsilon|_2/(2 \cdot 0.05^2)}$. Algorithm 1 shows our complete approach.

6. Experiments

We implement Algorithm 1 using graph cut algorithm for energy minimization as our basic method (denoted as "ours"). In order to further deal with challenging cases with many constant color objects, we adopt a multi-scale optimization scheme (denoted as "ours msc"), where the downsampled version of the problem is first solved and then the result is upsampled [12] to guide the disparity estimation for finer levels. In many experiments, the basic method already produces satisfactory results.

Algorithm 1: Robust Disparity Estimation from LF

Data: LF input $I_{(s,t)}$, disparity range $[d_{min}, d_{max}]$, max focal shift δ_{max}
Result: Disparity map o
Initialization $i = 0$, $o^0 = \mathbf{0}$, $E^0 = E(o^0)$, $\Delta E = 1e^6$;
Synthesis LF focal stack $\varphi_p(f)$;
Compute s^{in}_p and s^c_p for all p in center view;
while $i \leq$ *max iter and* $\Delta E \geq$ *min err update* **do**
  forall the *p in center view* **do**
    Render $\psi_p(\delta)$ for $\delta \in [0, \delta_{max}]$ based on o^i;
    Compute s^m_p ;
  Solve $o^{i+1} = \arg\min_o E(o)$ Eq. 13 by Graph-cut;
  $E^{i+1} = E(o^{i+1})$, $\Delta E = E^i - E^{i+1}$;
  $i = i + 1$;

We first experiment on the synthetic LF datasets used in [39, 37, 38, 33, 5] to validate the effectiveness of the proposed measures. We compare our technique with graph cut method (GC) involving classical data cost only (Eq. 11). Second, we evaluate on real LFs from [33] in comparison with the methods of Sun *et al*. [32], Wanner *et al*. [39] and Tao *et al*. [33]. Then, we test on our own real LFs captured by Lytro: indoor and outdoor sets. The noise level of indoor images is higher than that of outdoor set due to insufficient lighting. The sub-aperture images are extracted using MATLAB Light Field Toolbox v0.2 [9], which exhibit significant amount of noise.

Execution time. The main computation load is the focal stack computation and the local focal stack rendering. Current technique [7] shows that LF refocusing achieves real time with GPU programming. Similarly, the local focal stack rendering is also parallelizable. However, our current implementation is in Matlab, and runs on CPU at 3.4GHz with 12G memory. It takes around 20 mins for an LF of size $7 \times 7 \times 370 \times 370$ with 60 disparity labels, where the focal stack computation takes around 10 mins. We leave GPU implementation as future work.

Parameters. We can adjust the weighting parameters in Eq. 13 for optimal results according to the noise level of the input data. They are relatively stable inside each data set of similar SNR level. This is analogous to fine-tuning weights to balance data and smoothness terms in classical MRF. Table 1 lists all parameter settings used in our experiments. In all our experiments, we set δ_{max} to be $1/5$ of the overall disparity range.

Table 1. Parameter settings.

Data Set	Outdoor	Indoor	Tao's	Multi-scale	StillLife	Raytrix
λ_m	0.5	0.4	0.4	0.4	0.5	0.9
λ_c	0.8	0.9	0.9	0.9	0.9	0.5
$\lambda_{\mathcal{R}}$	0.05	0.05	0.05	0.08	0.05	0.1

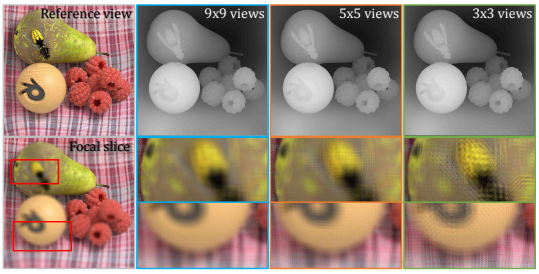


Figure 8. View number analysis. Although the refocusing results exhibit strong aliasing in sparsely sampled dataset, our method produces almost indistinguishable results.

View number analysis. In order to show the robustness of our algorithm to refocusing aliasings, we test on Still-Life [39] with sampling rates being 9x9, 5x5 and 3x3 views. Fig. 8 shows our results. Although the refocusing results exhibit strong aliasing in sparsely sampled dataset, our method produces almost indistinguishable results.

Table 2. Error comparison on datasets with Guassian noise.

	Noise variation 10/255			Noise variation 20/255		
Methods	Buddha	Cube	Mona	Buddha	Cube	Mona
SCam[5]	0.0703	0.0569	0.0973	0.2552	0.2001	0.2760
LAGC[40]	0.1325	0.5745	0.0771	0.9806	0.9249	0.4527
GCDL[39]	0.0610	0.0150	0.0347	0.3038	0.2109	0.3863
Ours	0.0173	0.0148	0.0206	0.0303	0.0154	0.0383

Noise analysis. To validate the robustness of our algorithm to noises, we add Gaussian noises (with noise variation 20/255 and 10/255) to several clean LF datasets from [39] and compare the mean square errors between our method and methods SCam [5], LAGC [40][6] and GCDL [39] in Table 2. LAGC [40] can be viewed as a sophisticated GC method with line-assisted high order regularization and occlusion reasoning. From the comparison, we can see that our results are much more robust to noises by incorporating noise-invariant data measures, while the other methods heavily rely on color matching across views and bring noises into their disparity maps.

Combined cost. As shown in the previous work [33, 22], combining different depth cues is beneficial in depth recovery. The cost metric from a single depth cue often suffers from having multiple competing local minimums in the cost profile. Combining multiple depth cues will solve the ambiguity by ruling out inconsistent local minimums between different cues. Fig. 9 shows a such example. For the occlusion boundary pixel marked by the blue circle, the cost profile (blue curve) from the focus cue has multiple competing local minimums. The preferred disparity label is not the true disparity. The correspondence matching cost profile (green curve) has a flat valley. Disparity from each single cost profile will be erroneous. Our combined cost profile

[6]We use the code from the authors' project page.

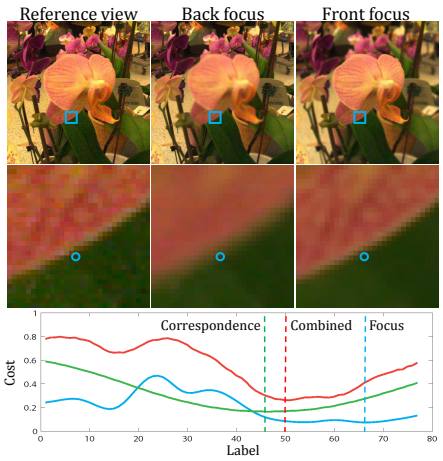


Figure 9. Cost profile on occlusion boundary. Only use the focus score(blue) or correspondence score(green) will lead to the wrong disparity estimation. Our combined score(red) implies the correct disparity.

(red curve) correctly reveals the true disparity[7], where the true disparity has low costs in both profiles.

Real examples. We compare our results with the results from Tao *et al.* [33], Sun *et al.* [32] and Wanner *et al.*'s GCDL [39] in Fig. 10. This dataset contains heavy noise. The results of Sun's and Wanner's methods are from the project page of [33]. Sun's and Wanner's methods fail to recover meaningful disparity maps because of ambiguous correspondence matching. While by combining multiple cues, Tao's algorithm gives a rough disparity estimation. However their results are overall blurry. Our multi-scale method clearly outperforms those methods. Unlike Tao's, our results have sharp boundaries and more details.

Fig. 11 shows the results on our data. Both GC and Tao's method lose fine structures, since the details are vulnerable to noises. However, our method recovers fine details with extracted noisy images. Even official Lytro software produces less satisfactory results, although it can access more accurate and cleaner sub-aperture images. Among the results, our iron wire is most complete. We recover complex plants well, such as the top leaf and the fine branch structures in the bonsai example. Our results preserve much clearer details in flower examples.

The results of Raytrix dataset are shown in Fig. 12. Compared with the results from Wanner's method [39], our results are much clearer.

Fig 13 shows the improvement from our multi-scale scheme for more challenging noisy images with objects of constant color. The correspondence matching cost is severely affected by noises. Our multi-scale successfully recover-

[7]The true disparity is validated manually by matching sub-aperture views in photoshop.

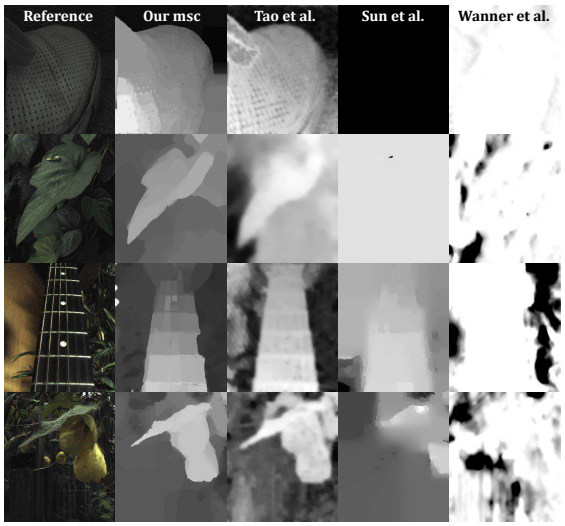


Figure 10. Comparison between our results and the results from Tao [33], Sun [32] and Wanner's GCDL [39] on the dataset from Tao et al[33].

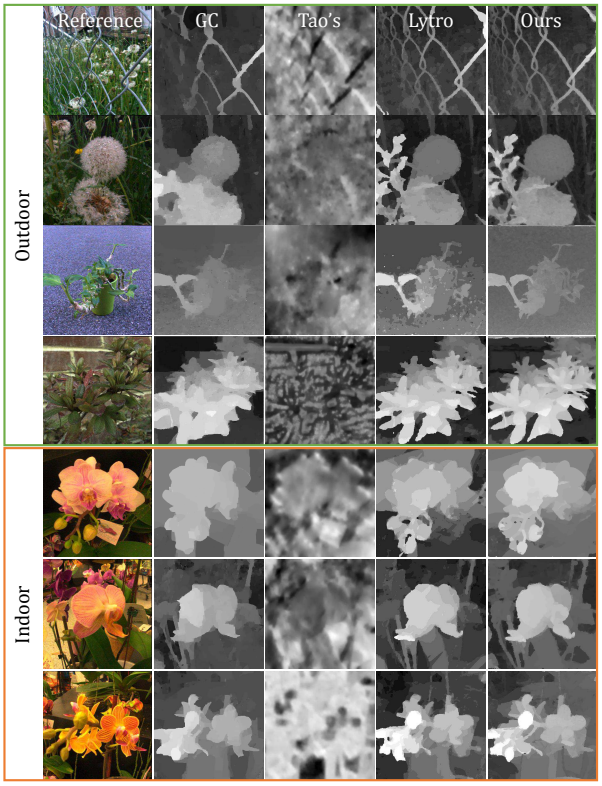


Figure 11. Disparity reconstruction results on indoor and outdoor datasets.

s the disparity map of those scenes, while the GC, Tao's method fail in finding the right correspondence.

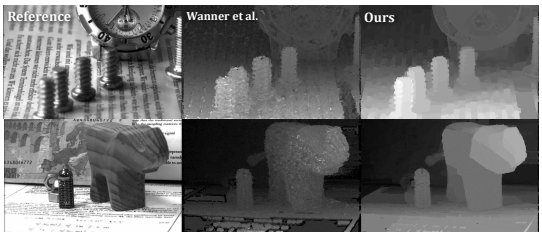


Figure 12. Disparity reconstruction results on Raytrix dataset.

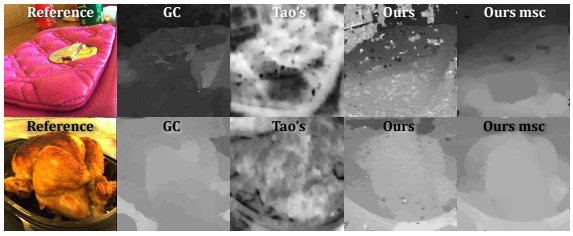


Figure 13. Disparity result comparison between GC, Tao's, ours and our multi-scale optimization method.

7. Conclusion

We have presented a new depth-from-light-field (DfLF) technique by exploring two new properties of the LF focal stack. We proposed the use of a symmetry property of the focal stack synthesized from the LF: if the focal dimension is parameterized in disparity, non-occluding pixels in the focal stack exhibit symmetry along the focal dimension centered at the in-focus slice. We showed that this symmetry property is valid even if the LF is noisy or undersampled, and as a result, useful as a new robust focus measure. We have further proposed a new data consistency measure by rendering a (local) focal stack from the hypothesized depth map and computing its difference with the LF synthesized focal stack. This new data consistency measure behaves much more robustly under noise than traditional color differences across views. We validated our approach on a large variety of LF data, captured using LF camera array and LF cameras; our results outperformeed state-of-the-art techniques.

One plan is to explore automatic parameter tuning for better energy minimization function for specific types of scenes. Of particular interest is example-based learning. In addition, we would like to investigate the use of contour detection to determine if the image/scene contains a small or large number of occlusion boundaries. This analysis can provide useful cues for adjusting λ_m, λ_c and $\lambda_{\mathcal{R}}$ in Eq. 13 accordingly. Currently, some of our results still appear rather flat due to the first-order smoothness term. Higher-order smoothness priors may be used for capturing more detailed scene geometry.

Acknowledgements

This project was partially supported by the National Science Foundation under grant IIS-1422477 and by the U.S. Army Research Office under grant W911NF-14-1-0338.

References

[1] Life in a different light. https://www.lytro.com/.

[2] Matlab light filed toolbox. http://marine-wp.acfr.usyd.edu.au/research/plenoptic-imaging/.

[3] The (new) stanford light field archive. http://lightfield.stanford.edu/.

[4] Raytrix: 3d light field camera technology. http://www.raytrix.de/.

[5] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu. Light field stereo matching using bilateral statistics of surface cameras. In *CVPR*, 2014.

[6] Y.-C. Chen, Y.-C. Wu, C.-H. Liu, W.-C. Sun, and Y.-C. Chen. Depth map generation based on depth from focus. In *Electronic Devices, Systems and Applications (ICEDSA), 2010 Intl Conf on*, pages 59–63. IEEE, 2010.

[7] G. Chunev, A. Lumsdaine, and T. Georgiev. Plenoptic rendering with interactive performance using gpus. In *SPIE Electronic Imaging*, 2011.

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Processing*, 2007.

[9] D. G. Dansereau, O. Pizarro, and S. B. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *CVPR*. IEEE, 2013.

[10] P. Favaro and S. Soatto. A geometric approach to shape from defocus. *IEEE TPAMI*, 2005.

[11] P. Favaro, S. Soatto, M. Burger, and S. J. Osher. Shape from defocus via diffusion. *IEEE TPAMI*, 2008.

[12] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*. IEEE, 2013.

[13] C. Frese and I. Gheta. Robust depth estimation by fusion of stereo and focus series acquired with a camera array. In *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 243–248. IEEE, 2006.

[14] I. Gheta, C. Frese, and M. Heizmann. Fusion of combined stereo and focus series for depth estimation. In *GI Jahrestagung (1)*, pages 359–363, 2006.

[15] I. Gheta, C. Frese, M. Heizmann, and J. Beyerer. A new approach for estimating depth by fusing stereo and defocus information. In *GI Jahrestagung (1)*, pages 26–31, 2007.

[16] S. W. Hasinoff and K. N. Kutulakos. Confocal stereo. In *ECCV*. Springer, 2006.

[17] S. Heber and T. Pock. Shape from light field meets robust pca. In *ECCV*, pages 751–767. Springer, 2014.

[18] S. Heber, R. Ranftl, and T. Pock. Variational shape from light field. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2013.

[19] M. H. Kamal, P. Favaro, and P. Vandergheynst. A convex solution to disparity estimation from light fields via the primal-dual method. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 350–363. Springer, 2015.

[20] C. Kim, A. Hornung, S. Heinzle, W. Matusik, and M. Gross. Multi-perspective stereoscopy from light fields. *ACM Trans. Graph.*, 2011.

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

[22] F. Li, J. Sun, J. Wang, and J. Yu. Dual-focus stereo imaging. *Journal of Electronic Imaging*, 19(4):043009–043009, 2010.

[23] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. In *CVPR*. IEEE, 2014.

[24] A. S. Malik, S.-O. Shim, and T.-S. Choi. Depth map estimation using a robust focus measure. In *ICIP*, volume 6, pages VI–564. IEEE, 2007.

[25] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. Active refocusing of images and videos. In *ACM Trans. Graph.* ACM, 2007.

[26] S. K. Nayar. Shape from focus system. In *CVPR*. IEEE, 1992.

[27] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE TPAMI*, 1994.

[28] S. K. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. *IEEE TPAMI*, 1996.

[29] 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*, 2(11), 2005.

[30] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi. Depth estimation and image restoration using defocused stereo pairs. *IEEE TPAMI*, 2004.

[31] Y. Y. Schechner and N. Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 2000.

[32] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*. IEEE, 2010.

[33] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *ICCV*. IEEE, 2013.

[34] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *CVPR*. IEEE, 2006.

[35] S. A. Valencia and R. M. Rodriguez-Dagnino. Synthesizing stereo 3d views from focus cues in monoscopic 2d images. In *Electronic Imaging 2003*, pages 377–388. International Society for Optics and Photonics, 2003.

[36] S. Wanner and B. Goldluecke. Spatial and angular variational super-resolution of 4d light fields. In *ECCV*, 2012.

[37] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE TPAMI*, 2013.

[38] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, 2013.

[39] S. Wanner, C. Straehle, and B. Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *CVPR*, 2013.

[40] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu. Line-assisted light field triangulation and stereo matching. In *ICCV*, 2013.

[41] L. Zhang, S. Vaddadi, H. Jin, and S. K. Nayar. Multiple view image denoising. In *CVPR*. IEEE, 2009.

[42] C. Zhou, O. Cossairt, and S. Nayar. Depth from diffusion. In *CVPR*. IEEE, 2010.