DEPTH FUSED FROM INTENSITY RANGE AND BLUR ESTIMATION FOR LIGHT-FIELD CAMERAS

Yatong Xu, Xin Jin and Qionghai Dai

Shenzhen Key Lab of Broadband Network and Multimedia,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

ABSTRACT

Light-field cameras attract great attention because of its refocusing and perspective-shifting functions after capturing. The special 4D-structured data contains depth information. In this paper, a novel depth estimation algorithm is proposed for light-field cameras by fully exploiting the characteristics of 4D light-field data. A novel tensor, intensity range of pixels within a microlens, is proposed, which presents strong correlation with the transition on focus, especially for texture-complex regions. Meanwhile, the other tensor, defocus blur amount is utilized to estimate the focus level, which generates more accurate depth estimation especially for homogeneous regions. Then, the depths calculated from the two tensors are fused according to the variation scale of intensity range and the minimal defocus blur amount under spatial smoothness constraints. Compared with the representative approaches, the depth generated by the proposed approach presents richer details for texture regions and higher consistency for unified regions.

Index Terms— depth estimation, Light-field, intensity range, depth fusion, confidence measure

1. INTRODUCTION

The newly released commercial light-field cameras, Lytro [1] and RayTrix [2], have attracted great attentions. Based on the theory of light-field, this kind of cameras are capable of refocusing and perspective-shifting simultaneously from a single shot with only one camera [3]. Furthermore, depth estimation with light-field cameras has been regarded as a much cheaper and easier way for ordinary users.

The existing methods of depth estimation for light-field cameras can be mainly classified into two categories: stereo matching approaches [4-7] and light-field approaches [9, 12-13]. Stereo matching approaches calculate depth from the correspondence relationship among sub-aperture images acquired by light-field cameras [4]-[7]. However, the computational complexity of such algorithms is extremely high and the quality of the depth is subject to the resolution of the input sub-aperture images, which is much lower, compared to the images captured by multi-view systems. Thus, it greatly affects the efficiency of stereo matching [8]. Some approaches updated stereo matching algorithms, e.g. considering the line structure of rays [9]. But they still only use the correspondence relationship in the light-field data. Although Light-field approaches utilize correspondence together with defocus information contained in the light-field [10, 11], the estimated depth still lack details in homogeneous regions, e.g. different cost functions are proposed by Min-Jung Kim et al. [12] for different cues to estimate the depth and the algorithm proposed by Tao et al. [13] further combines the confidence measures of the two cues to improve the accuracy of the estimated depth. Nevertheless, both of them fail when the captured scene is texture-less.

In this paper, a novel depth estimation algorithm is proposed for light-field cameras. By analyzing rendered light-field images with focus variation in the constructed volume, a novel tensor, intensity range of pixels within a microlens, is proposed, which indicates the focusing distance accurately, especially for the regions with complex texture. Moreover, the other tensor, defocus blur amount measured by blur estimation, aids to calculate the accurate focus distance for different objects in the scene, especially for the homogeneous regions. Then, based on the variation scale of intensity range and the minimal defocus blur amount from blur estimation, depths estimated by the two tensors are fused via global optimization with constraints of spatial smoothness. The proposed method generates the depth with richer transition details and higher consistency, compared with state-of-the-art works.

The rest of the paper is organized as follows. The framework of the proposed algorithm is illustrated in section 2. Section 3 describes depths estimation from the two tensors: intensity range and blur estimation, respectively. Section 4 illustrates the depth fusion and optimization. Experimental results are shown in section 5. And the conclusions are drawn in section 6.

2. THE PROPOSED FRAMEWORK

The framework of the proposed algorithm is in Fig. 1. First, *Refocusing* is performed to construct a volume from a single shot captured by a light-field camera. Point spread function (PSF) proposed by Ng et al. [14] is exploited during *Refocusing* as:

$$L_z(x,y,u,v) = L_o\left(x+u\left(1-\frac{1}{z}\right), y+v\left(1-\frac{1}{z}\right), u, v\right), \quad (1)$$

where L_0 is the rectification of the captured image[15]; L_z is the refocused image at depth level z; x, y are spatial coordinates and u, v are angular coordinates on the image plane. Thus, a number of refocused images are generated and organized according to the focusing plane varying from close to far to form a volume, which will be used for *Tensor Extraction*. Meanwhile, the central pixel of each microlens is picked out from L_0 to accomplish *Central Sub-aperture Image Acquisition* for calculating the smoothness constraints in the following processing.

Then, *Tensor Extraction* is applied to the volume of refocused images generated above to extract two variants which present high correlation with the variation in focusing plane. The first variant, intensity range is proposed and verified based on a comprehensive analysis on the light-field data. Exploiting the minimum value of

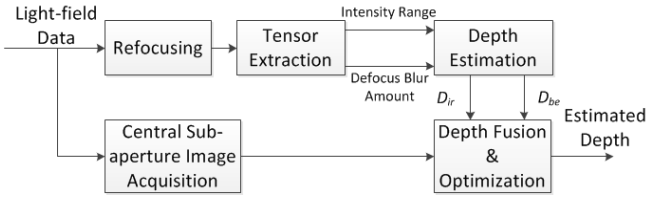


Fig. 1. Framework of the proposed method.

intensity range during refocusing, a depth image, D_{ir}, is calculated by *Depth Estimation*. The second variant, defocus blur amount, is used to measure the focus level of each pixel during the focus variation. A representative and efficient blur estimation algorithm proposed in [18] is adopted in this paper to the measure the defocus blur amount of images generated by *Refocusing* and integrated in the angle domain. Utilizing the minimum defocus blur amount, another depth image, D_{be}, is also calculated by *Depth Estimation*. The definition of the tensors and the related analyses will be described in detail in Section 3.

Finally, the two estimated depth, D_{ir} and D_{be}, are fused according to their accuracy under the neighborhood smoothness constraints via *Depth Fusion & Optimization*. The accuracy of D_{ir} and D_{be} is measured based on the variation scale of intensity range and the minimum defocus blur amount from blur estimation, respectively. The neighborhood smoothness constraints are set considering the gradient of the central sub-aperture image. The optimization is implemented according to [16]. By fusing the two depth maps, the final estimated depth presents high consistency and accuracy, e.g. decreasing the variance within the region of the same depth and sharpening the boundaries.

3. TENSOR EXTRACTION AND DEPTH ESTIMATION

3.1. Depth from Intensity Range

In order to estimate the depth with rich details and high accuracy simultaneously for light-field cameras, an efficient tensor strongly correlated with the variation in focusing distance is investigated.

According to the imaging theory of light-field cameras, as the focusing point moves away from a specific position in the real 3D space, the pixels corresponded to the focusing point scatter from one microlens to several microlenses around [14]. Inversely saying, if the spatial point is focused well, the intensity range of the corresponding pixels should be lower than that when the point is out-of-focus.

Therefore, intensity range $R_z(x, y)$ is proposed and extracted from the constructed volume, composed of a number of refocused images, at every hypothetical depth level z as:

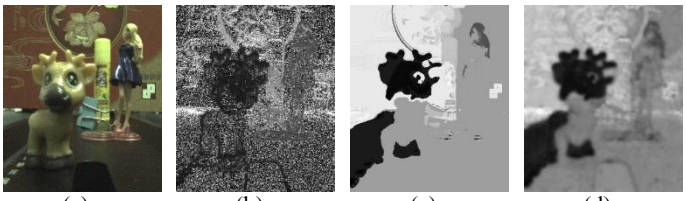
$$R_z(x,y) = I_{\max_{u,v}}(x,y,u,v) - I_{\min_{u,v}}(x,y,u,v), \quad u,v \in M \tag{2}$$

where $I(x, y, u, v)$ is the pixel intensity at (u, v) within the microlens (x, y) in L_z and M is the set of pixels within the microlens. Then, the depth from intensity range at pixel (x, y), $D_{ir}(x, y)$, is estimated by:

$$D_{ir}(x,y) = \arg\min_{z} R_z(x,y), \tag{3}$$

3.2. Depth from Defocus Blur Amount

The depth from intensity range, D_{ir}, reveals more accurate estimation in texture-complex regions. To further improve the dep-



(a) (b) (c) (d)

Fig. 2. (a) Central sub-aperture image; Depth from: (b) Intensity range; (c) Blur estimation; (d) Depth fusion and optimization.

-th accuracy in texture-less regions, a tensor called defocus blur amount is proposed.

The tensor, defocus blur amount, is measured by blur estimation [18] on the refocusing images integrated in the angle domain. $\overline{L_z}(x,y)$ is given by:

$$\overline{L_z}(x,y) = \frac{1}{N}\sum_{(u,v)} L_z(x,y,u,v), \tag{4}$$

where N is the number of pixels within the same microlens. The ratio between the gradients of $\overline{L_z}(x,y)$ and its re-blurred image, which is formed by using a Gaussian kernel at edge locations and then propagated according to [18], is calculated. Thus, defocus blur amount maps, B_z, corresponded with $\overline{L_z}(x,y)$ at each depth level z are generated. Then, the depth estimated from defocus blur amount at pixel (x, y), $D_{be}(x, y)$, is given by:

$$D_{be}(x,y) = \arg\min_{z} B_z(x,y), \tag{5}$$

which extracts the depth level z corresponding to $B_z(x, y)$ with the minimum defocus blur amount as the depth of pixel (x, y).

D_{ir} and D_{be} estimated for the sample scene shown in Fig. 2 (a) are shown in Fig. 2 (b) and (c), respectively. It is obvious that D_{ir} benefits regions with complex texture, while D_{be} provides higher consistency and accuracy for unified regions. Therefore, to exploit the advantages from both of them, an optimization model is proposed by analyzing the response of $R_z(x, y)$ and $B_z(x, y)$ under the smoothness constraints of the texture.

4. DEPTH FUSION AND OPTIMIZATION

In order to fuse D_{ir} and D_{be} to strengthen the final estimated depth D_{final} by preserving clear boundaries and the consistency in homogeneous regions, an optimization model is proposed based on the pixel-wise measurement of the accuracy of D_{ir} and D_{be}, and the neighborhood smoothness constraints. The model is given by:

$$\begin{aligned}\underset{D_{final}}{\text{minimize}} & \sum_{(x,y)} \left| D_{final} - \left(\lambda_{ir} C_{ir} D_{ir} + \lambda_{be} C_{be} D_{be}\right)\right|_{(x,y)} \\ & + \lambda_{flat} \sum_{(x,y)} \left(\left| \frac{\partial D_{final}}{\partial x} - \frac{\partial G}{\partial x} \right| + \left| \frac{\partial D_{final}}{\partial y} - \frac{\partial G}{\partial x} \right| \right)_{(x,y)} + \\ & \lambda_{smooth} \sum_{(x,y)} \left(\left| \frac{\partial^2 D_{final}}{\partial x^2} - \frac{\partial^2 G}{\partial x^2} \right| + \left| \frac{\partial^2 D_{final}}{\partial y^2} - \frac{\partial^2 G}{\partial y^2} \right| \right)_{(x,y)},\end{aligned} \tag{6}$$

where C_{ir} and C_{be} are the confidence map which measures the accuracy of D_{ir} and D_{be}, respectively; λ controls the weight between D_{ir} and D_{be}; λ_{flat} and λ_{smooth} control the Laplacian constraint and the second derivative kernel respectively to enforce the flatness and overall smoothness of the final estimated depth. Gradient G extracted from the central sub-aperture image is applied as constraints to improve the depth consistency in the hom-

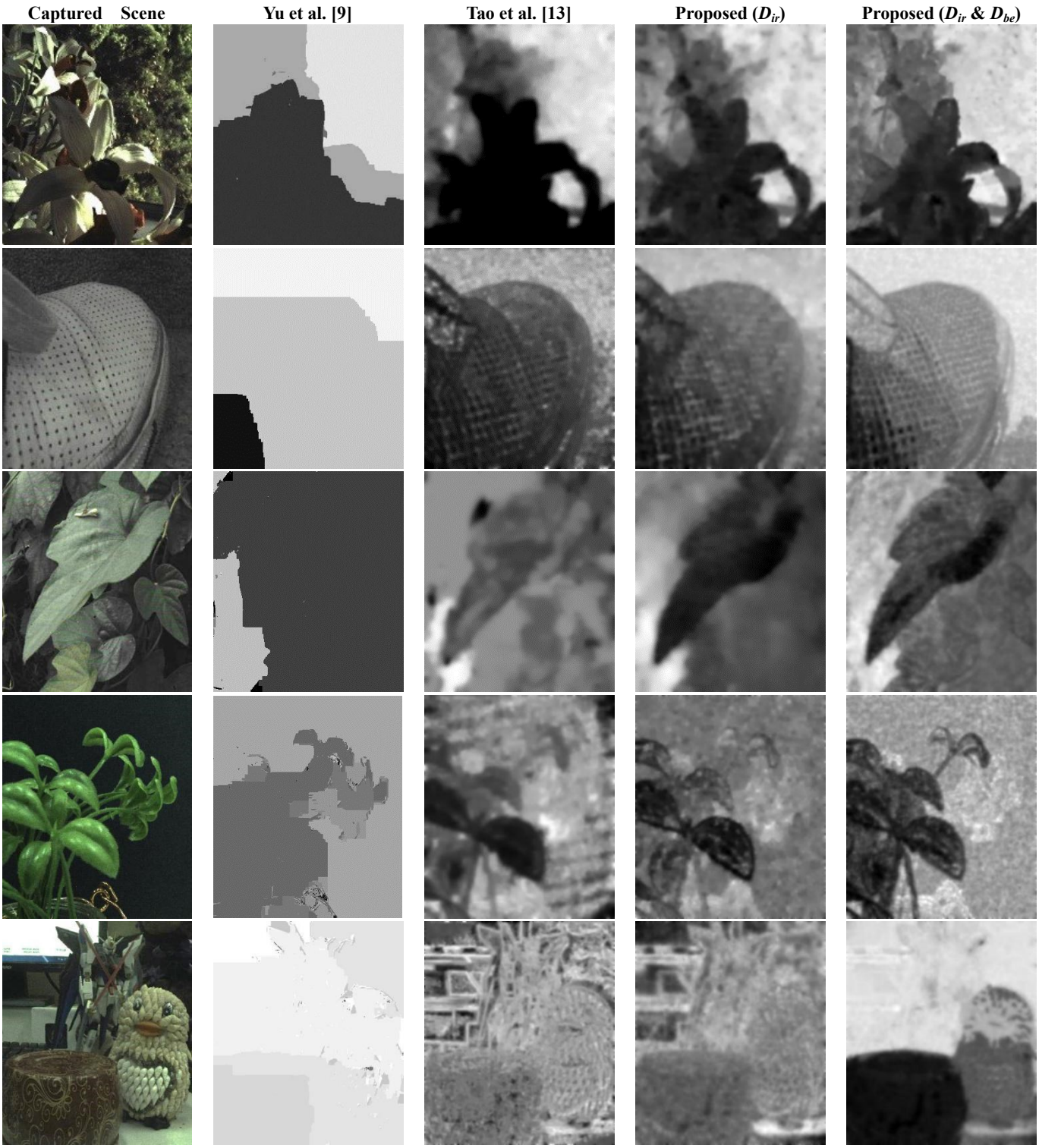


Fig. 3. Experimental comparison of indoor and outdoor scenes.

-ogeneous regions while preserving boundaries simultaneously. The definition of C_{ir} and C_{be} are given as follows.

4.1. Confidence Map of Intensity Range

In order to measure the accuracy for the depth estimated by intensity range, the response of the defined tensor, intensity range, is analyzed. It is found that if $R_z(x, y)$ presents a large variation scale along z, i.e. the difference between the minimum and maximum of $R_z(x, y)$ is big, it always leads to a more accurate $D_{ir}(x, y)$. Thus, $C_{ir}(x, y)$ is defined as:

$$C_{ir}(x,y) = \text{NORMALIZE}\left(\max R_z(x,y) - \min R_z(x,y)\right), \quad (7)$$

The measure of $C_{ir}(x, y)$ produces a high value when there is a big difference between the minimum and maximum of $R_z(x, y)$. Accurate depth is generated by utilizing C_{ir} to strengthen the correct estimations and degrade the incorrect estimations of D_{ir} via the global optimization.

4.2. Confidence Map of Blur Estimation

In order to measure the accuracy for the depth estimated by defocus blur amount, the response of the defined tensor is also analyzed. Since lower defocus blur amount corresponds to a better focus, we regards that the depth retrieved from lower defocus blur amount presents higher confidence. Thus, $C_{be}(x, y)$ is defined by:

$$C_{be}(x,y) = 1 - \text{NORMALIZE}\left(\min_z B_z(x,y)\right). \quad (8)$$

C_{be} produces high values for pixels focused better during refocusing, while produces low values for blurry pixels so that to enhance the accurate estimation of D_{be} and degrade the inaccurate ones.

Applying the fusing and optimization to D_{ir} and D_{be}, D_{final} for the sample scene in Fig. 2 (a) is shown in Fig. 2 (d). Compared with D_{ir} and D_{be}, shown in Fig. 2 (b) and (c), D_{final} provides richer transition details for depth discontinuity and higher consistency for depth uniformity.

5. EXPERIMENTAL RESULTS

The effectiveness of the proposed algorithm is demonstrated by comparison with state-of-the-art methods proposed by Yu et al. [9] and Tao et al. [13]. Yu et al. [9] is representative in adapting stereo matching algorithm to depth estimation using light-field data. Tao et al. [13] is a representative light field approach which combines defocus and correspondence cues to estimate dense depth with a light field camera. All images in the paper are captured by Lytro1.0 [1]. For Yu et al. [9], the disparity varies among [-2, 2] pixels, with the step as 0.2 pixels, σ of Gaussian filter is 1.0 and the direction parameter is set to fit the arrangement of the light-field of Lytro1.0 [1]. Other parameters are set to default values. The light- field data of the first three scenes in Fig. 3 are downloaded from [17].

Fig. 3 compares the estimated depths of the scenes on the leftmost column. The processing results of Yu et al. [9] are shown in the second column from the left. It is obvious that it provides the major depth levels for each scene, while loses all the details in depth transition because of inefficient line-structure detecting. The processing results of Tao et al. [13] are shown in the third column from the left. Although they can provide more details in depth transition relative to that of Yu et al. [9], the granularity of depth along the variation in distance is still very coarse. Obvious depth errors happen where the tensors based on contrast and angular variance both fail. The second column from the right shows the depths estimated only by intensity range. Compared with Yu's and Tao's results, it provides more depth transition details. It is also observed that some errors exist in regions lack of texture, especially for the last scenes. The depths estimated by fusing D_{ir} and D_{be} are shown in the rightmost column. The comparison between the last two columns gives a self-proof that by fusing the depth from blur estimation, the accuracy and consistency of the estimated depth get improved, especially for the texture-less regions. It can be seen that the proposed fusion method is effective in producing much richer depth details, clearer boundaries with more consistent depth.

6. CONCLUSIONS

In this paper, an efficient depth estimation method is proposed for light-field cameras. Two novel tensors: intensity range of pixels within a microlens and defocus blur amount are proposed to track the focus variation. Depths calculated from the two tensors are fused according to the variation scale of intensity range and the minimum defocus blur amount from blur estimation via global optimization with the constraints of neighborhood smoothness. The effectiveness of the proposed algorithm is demonstrated by the comparison with the existing representative approaches. Much richer transition details and higher consistency in homogeneous regions together with clearer object boundaries are achieved in the estimated depth, which will benefit the subsequent applications in the future.

7. ACKNOWLEDGMENT

This work was supported in part by the NSFC-Guangdong Joint Foundation Key Project (U1201255) and project of NSFC 61371138,China.

8. REFERENCES

[1] "Lytro - Home", https://www.lytro.com/.

[2] "Raytrix | 3D light field camera technology", http://www.raytrix.de/.

[3] M. Levoy, "Light fields and computational imaging," IEEE Computer, 2006, 39(8): 46-55.

[4] E. H. Adelson and J. Y. Wang, "Single lens stereo with a plenoptic camera," IEEE Transactions on Pattern Analysis and machine intelligence (TPAMI), vol. 14, no. 2,pp. 99–106, 1992.

[5] C. Perwass and P. Wietzke, "Single lens 3D-camera with extended depth-of-field," In Proceedings of the conference on Society of Photo-Optical Instrumentation Engineers (SPIE Elect. Imaging), 2012.

[6] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," In Special Interest Group for Computer Graphics (SIGGRAPH), 2013.

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

[8] T. Georgiev, Z. Yu, A. Lumsdaine, and S. G. Qualcomm, "Lytro camera technology: theory, algorithms, performance analysis," In Proceedings of the conference on Society of Photo-Optical Instrumentation Engineers (SPIE Elect. Imaging), 2013.

[9] Z. Yu, X. Guo, and J. Yu, "Line assisted light field triangulation and stereo matching," in IEEE International Conference on Computer Vision (ICCV), 2013.

[10] 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.

[11] V. Vaish, R. Szeliski, C. Zitnick, S. Kang, and M. Levoy, "Reconstructing occluded surfaces using synthetic apertures: stereo, focus

and robust measures," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[12] M. J. Kim, T. H. Oh, and I. S. Kweon, "Cost-aware depth estimation for Lytro camera," In IEEE Conference on Image Processing (ICIP), 2014.
[13] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in IEEE International Conference on Computer Vision (ICCV), 2013.

[14] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Computer Science Technical Reports (CSTR) 2005-02, 2005.

[15] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[16] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3D object dataset: putting the kinect to work," in IEEE International Conference on Computer Vision (ICCV), 2011.

[17] "Depth from Combining Defocus and Correspondence Using light-Field Cameras – U.C. Berkeley Computer Graphics Reserach", http://graphics.berkeley.edu/papers/Tao-DFC-2013-12/index.html.

[18] S. Zhuo and T. Sim, "Defocus map estimation from a single image," IEEE Pattern Recognition, 2011, 44(9): 1852-1858.