Camera Pose Free Depth Sensing Based on Focus Stacking

Kai Xue, Yiguang Liu(✉), Weijie Hong, Qing Chang, and Wenjuan Miao

Vision and Image Processing Lab(VIPL), College of Computer Science,
Sichuan University, Chengdu 610065, People's Republic of China
lygpapers@aliyun.com

Abstract. Binocular or multi-view depth imaging usually fails when camera pose is fixed, but depth need to be sensed with single fixed camera in some scenarios. To tackle this problem, we present a camera pose free depth sensing based on focus stacking. We first compute the mapping between scene depth and focus ring. A sharpness function based on discrete Fourier transform (DFT) is provided to calculate the in-focus parts, and parallax method is used to obtain the mapping between object space and image space. After the camera mapping calculation, we calculate the depth of the scene from the image distance map (IDM), which is generated by fusing in-focus areas of different images in the focal stacks. To reconstruct the scene, a depth map and an all-in-focus (AiF) image are combined from the focal stacks and IDM. Experimental results show that the proposed method is effective and robust compared to some binocular stereo and focus stacking method, and the depth sensing accuracy of our method is over 98%.

Keywords: Depth sensing · Focus staking · Camera pose

1 Introduction

Depth sensing from images is a traditionally hot research field. Generally, the depth information is computed by stereo matching, in which the pixels are matched between images, and the depth are obtained by solving the geometry relation between the camera poses and matched pixels. The stereo matching-based methods are facing many difficulties in practice, such as mutual occlusion, textureless region, none diffuse reflection, etc. In this paper, we propose a novel method for depth sensing with better convenience in general use, since it avoids solving the geometry relation of the cameras and it does not need the time-consuming pixel-wise matching. The proposed method is based on focus stacking, which is primally designed to gain large depth of field (DoF) in one image by fusing images captured at different focusing distances. It is especially useful to compute AiF images when the DoF is limited [1,8,14]. Focus stacking

Supported by NSFC under Grant 61860206007 and 61571313.

© Springer Nature Switzerland AG 2019

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 181–192, 2019.
https://doi.org/10.1007/978-3-030-34113-8_16

has been used in many other applications in image processing and computational photography [2,4,5], such as image denoising [17] and 3D surface profiling [3]. And there are also some works proposed to improve its efficiency [13].

Camera imaging is a bijection from object space, generating an inverted, shrunken and distorted real image to image space. Taking photos is a sampling process of the image space with the shift of image sensor. However, it is not a standard sampling process, for the image sensor receive and block all the light beams casting on it at any location, as shown in Fig. 1. In-focus parts produce the best sharpness when the objects are exactly on the focal plane. Others are differently blurred into circles of confusion (CoC) related to the distance to the camera.

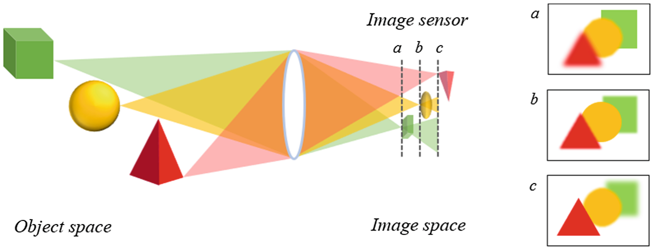

Fig. 1. A simplified model of camera imaging. Three objects in object space at different depth and their images in image space. a–c are 3 positions of the image sensor, and right are pictures taken at these positions.

As shown in Fig. 2, in-focus points in image space can be described as $(h_0, v_0) = (Hf_0/(d - f_0), df_0/(d - f_0))$, where H and h_0 are radial distances from the object points and its corresponding image points to the optical axis respectively. The in-focus image distance v_0 is related to the depth d and focal length f_0. In this condition, all light beams from objects converge to a single sensor point, which leads to image pixels with best sharpness. With the altering of image distance, defocus blur is generated. The defocus amount of a pixel denoted by c is defined as the diameter of CoC [10], which is positive related to the image distance v when v_0 is fixed. N_s is the f-stop number:

$$c = \frac{f_0(v - v_0)}{N_s v_0} \tag{1}$$

The defocus blur can be denoted by a convolution of a sharp image with the point spread function (PSF). The PSF is usually approximated by a Gaussian function. And the standard deviation σ measures the diameter of CoC.

The approach we present is to reconstruct the image in image space, and invert the camera mapping from image space to the object space. Figure 3 shows the framework of the presented method. There are two key points to reconstruct the scene. We first calculate the inverse mapping $(H, d) = (h_0 f_0/(v_0 - f_0), v_0 f_0/v_0 - f_0)$ between image distance and scene depth, in which sharpness

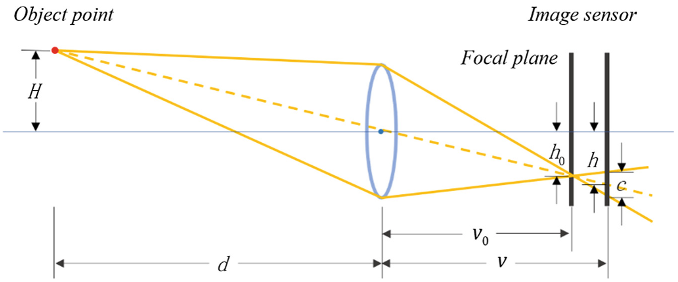

Fig. 2. Description of in-focus and defocus image points.

evaluation function and parallax method are used to confirm the position of the image sensor plane. Then we reconstruct the scene with the focal stacks. A defocus blur estimation method [18] is improved to segment in-focus regions of the focal stacks and fuse an IDM to mark the in-focus patches and an AiF image to record their colors. Now we have all the 3D information which is used to reconstruct the scene.

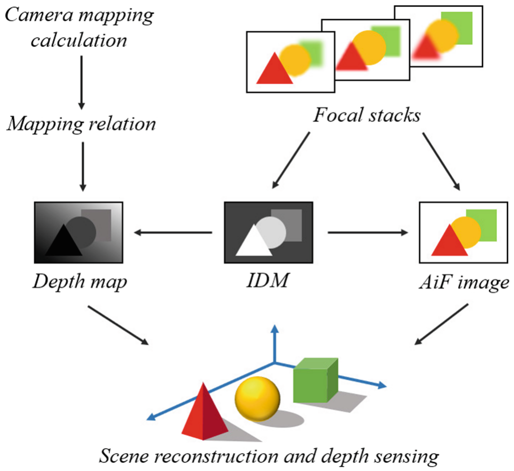
Camera mapping calculation
Mapping relation
Focal stacks
Depth map
IDM
AiF image
Scene reconstruction and depth sensing

Fig. 3. Framework of the proposed depth sensing method.

2 Camera Mapping Calculation

We calculated the inverse mapping between image distance and object distance in the INTRODUCTION, which is undifferentiated to each camera. But the positions of image sensor are complicated to learn which is related to camera types and conditions. Camera mapping calculation is to confirm the relation between scene depth and its corresponding focus ring positions. In this section, we use the parallax method to calculate the image distance. Before this, sharpness evaluation function is used to evaluate the clarity of local areas and confirm the in-focus parts in the focal stacks.

2.1 Sharpness Evaluation Function

Image sharpness is a measurement of the focus degree. There are more details and edge features in focused images compared to defocused ones. In the view of pixels on the images, clear images change dramatically especially at complex texture areas. The defocus blur modeled as a convolution of a sharp image with the PSF. We can evaluate the sharpness of images by the complexity of pixels at a certain area. Sharpness evaluation function is used for filtering the in-focus parts in the focal stacks.

Two objects with complex texture and flat plane are chosen as aim objects in the experimental scene. The planes perpendicular to the optical axis provide homogeneous depth areas and the complex textures offer more features. Positions of the camera and two aim objects in different depth were fixed. Shift the image plane and focus on two aim objects respectively. During this process, we assume $a_s(x, y)$ as the aim object area on the image plane, and s is the displacement of focusing ring. It is always hard to distinguish the best in-focus position with human eyes. There are many sharpness evaluation functions to solve this problem, especially in space domain, such as gradient square function, Roberts gradient function, Brenner function and Laplacian function [15,16].

We employ the DFT in frequency domain into the sharpness evaluation function, which shows superiorities to focal stacks than other methods. For the defocus blur can be treated as a PSF convolution of the in-focus image, and the degree of the defocus blur is related to diameter of CoC. In the frequency domain, PSF like a lowpass filtering (LSP) reduce the high frequency parts. And with the expansion of CoC, more high frequency parts are abandoned. We obtain the frequency image of $a_s(x, y)$ after Fourier transformation $A_s(x, y) = \mathcal{F}\{a_s(x, y)\}$. Then we compare the aim parts with different s and calculate our sharpness evaluation function

$$\begin{cases} f_s(x,y) = \sum_{x=0}^{Max(x)} \sum_{y=0}^{Max(y)} A_s(x,y) W(x,y) \\ W(x,y) = \left(\frac{Max(x)-x}{\sum_{x=0}^{Max(x)} x}, \frac{Max(y)-y}{\sum_{y=0}^{Max(y)} y}\right) \end{cases}, \tag{2}$$

where $W(x, y)$ is weight matrix to provide higher weights to the high frequency parts. In this subsection, we obtain the best in-focus positions of two aim objects with their corresponding focus ring positions denoted by s_a and s_b respectively.

2.2 Parallax Method

To distinguish the two image planes, we invert the aim1 image plane at v_a in front of the optic center, as shown in Fig. 4. Two objects reflect their images at

(x_{1L}, y_1) and (x_{2L}, y_2) on the left camera sensor respectively, and reflect their images at (x_{1R}, y_1) and (x_{2R}, y_2) on the right camera sensor respectively. As the focus ring positions are known, we calculate the corresponding image distances v_a and v_b using parallax method in Eq. (3), where $\varDelta$ is unit pixel length of the image sensor plane.

$$\begin{cases} v_a = \frac{(x_{1L}-x_{1R})\varDelta+B}{B} f_0 \\ \\ v_b = \frac{(x_{2L}-x_{2R})\varDelta+B}{B} f_0 \end{cases} \tag{3}$$

We obtain the two best image distances of in-focus image planes with their responding focus ring displacements s_a and s_b. The relation between them is a liner mapping $v = k^{'}(s - s_b) + v_b, k^{'} = (v_a - v_b)/(s_a - s_b)$. Combining the inverse mapping function, we can acquire the position of the focus ring and their corresponding scene depth.

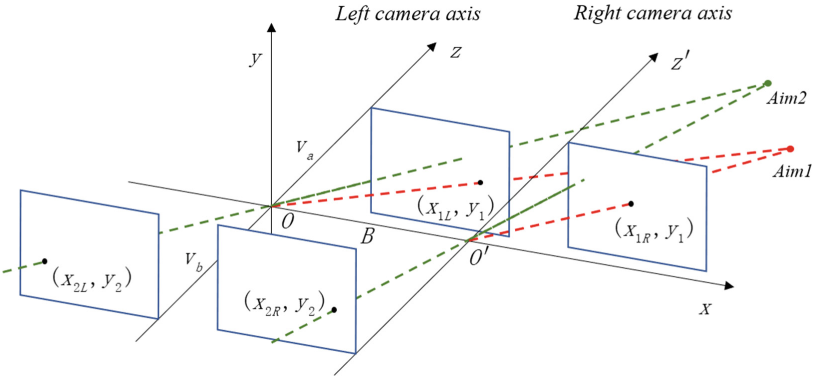

Fig. 4. The model of parallax method. World coordinate system with the left optic center as origin is established, whose x-axis is horizontal direction, y-axis is vertical direction and z-axis is depth direction. $O^{'}$ is the right camera optic center and B is the baseline between the two optic centers. v_a and v_b are the optimum image plane distances of two aim points.

$$u = \frac{{f_0}^2}{k'(s - s_a) + v_a - f_0} + f_0 \tag{4}$$

It should be noted that the mapping calculation only needs to conduct once for every camera, when the mapping between in-focus images plane and responding focus ring displacement is unknown. If camera mapping calculation is finished during the camera manufacturing process, no more camera mapping calculation is needed to users and depth sensing can be directly conducted.

3 Depth Sensing

3.1 Defocus Model of Focus Stacks

By shifting the image plane of an aperture camera and coaxially photographing the scene, we can obtain the focal stacks with different image distances. Accord-ing to Eq. (1), the same image coordinates in focal stacks will be defocus blurred with different sizes of PSF caused by the image distance. The defocus blur image can be described as a sharp image convoluted by a PSF, which is usually approximated by a Gaussian function. We assume the defocus blur as follows:

$$\begin{cases} I_i(x,y) = P(x,y) \star g_i(x,y) \\ g_i(x,y) = \frac{1}{2\pi\sigma_i^2} exp[-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma_i^2}] \\ \sigma_i(x,y) = \frac{kf_0[v_i - v_0(x,y)]}{N_s v_0(x,y)}, i = 1, 2, \cdots, n \end{cases}, \tag{5}$$

where i is the serial number of image in the focal stacks and v_i is the image distance of the $No.i$ image. $v_0(x,y)$ is the latent in-focus parts at (x,y) of the image, which is only related to the depth of the scene. $I_i(x,y)$ is the blurred image, modeled as a convolution of the latent sharp image $P(x,y)$ and the Gaussian function map $g_i(x,y)$ of the whole $No.i$ image with different σ_i sizes, which is related to the image coordinates. k is a coefficient measures the defocus blur amount c.

3.2 IDM

We improve Zhuo et al.'s method [18] to select the in-focus parts, which is an effective approach to estimate the amount of spatially varying defocus blur at edge locations. Zhuo et al.'s method [18] can produce a continuous defocus map with less noise, which is quite robust especially in detecting the spatially adjacent in-focus objects. However, it can't distinguish defocus blur between front and back areas. We add a contrast section to compare adjacent images of the focal stacks to verify the spatial adjacent parts.

By propagating the blur amount at edge locations to the entire image based on Gaussian gradient ratio, a full defocus map obtained. Then Canny edge detector is used to perform the edge detection, and calculate the standard deviation of the PSF to obtain the sparse depth map. Finally, we apply the joint bilateral filtering (JBF) [12] to correct defocus estimation errors and interpolate the sparse depth to full defocus map [9,11].

In our method, we use the JBF method [12] to deal with the focal stacks first. And then a threshold is given to dispose the most defocus parts, especially effective on the spatially adjacent objects. A comparation is used to select the most in-focus parts among the neighbor areas. Every image corresponding to one focus ring position, and with which an IDM is fused from the defocus map stacks.

Values on the IDM are labels, which mark the clearest parts on all images in the focal stacks. Need to know that "the best image" is unequal to "the clearest image" because of the discrete sampling illustrated in the INTRODUCTION. Therefore, the sample frequency determines the IDM accuracy. It can be an index to guide the most in-focus parts in the stacks and by which an AiF image is fused. We can calculate the coordinates of the points in world space according to Figs. 2 and 4:

$$(X, Y, Z) = (\frac{-x\Delta f_0}{v - f_0}, \frac{-y\Delta f_0}{v - f_0}, \frac{vf_0}{v - f_0}), \tag{6}$$

where Z is the depth of the scene. Then we reconstruct the scene with these conditions.

4 Experiments and Analysis

In this section, the proposed method is evaluated on the real scene. The main devices we used are Cannon EOS 80D SLR camera with a 135 mm focus, a camera holder whose moving unit is up to 0.01 mm, and a PC to control the images sensor shifts. Besides, two bookends clamped with two printed leopard pictures are used as the mapping calculation targets. A pavilion and a path in Sichuan University in China are chosen for camera mapping calculation and depth sensing respectively.

4.1 Camera Mapping Calculation

A software from the Cannon official website on a PC are used to control the camera shift, reducing shakes during photographing and improving the accuracy of focal plane's movements. Camera mapping calculation is conducted in a pavilion in Sichuan University in China.

We divide the image sensor moving distance into sparse unit displacements and we fix the camera and coaxially photograph at every unit to obtain focal stacks. Top two rows of Fig. 5 show some of the focal stacks. We can see the focal plane move from the camera nearby to the white wall behind including the two bookends. Then we choose the aim area of the front bookend at the same position in the stacks, as the yellow boxes shown in the top row. The aim area is processed with DFT shown in the bottom row of Fig. 5, and we can see it's apparently different in the frequency domain of the images with different sizes of CoC, especially in the high frequency areas. Then we do the same process to the back bookend and the frequency domain maps of the two aim areas are obtained.

We compare the frequency domain of the aim areas in the focal stacks with the Eq. (2), where the higher frequency parts share more weights. Figure 6 left one shows the sharpness of the focal stacks of the front and back bookend planes. Two target objects are combined into one coordinate system to illustrate focal position relations of the two objects, and we know that the peak of sharpness

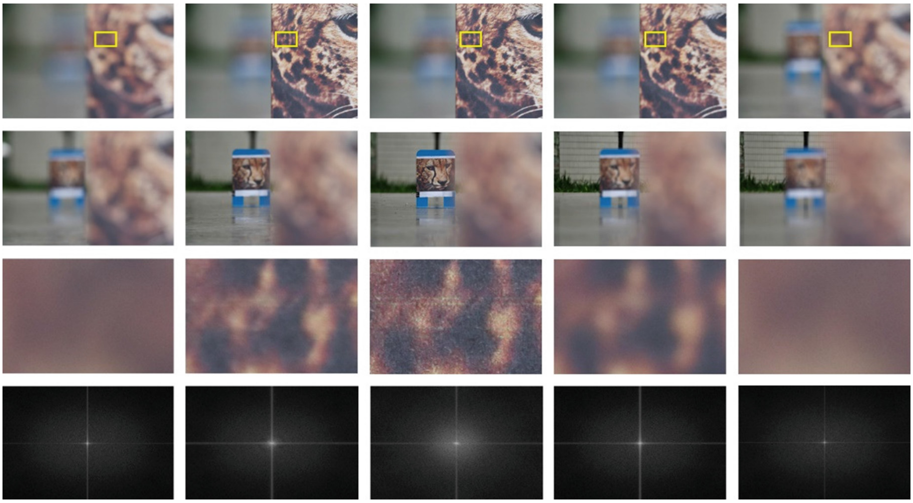

Fig. 5. Top row: several images of the front bookend at different image distances, which are 165.25 mm, 163.35 mm, 161.45 mm, 159.55 mm and 157.65 mm from left to right. Second row: several images of the back bookend at different image distances, which are 144.35 mm, 142.45 mm, 140.55 mm, 138.65 mm and 136.75 mm from left to right. Third row: the responding details boxed out in the top row. Bottom row: the DFT of the third row.

evaluation function curve is single and symmetrical. That is because the diameter of CoC c positively correlates to the distance between focal plane and image plane (see Eq. (1)), which make the blur degree symmetrical about the in-focal image plane.

We get the two sharpest image positions from the curve in Fig. 6 left and photograph the scene at in-focus plane. Shift the camera holder to simulate the parallax method. Images from left and right camera can be seen in Fig. 6 right. The yellow lines connect corresponding points between the two cameras. We calculate image distances of the two bookends with parallax method from Eq. (3) and the depth of the front and back bookends. The two in-focus distances of the two bookends are $v_a = 140.02$ mm and $v_b = 161.77$ mm, and the object distances are $Z_a = 815.72$ mm and $Z_b = 3763.22$ mm. The depth between them that we measured with tape is $d_{ab} = 2900$ mm, which is close to the computation and the accuracy is 98.36%. The linear relation between image plane shift and image distance is $v = -0.38s/\delta + 176.65$, where δ is the unit displacement of the focus ring.

4.2 Depth Sensing

A path scattered with several stones and branches at different depth are chosen as the test scene. Devices are still the Cannon EOS 80D SLR camera with a 135 mm fixed focus and a PC to control the shift of the image plane.

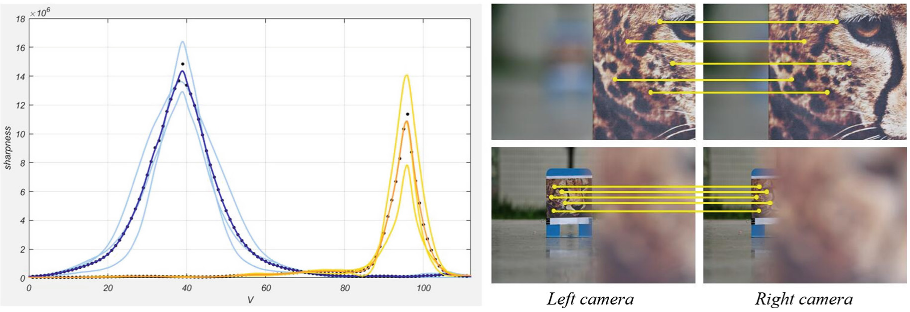

Fig. 6. Left: the sharpness evaluation function curve. Two peaks represent the front and back bookends respectively. The left peak is the front one, and right peak is the behind one. The yellow and light blue lines are some sample points we chosen. The orange and dark blue lines are their mean values respectively. Right: top two are images of the front bookend from right and left cameras at in-focus distance v_a and bottom two are images of the back bookend from right and left cameras at in-focus distance v_b. (Color figure online)

In our experiments, we photograph the scene with the image plane moving far away gradually and some of the results shown in the top row of Fig. 7. The focal stacks are $\{P_i, s_i\}, \forall_i = 1, 2, ..., N$. As shown in the bottom row of Fig. 7, defocus maps of images in focal stacks darken the aim depth areas and light up the adjacent defocus areas. We exclude the defocus light patches and compare with the contiguous defocus map to select the most in-focus parts of the whole image. The final IDM, depth map and the AiF image calculated from the IDM as shown in Fig. 8.

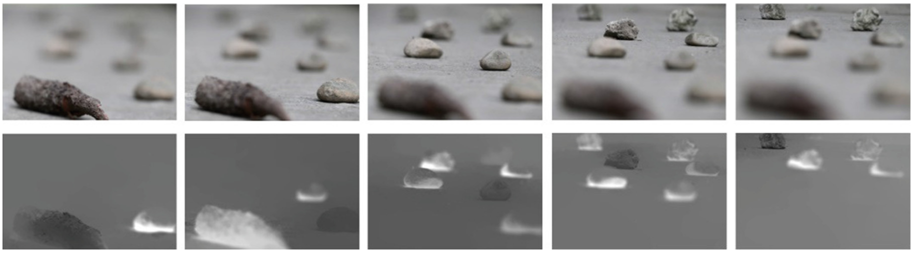

Fig. 7. Top row: samples of focal stacks. Bottom row: our defocus maps of the top row.

Figure 9 shows the experimental scene and the different views of the reconstruction of the scene. We can see the final scene reconstruction illustrates the spatial relationship of the fused image, which is similar to the real scene. Objects close to the camera show more details because of the depth and image distance mapping. Layered reconstructed road is caused by the discrete sampling, but

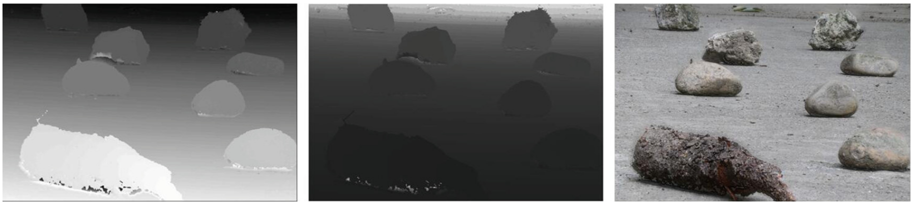

Fig. 8. Left: the IDM. Middle: depth map. Right: defocus map.

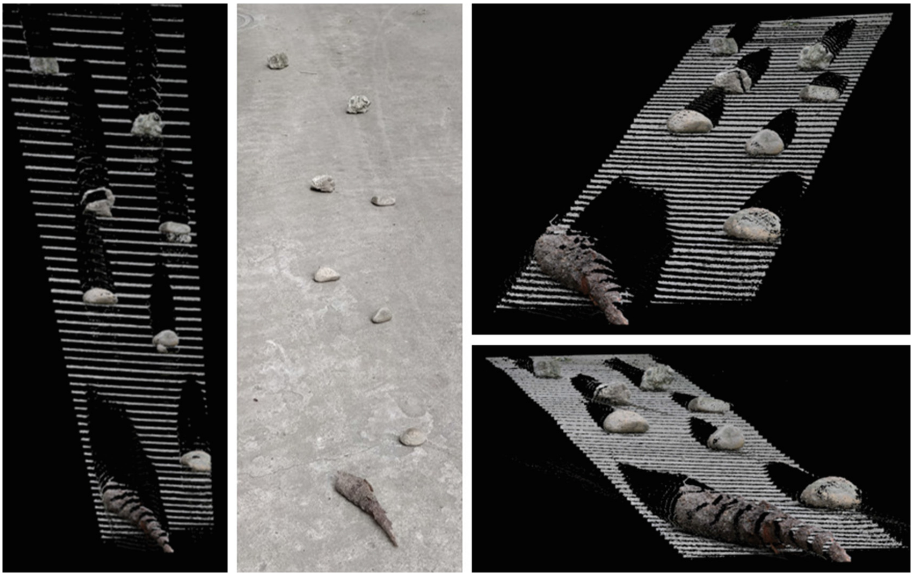

Fig. 9. Middle: the test scene photographed by another camera. Others: reconstructed scene in different directions.

provide a precise distance detection of each object without camera poses. The detection depth is related to the camera features. Under our experimental conditions, the best sensing range is from 0.4 m to 25 m.

To demonstrate the effectiveness of our method, we present contrast experiments and results are shown in Fig. 10. Three methods [6,7,10] and the ground truth generated with RealSense of Intel are contrasted with ours. Compared to the binocular methods, our method provides complete outlines. Compared to the Levin et al.'s method [10] which we experimented on focal stacks, our method provides clear boundaries and more layers from the stacks. Bisides, we calculate the real distance between objects in space compared with the real depth obtained from the RealSense, which shows that depth sensing accuracy of in-focus areas is over 98% in the experimental scope.

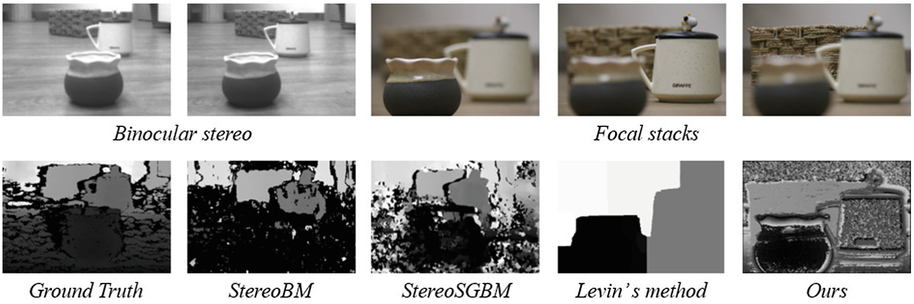

Fig. 10. Top row: the experimental scene photographed via binocular stereo, and the focal stacks of the same scene. Bottom row: depth map of ground truth, three depth sensing method [6,7,10] and ours.

5 Conclusion

In this paper, we proposed a camera pose free depth sensing method, in which the depth information was inferred from differently blurred images captured by aperture camera with focus stacking. DFT sharpness evaluation function and parallax method were employed in the camera mapping calculation. Then we realized depth sensing by fusing an IDM and an AiF image with focus stacking. The experimental results show that the proposed method is robust and accurate for depth sensing compared to the binocular and other focus stacking methods. In the future, we will concentrate on improving spatial continuity of the reconstructed scene, less focal stacks and better calculation algorithm.

References

1. Akira, K., Kiyoharu, A., Tsuhan, C.: Reconstructing dense light field from array of multifocus images for novel view synthesis. IEEE Trans. Image Process. **16**(1), 269 (2007). A Publication of the IEEE Signal Processing Society
2. Alonso, J.R., Fernández, A., Ferrari, J.A.: Reconstruction of perspective shifts and refocusing of a three-dimensional scene from a multi-focus image stack. Appl. Opt. **55**(9), 2380 (2016)
3. Fan, C., Weng, C., Lin, Y., Cheng, P.: Surface profiling measurement using varifocal lens based on focus stacking. In: 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–5, May 2018. https://doi.org/10.1109/I2MTC.2018.8409820
4. Grossmann, P.: Depth from focus. Pattern Recogn. Lett. **5**(1), 63–69 (1987)
5. Gulbins, J., Gulbins, R.: Photographic Multishot Techniques: High Dynamic Range, Super-Resolution, Extended Depth of Field, Stitching. Rocky Nook, Inc., Santa Barbara (2009)
6. Heiko, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2007)
7. Konolige, K.: Small vision systems: hardware and implementation. In: Shirai, Y., Hirose, S. (eds.) Robotics Research, pp. 203–212. Springer, London (1998). https://doi.org/10.1007/978-1-4471-1580-9_19

8. Kuthirummal, S., Nagahara, H., Zhou, C., Nayar, S.K.: Flexible depth of field photography. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 58–71 (2010)
9. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. **23**(3), 689–694 (2004)
10. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 228–242 (2007)
11. Lischinski, D., Farbman, Z., Uyttendaele, M., Szeliski, R.: Interactive local adjustment of tonal values, pp. 646–653 (2006)
12. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M.F., Toyama, K.: Digital photography with flash and no-flash image pairs. ACM Trans. Graph. **23**(3), 664–672 (2004)
13. Sakurikar, P., Narayanan, P.J.: Focal stack representation and focus manipulation. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR) (2017)
14. Xu, G., Quan, Y., Ji, H.: Estimating defocus blur via rank of local patches. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5371–5379 (2017)
15. Zhang, L.X., Sun, H.Y., Guo, H.C., Fan, Y.C.: Auto focusing algorithm based on largest gray gradient summation. Acta Photonica Sinica **42**(5), 605–610 (2013)
16. Zhao, H., Bao, G.T., Wei, T.: Experimental research and analysis of automatic focusing function for imaging measurement. Opt. Precis. Eng. **12**, 531–536 (2004)
17. Zhou, S., Lou, Z., Yu, H.H., Jiang, H.: Multiple view image denoising using 3D focus image stacks. In: Signal & Information Processing (2016)
18. Zhuo, S., Sim, T.: Defocus map estimation from a single image. Pattern Recogn. **44**(9), 1852–1858 (2011)