applied optics

Active depth estimation from defocus using a camera array

TIANYANG TAO,[1,2,3] QIAN CHEN,[1,2,4] SHIJIE FENG,[1,2,3] YAN HU,[1,2,3] AND CHAO ZUO[1,2,3,*]

[1]School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, China
[2]Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China
[3]Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China
[4]e-mail: chenqian@njust.edu.cn
*Corresponding author: surpasszuo@163.com

Received 15 March 2018; revised 19 May 2018; accepted 22 May 2018; posted 22 May 2018 (Doc. ID 326220); published 11 June 2018

This paper introduces an active depth estimation method from defocus using a camera array. High-frequency phase-shifted sinusoidal fringe patterns are projected onto the surface of the object, making low-texture areas of the object surface easily distinguishable. Based on the light field measurement captured by a 5 × 5 camera array, a synthetic aperture refocusing of the fringe images can be realized after the camera array is properly calibrated and rectified. The fringe modulations at different depths are calculated based on the computationally refocused images, which are used as depth cues to reconstruct the 3D shape of the measured object. We implemented some experiments to verify the effectiveness of the proposed method.

OCIS codes: (120.0120) Instrumentation, measurement, and metrology; (150.6910) Three-dimensional sensing; (110.5200) Photography; (150.0150) Machine vision.

https://doi.org/10.1364/AO.57.004960

1. INTRODUCTION

Light field imaging is an emerging technology in computational photography areas. While traditional photography simply captures a 2D projection of the 3D world, light field imaging can record not only a spatial dimension, but also the angular dimension of rays impinging on the image sensor. This enables images to be formed postcapture, where properties such as a viewing perspective, the aperture size, and the focal plane position are varied. The light field can be captured in various ways, for example, camera arrays [1–5], micro-lens arrays [6–8], programmable aperture [9–12], and attenuation mask [13,14] techniques. Benefiting from light field imaging, several interesting imaging modalities and applications have been demonstrated, such as seeing through partly occluding environments [1,2], refocusing at any depth after capturing [8], generalizing high-dynamic and high-resolution images [3,5], creating a virtual high-speed camera [4], and estimating depth information [15].

In contrast to traditional stereo matching, the baseline between the adjacent views in the light field is typically narrow, which makes it difficult to recover accurate disparity from two views based on traditional stereo-matching methods. Therefore, instead of stereo-matching methods, constraints and cues that take advantage of all the views together are used to estimate the depth map with improved accuracy from a light field image [16]. Most depth estimation methods of light field imaging are based on this principle, and one of the most important concepts in these methods is the epipolar plane images (EPI), which was first proposed by Bolles [17]. They stacked all the views from camera motion to get the spatiotemporal solid of data, and then the 2D EPIs could be sliced from this 3D solid of data. The matching points of the object in different views form a straight line in the EPIs, and the slope of the line reflects the depth of these matching points. Bolles *et al.* extracted edges of featured lines in the EPIs to reconstruct the 3D structure of the measured scene [17]. However, the employed basic line fitting is not robust enough for a dense reconstruction. Based on this work, Criminisi *et al.* [18] attained the goal of achieving a dense scene reconstruction by decomposing the scene into EPI tubes, as well as analyzing and removing specular highlights. This problem was also studied in Ref. [19], where an energy minimization framework was introduced for a light field. Using the same principle in Ref. [19], Jeon *et al.* [20] realized the subpixel-level accuracy by combining a phase-shift method. Recently, Wanner and Goldluecke [21] used a structure tensor to compute the vertical and horizontal slopes in EPIs, and they formulated the depth map estimation problem as a global optimization approach that was subject to the epipolar constraint. Lin *et al.* [22] made use of the geometric constraints of 3D lines to further improve the reconstruction quality.

Most EPI-based algorithms need several pre-processing steps and involve a global optimization to obtain sufficiently smooth results. It is often challenging to scale such approaches to a significantly high-resolution image due to a very high computational cost of global optimization. The second difficulty of approaches based on global optimization is how to avoid error propagation. In addition, the EPI-based methods are only robust in high-texture areas but fail in weak-texture cases.

Besides the correspondence cue in EPIs, some other cues, for example, the defocus cue [23–25] and the shading cue [26,27], have also been applied for depth extraction. The depth from defocus is the prominent one, benefitting from the property of the light field. The light field image can be refocused on any area by integrating the light field along an arbitrary angular or performing an inverse fourier transform with the slice of 4D fourier spectrum of the light field [28]. After refocusing at a new depth, the light rays from any in-focus scene point will be recorded in the same spatial location. We can refocus the light field function at different depth candidates. The absolute difference between the central view and its angular views reflects the depth probability. Then, using the operator such as the Laplacian operator, we can extract the defocus degree of all candidates, and the depth can be derived from the position of minimum defocus (maximum focus) degree.

However, out-of-focus regions, such as certain high-frequency regions and bright lights, may yield higher contrast. The ambiguities in maximum focus degree may cause mistakes of depth estimation. Another problem is how to select the size of analyzed patch (or windowed operator) considering that the size of the defocus blur is difficult to be qualified. In most cases, the defocus cue is used as an additional cue to enhance the robustness of an EPI-based method [25,27].

The purpose of this paper is to introduce a novel defocus-based depth estimation method that uses active illumination. The depth estimation combining active illumination is rarely referred in light field imaging, and the existing active-illumination-based works tend to just mark the weak-texture areas with some regular or irregular patterns to perform the same subsequent algorithms as passive methods. In this work, the sinusoidal fringe pattern is provided not only to project artificial textures onto the surface of the object, but also to help decrease the computational complexity. We analyze the relationship between the defocus degree and the modulation of the sinusoidal fringe of the captured images. Compared with the available EPI-based and active method, there are three main contributions in our work: (1) our method is a pixel-wise method where we can concurrently detect the defocus degree of each pixel. Thus, there is no error propagation, and the most computational task can be dealt with a graphics-processing unit. (2) Replacing the traditional defocus cues with modulation, no window-based defocus operator is required in our method so that we do not need to consider the problem of the size of window operator and the subsequent trade-off between the spatial resolution and the window size. (3) The precision of displacement can be refined by weighted average or the Gaussian curve fitting due to the regular relationship between the modulation variance and the defocus degree.

2. PRINCIPLE

The main principle is organized as follows. The three-step stereo rectification is first presented in Section 2.A.1, followed by the image rectification for a camera array in Section 2.A.2. Based on the rectified images, the process of digital refocusing is introduced in Section 2.B, and then in the same section, we extract the defocus cues from these refocusing images at different depths by detecting the maximum modulation of phase-shifting fringes. The defocus cues are finally refined by interpolation to improve the depth accuracy; this is depicted in Section 2.C.

A. Rectification of Camera Array

The light field in our work is captured by a camera array (Profusion 25 M). Considering the manufacturing error of the camera array, an image rectification is inevitable before the refocusing process. The rectification for the camera array should simultaneously align the matching points from horizontally and vertically distributed cameras to the same rows or columns, which is an extended version from conventional stereo rectification [29,30] where only the row or column alignment is required. In this section, we first introduce the stereo rectification and then apply the modified rectification to a camera array.

1. Three-Step Stereo Rectification

In a stereo system, an arbitrary point $P = [X, Y, Z]^T$ can be mapped into the left and right camera coordinate systems by the following equations:

$$P_l = R_l P + T_l, \qquad P_r = R_r P + T_r, \tag{1}$$

where R_l, R_r, T_l, and T_r represent the rotation and translation matrices from the world coordinate system to the left and right camera coordinate systems, respectively. Point P is identified in the left and right camera coordinate systems as $P_l = [X_l, Y_l, Z_l]^T$ and $P_r = [X_r, Y_r, Z_r]^T$, respectively. In this paper, we use the subscript l and r to distinguish the variables for the left camera and the right camera. The projections of P_l and P_r in the 2D pixel coordinate systems are denoted as $p_l = [u_l, v_l]^T$ and $p_r = [u_r, v_r]^T$, and

$$u_l = \alpha_l \frac{X_l}{Z_l} + u_l^c,$$
$$v_l = \beta_l \frac{Y_l}{Z_l} + v_l^c,$$
$$u_r = \alpha_r \frac{X_r}{Z_r} + u_r^c,$$
$$v_r = \beta_r \frac{Y_r}{Z_r} + v_r^c, \tag{2}$$

where α and β are scale factors in the u and v axes, and (u_l^c, v_l^c) and (u_r^c, v_r^c) are the coordinates of principle points. The so-called stereo rectification suggests that we should have $v_l = v_r$ (for the horizontally distributed cameras) or $u_l = u_r$ (for the vertically distributed cameras). In this subsection, the rectification for $v_l = v_r$ is introduced. $\alpha_l = \alpha_r = \beta_l = \beta_r$, and $v_l^c = v_r^c$ can be obtained by replacing one of them with another. However, Y_l/Z_l and Y_r/Z_r are variables, so to make sure that an arbitrary point P meets $v_l = v_r$, we must have

$Y_l/Z_l = Y_r/Z_r$. This is the most important process of rectification.

As shown in Fig. 1(a), blue coordinate systems represent the initial camera coordinate systems, where O_l and O_r are the centers of lenses, and $X_lO_lY_l$ and $X_rO_rY_r$ are the planes of lenses. The relative rotation R_{lr} and translation difference T_{lr} between the left and right camera coordinate systems satisfy the following relationship:

$$P_r = R_{lr}(P_l - T_{lr}). \tag{3}$$

Combining Eqs. (1) and (3), we can obtain $R_{lr} = R_rR_l^{-1}$, as well as $T_{lr} = T_l - R_{lr}^{-1}T_r$. However, $P_l = [X_l, Y_l, Z_l]^T$ and $P_r = [X_r, Y_r, Z_r]^T$ cannot be derived from $p_l = [u_l, v_l]^T$ or $p_r = [u_r, v_r]^T$ separately, but only their normalized coordinates $P'_l = [X_l/Z_l, Y_l/Z_l, 1]^T$ and $P'_r = [X_r/Z_r, Y_r/Z_r, 1]^T$ can be obtained. The following rectification process can be further divided into three steps. Let the right camera coordinate system be the reference coordinate system. In the first step, the coordinate system of the left camera should be rotated around O_l to become parallel to that of the right camera. Then we update P'_l and $T_{lr} = [\Delta X_{lr}, \Delta Y_{lr}, \Delta Z_{lr}]^T$ by

$$\begin{aligned} P'_l &\Leftarrow R_{lr}P'_l, \\ T_{lr} &\Leftarrow R_{lr}T_{lr}, \end{aligned} \tag{4}$$

where $\Leftarrow$ is an assignment operator. Here, only a translation T_{lr} still exists between the two camera coordinate systems. The second step intends to eliminate the difference between Y_l and Y_r by rotating the two camera coordinate systems around their Z_l and Z_r axes with θ_Z, respectively. θ_Z can be derived from $\tan\theta_Z = \Delta Y_{lr}/\Delta X_{lr}$, and its 3×3 rotation matrix is R_Y. Then Eq. (4) is rewritten as

$$\begin{aligned} P'_l &\Leftarrow R_ZR_{lr}P'_l, \\ P'_r &\Leftarrow R_ZP'_r, \\ T_{lr} &\Leftarrow R_ZR_{lr}T_{lr}. \end{aligned} \tag{5}$$

This step is shown in Fig. 1(b), and the result is shown in Fig. 1(c), where we can easily find $Y_l = Y_r$. To make $Y_l/Z_l = Y_r/Z_r$, we should further align Z_l and Z_r by rotating the two camera coordinate systems around their Y_l and Y_r axes with θ_Y in the third step, as shown in Fig. 1(c). Here, $\tan\theta_Y = \Delta Z_{lr}/\Delta X_{lr}$, and its 3×3 rotation matrix is R_Y. P'_l and P'_r are rewritten as

$$\begin{aligned} P'_l &\Leftarrow R_YR_ZR_{lr}P'_l, \\ P'_r &\Leftarrow R_YR_ZP'_r, \\ T_{lr} &\Leftarrow R_YR_ZR_{lr}T_{lr}. \end{aligned} \tag{6}$$

Figure 1(d) displays the result of the third step of rectification. By dividing P'_l and P'_r by $P'_l(3)$ and $P'_r(3)$, respectively, $Y_l/Z_l = Y_r/Z_r$ is established, so does $v_l = v_r$.

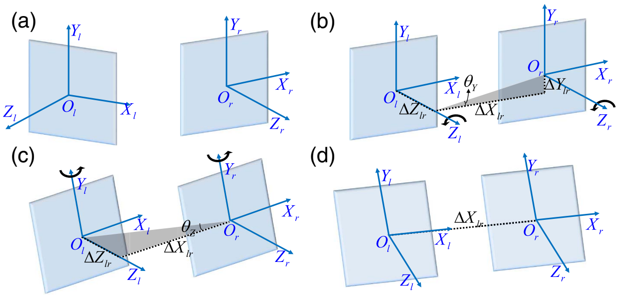

Fig. 1. Processes of stereo rectification. (a) The initial left and right camera coordinate systems. (b) The two camera coordinate systems after the left camera coordinate system being rotated by R_{lr}. (c) The two camera coordinate systems after being rotated by R_{lr}, as well as R_Z. (d) The two camera coordinate systems after being sequentially rotated by R_{lr}, R_Z, and R_Y.

2. Rectification for a Camera Array

It should be noted that in the previous stereo rectification, the rotation transforms are implemented on the two camera coordinate systems. We can not directly apply this rectification to the case of a camera array since, at most, four stereo rectification processes, including both horizontal and vertical rectifications, are required. The scheme to simultaneously rectify horizontally and vertically distributed cameras is proposed in Ref. [31]. But this scheme cannot be used directly when more than three cameras are used. By calculating a common baseline of multiple cameras, the rectification methods for N cameras are also available [32,33]. However, the N cameras in these methods must be arranged in a line. Fortunately, the camera array used in this paper has already been roughly rectified. This is a useful preposing configuration. Let us return to Eq. (4). There exists only a translation between the two cameras after rotating the left camera coordinate system. All the points in these two coordinate systems satisfy the following relationships:

$$\begin{aligned} X_l - \Delta X_{lr} &= X_r, \\ Y_l - \Delta Y_{lr} &= Y_r, \\ Z_l - \Delta Z_{lr} &= Z_r. \end{aligned} \tag{7}$$

The so-called rough rectification means that ΔY_{lr} (ΔX_{lr}) and ΔZ_{lr} between the horizontally (vertically) distributed cameras can be ignored when Z_l is large enough. Keep in mind that when we select the top right camera to be the referenced camera (shown in Fig. 2), then the relationships between the camera coordinate systems of this referenced camera and other horizontal distributed cameras can be written as the following approximate formulas

$$\begin{aligned} X_{1j} - \Delta X_{1j} &= X_{11}, \\ Y_{1j} \approx Y_{1j} - \Delta Y_{1j} &= Y_{11}, \\ Z_{1j} \approx Z_{1j} - \Delta Z_{1j} &= Z_{11}. \end{aligned} \tag{8}$$

The subscripts $1j$ and 11 denote the top row and top right of the camera array shown in Fig. 2, respectively. Based on Eq. (8), we have $Y_{1j}/Z_{1j} \approx Y_{11}/Z_{11}$, which suggests the feasibility of rectification just using Eq. (4). For the vertically distributed cameras, we have similar formulas:

$$\begin{aligned} X_{i1} \approx X_{i1} - \Delta X_{i1} &= X_{11}, \\ Y_{i1} - \Delta Y_{i1} &= Y_{11}, \\ Z_{i1} \approx Z_{i1} - \Delta Z_{i1} &= Z_{11}. \end{aligned} \tag{9}$$

$X_{i1}/Z_{i1} \approx X_{11}/Z_{11}$ can be easily derived from Eq. (9). The general relationships between the referenced camera and an arbitrary camera among the others can be formed as

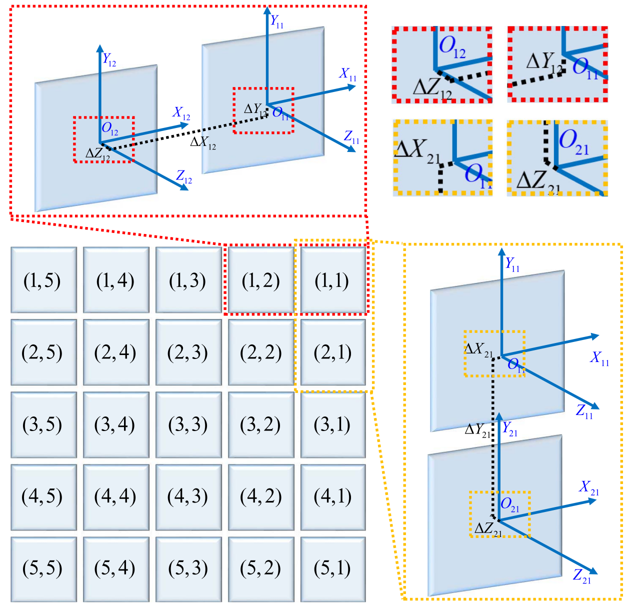

Fig. 2. Diagram of camera array used in our work where the number coordinates represent different cameras in the array. The red and yellow dotted rectangles display the related positions of horizontally distributed cameras and vertically distributed cameras, respectively.

$$\begin{aligned} X_{ij} - \Delta X_{1j} &= X_{11}, \\ Y_{ij} - \Delta Y_{i1} &= Y_{11}, \\ Z_{ij} \approx Z_{ij} - \Delta Z_{ij} &= Z_{11}. \end{aligned} \tag{10}$$

Now it is verified that Eq. (4) is sufficient enough to rectify all the cameras of a roughly rectified camera array. Combining Eqs. (2) and (10), we have

$$\begin{aligned} \Delta u_{ij} &= u_{ij} - u_{11} = \alpha_{11} \frac{\Delta X_{1j}}{Z_{11}}, \\ \Delta v_{ij} &= v_{ij} - v_{11} = \alpha_{11} \frac{\Delta Y_{i1}}{Z_{11}}, \end{aligned} \tag{11}$$

where Δu_{ij} and Δv_{ij} represent the pixel translations between matching points. Substituting $i = 1$ and $j = 2$ into Eq. (11), we can obtain $\Delta u_{12} = \alpha_{11} \Delta X_{12}/Z_{11}$, and they can be transformed into $\alpha_{11}/Z_{11} = \Delta u_{12}/\Delta X_{12}$. Substituting them into Eq. (11), we will get Eq. (12):

$$\begin{aligned} \Delta u_{ij} &= \Delta u_{12} \frac{\Delta X_{1j}}{\Delta X_{12}}, \\ \Delta v_{ij} &= \Delta u_{12} \frac{\Delta Y_{i1}}{\Delta X_{12}}. \end{aligned} \tag{12}$$

B. Depth Estimation from Defocus

For an arbitrary point in space, its light rays spread around and can be captured from different views. One camera can only gather the rays within a small range of angles due to its limited lens size. If the point is located at the object plane, these rays will converge at one point in the focal plane, and we can get an enhanced image; otherwise, the rays will spread to a diffuse spot, and a blurry image will be obtained. Since this imaging process is not irreversible, only an intensity map is finally available for us. However, we need some other useful information, such as the angle of the light rays, to derive the depth of this space point. This useful information cannot be accessible by using one camera, but it can be obtained by using a camera array. The camera array will capture light rays from a wider range of angles, and more importantly, these light rays from different angles are reserved in different images by the cameras with different views. Using these images, we can realize digital refocusing, extract the angle information, and even estimate the depth map. This method just supplies an easy way to search the corresponding points or extract the angle information from these images.

Digital refocusing is realized by translating and adding the images from different cameras. If we select camera (1, 1) in Fig. 2 to be the referenced camera, the refocused image can be formulated as

$$I(\Delta u_{12}) = \sum_{j=1}^{5} \sum_{i=1}^{5} I_{ij} \left(u - \frac{\Delta X_{1j}}{\Delta X_{12}} \Delta u_{12}, v - \frac{\Delta Y_{i1}}{\Delta X_{12}} \Delta u_{12} \right), \tag{13}$$

where $\Delta X_{11} = 0$ and $\Delta Y_{11} = 0$, and I_{ij} represents the rectified images. Equation (11) gives the relationship between the pixel translation Δu_{12} and the depth Z_{11} as

$$Z_{11} = \alpha_{11} \frac{\Delta X_{12}}{\Delta u_{12}}. \tag{14}$$

Equations (13) and (14) suggest that for the given Δu_{12}, the point at the depth of Z_{11} will be focused, while the points at other depths will be out of focus. We record the defocus (or focus) degree of each point of $I(\Delta u_{12})$ for a series of Δu_{12}. Then, we detect the minimum (or maximum) value of the defocus (or focus) degree of each point and finally derive the depth from the minimum (or maximum) defocus cues. However, the essential problem in this depth-estimation method is how to accurately detect pixel-wise defocus cues, especially the point located in a low-texture area.

Active illumination can help distinguish the defocus degree at different translation Δu_{12}, even for low-texture areas. However, in the available depth estimation method, the illuminated pattern is confined to paste textures on the object surface, and the subsequent process still relies on a windowed analysis rather than a pixel-based analysis. This paper introduces a pixel-wise depth estimation method based on defocus cues by projecting sinusoidal pattern onto the objects. The three-step phase-shifting sinusoidal patterns captured by the camera can be formulated as

$$\begin{aligned} I_{ij}^1 &= A + B \cos \Phi, \\ I_{ij}^2 &= A + B \cos(\Phi + 2\pi/3), \\ I_{ij}^3 &= A + B \cos(\Phi + 4\pi/3), \end{aligned} \tag{15}$$

where I_{ij}^1, I_{ij}^2, and I_{ij}^3 represent the image intensities, A is the average intensity, B is the fringe contrast, and Φ the modulated phase. Φ is a primary parameter to derive the accurate depth in fringe-projection profilometry, but it is not suitable for implicating a defocus degree due to its irregular ambiguities, as shown in Fig. 3. Besides Φ, we can easily deduce modulation information M by Eq. (16).

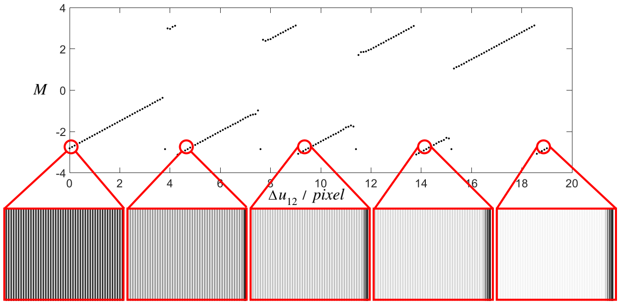

Fig. 3. Diagram of the relationship between the phase and defocus degree, where the dotted curve represents the phase variance along with Δu_{12}; red circles are the regions with the phase nearing –2.8, and the images in red rectangles are the corresponding refocused images I^1.

$$M = \frac{\sqrt{\sum_{m,n=1}^{3} (I^m(\Delta u_{12}) - I^n(\Delta u_{12}))^2}}{3}, \quad m \neq n. \tag{16}$$

There is a regular relationship between the defocus degree and modulation M [34]. The larger defocus degree corresponds to the smaller M in a pre-defined interval, which can be easily understood in the frequency domain. We make a simulation to conveniently explain this phenomenon. In this simulation, the sinusoidal wavelength is 19 pixels, and the refocused image is a fused version of five sinusoidal images with different pixel translations. The varying curve of modulation M of the left top point in refocused image is shown in Fig. 4. The greatest focus degree with $\Delta u_{12} < 19$ emerges at $\Delta u_{12} = 0$, where its modulation M is the largest. However, due to the periodicity of sinusoidal wave, there are some other maximum peaks of modulation M when Δu_{12} is integral multiples of the fringe period 19. This is an inherent problem of periodic textures. A straightforward strategy to relieve this problem is to extend the sinusoidal wavelength, as shown in Figs. 5(a) and 5(c). But from Figs. 5(b) and 5(d), it can be found that if we add some noise to the fringe patterns, this strategy will make the modulation peaks become unconspicuous. The noise is inevitable in the experiment so that simply extending the wavelength will cause some errors in detecting the maximum M. There is a trade-off between the quality and the ambiguity of the refocused images. Considering the focus volume of the projector, we restrict the scene to be measured at the depth range $Z_{11} \in [Z_{\min}, Z_{\max}]$, and then Δu_{12} should satisfy

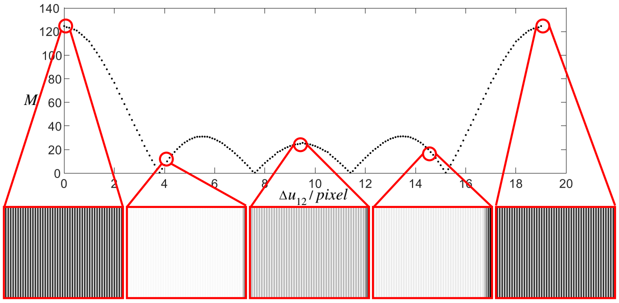

Fig. 4. Diagram of the relationship between modulation and the defocus degree, where the dotted curve represents the modulation variance along with Δu_{12}, red circles are the regions with different modulation, and the image in red rectangles are the corresponding refocused images I^1.

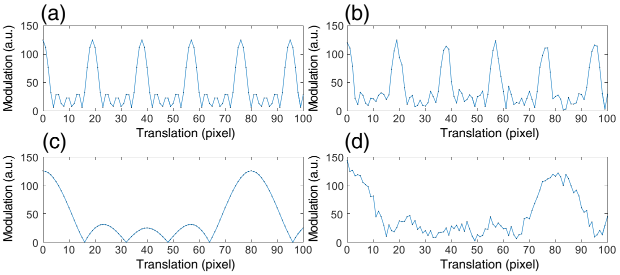

Fig. 5. Diagram of the varying curve of M of different wavelengths along with Δu_{12}. (a) The varying curve of M for 19-pixel wavelength. (b) The varying curve of M for 19-pixel wavelength by adding Gaussian noise. (c) Varying curve of M for 80-pixel wavelength. (b) Varying curve of M for 80-pixel wavelength by adding Gaussian noise.

$$\alpha_{11} \frac{\Delta X_{12}}{Z_{\max}} \leq \Delta u_{12} \leq \alpha_{11} \frac{\Delta X_{12}}{Z_{\min}}. \tag{17}$$

In this depth volume, the smallest wavelength is set to $\alpha_{11} \Delta X_{12} \left(\frac{1}{Z_{\min}} - \frac{1}{Z_{\max}}\right)$ to eliminate the refocused ambiguity. The optimal wavelength is selected around the small interval of $\alpha_{11} \Delta X_{12} \left(\frac{1}{Z_{\min}} - \frac{1}{Z_{\max}}\right)$. Taking the optimal wavelength, we can easily detect the maximum focus degree and its corresponding Δu_{12} of each point in the referenced image, and finally estimate the depth from Eq. (14).

C. Interpolation of Displacement

It is suggested in Eq. (14) that the precision of estimated depth Z_{11} depends on the precision of translation Δu_{12}. Supposing Δs is the varying step of Δu_{12}, then the varying rate ΔZ_{11} of depth can be derived combining Eq. (14):

$$\Delta Z_{11} = \alpha_{11} \frac{\Delta X_{12}}{\Delta u_{12}} - \alpha_{11} \frac{\Delta X_{12}}{\Delta u_{12} + \Delta s} \approx \alpha_{11} \frac{\Delta X_{12} \Delta s}{\Delta u_{12}^2}. \tag{18}$$

We can decrease Δs to make the variance of Z_{11} more successive. However, limited to the precision of sub-pixel translation peak and computational cost of multiple refocusing processes, Δs can not be decreased limitlessly. In this paper, we set $\Delta s = 0.2$ pixel and explore a simple strategy to refine Δu_{12}. Benefitting from the regular relationship between M and Δu_{12} in Fig. 4, all the value of M as well as Δu_{12} near the maximum peak can be used to refine Δu_{12} by polynomial fitting or weighted average. In this paper, the final translation $\Delta \bar{u}_{12}$ is refined by

$$\Delta \bar{u}_{12} = \frac{\sum_{\Delta u_{12} = \Delta u_{12}^m - 3}^{\Delta u_{12} = \Delta u_{12}^m + 3} M \Delta u_{12}}{\sum_{\Delta u_{12} = \Delta u_{12}^m - 3}^{\Delta u_{12} = \Delta u_{12}^m + 3} M}, \tag{19}$$

where Δu_{12}^m is the translation corresponding to maximum M. The value of M is used as weighted coefficient for corresponding Δu_{12}. The translation Δu_{12} is refined by making full use of the information near the maximum M.

3. EXPERIMENTS

To verify the feasibility of the proposed method, we conducted some experiments for depth estimation using a camera array (Profusion 25M) and a projector (LightCrafter 4500). The depth volume in these experiments is $Z_{11} \in [250, 450]$, and the selected wavelength of the fringes is 19 pixels. All the cameras of Profusion 25 M are calibrated by [35], and we got the parameters $\alpha_{11} = 909$ pixels, as well as $\Delta X_{12} = 12$ mm.

A. Rectification Evaluation

The rectification is required to pre-process the raw before refocusing process. Here the rectification is executed by the method described in Section 2.A.2. Figure 6 displays the rectified results. The corner point selected by red rectangles has been captured by all the cameras, so it can be regarded as the matching point to evaluate the effectiveness of the rectification by detecting if it is located in the same line. The top line in Fig. 6 tends to verify the horizontal rectification, and the effectiveness of vertical rectification is indicated by the right line in Fig. 6. The rectangle regions are enlarged; thus, we can find that the both the matching points in the top row and the right column are rectified in the same line.

B. Refocusing and Depth Estimation Experiment

The digital refocusing is implemented according to Eq. (16) with the rectified images. A Kitty plaster is first measured, and the result is shown in Fig. 7. We tracked the modulation variance of three points labeled with the red, green, as well as the blue color to display the relationship between maximum modulation and different depths. The related results are shown in Figs. 7(b)–7(e), where different depths correspond to conspicuously different translation Δu_{12}. Note that the maximum value of the three R, G, B curves is different because of a different reflectivity of the measured surface and a different degree of projector defocus. The defocus degree of the object itself is difficult to detect using passive methods, but it can be actually qualified with the help of easily estimated resorting to sinusoidal patterns.

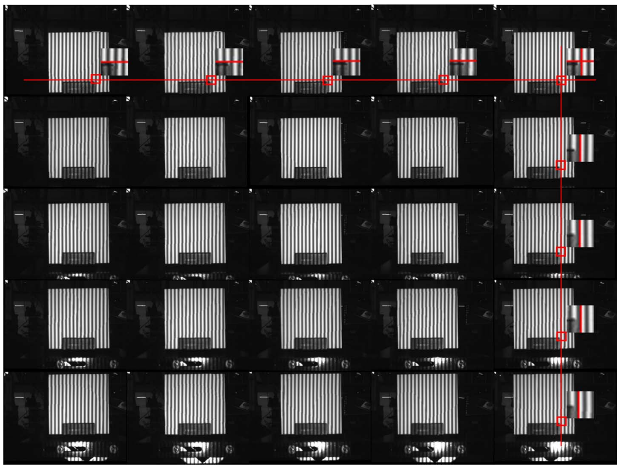

Fig. 6. Results of rectification, where the two red lines represent the referenced straight line to detect if the matching points are rectified at a line, and the sub-images at the top and right line are the enlargement images of the red rectangles.

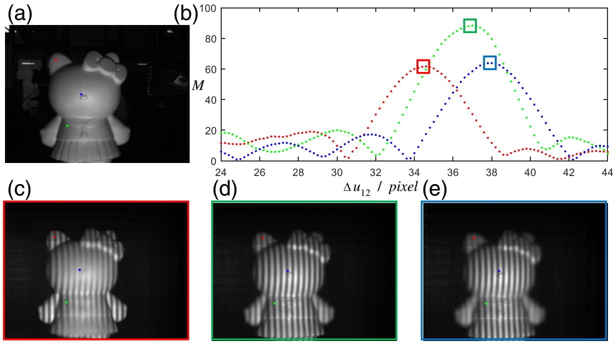

Fig. 7. Refocused images with different translation Δu_{12}. (a) Measured object captured by the referenced camera. (b) Modulation curves for the red, green, and blue points in (a). (c)–(e) Refocused images at the maximum modulation of the curves in (b).

To synthetically evaluate the proposed method, three different objects shown in Figs. 8(a)–8(c) are measured. The depths of all these three objects are difficult to accurately estimate due to the weak textures. For example, the plate in Fig. 8(c) may be roughly regarded as a whole with the same depth by graph cut. However, benefitting from the sensitivity of modulation to the defocus degree, the proposed method can get accurate depths shown in Figs. 8(d)–8(f). The regular relationship between modulation and pixel translation gives us the chance to make full use of the redundant information to refine the pixel-translation map, as well as the depth map. Figures 8(g)–8(l) display the great improvement by using the interpolation algorithm descibed in Eq. (19).

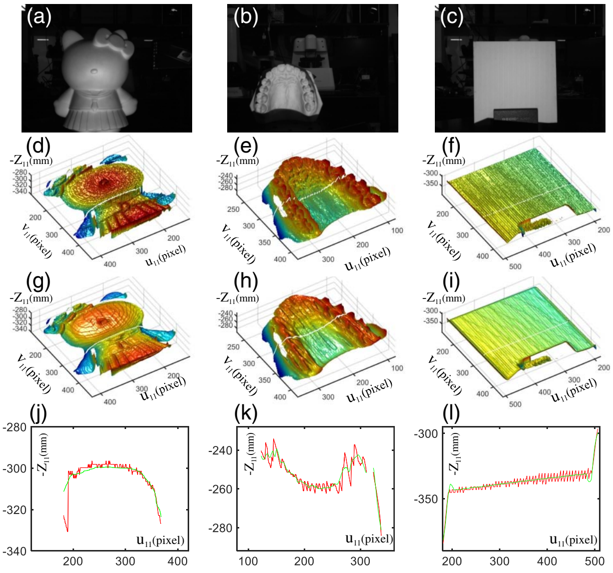

Fig. 8. Depth estimation results of three measured objects. (a)–(c) Three objects to be measured. (d)–(f) Depth maps without interpolation described in Eq. (19). (g)–(i) Depth with interpolation in Eq. (19). (j)–(l) Detailed depth information corresponding to the line labeled with the white color in (d)–(i), the red curves come from (d)–(f), and the green curves from (g)–(i).

The depth of Kitty plaster was also estimated by a windowed Fourier transform (WFT) method to make a comparison. We first selected an image with the focus on the arms of the Kitty, as shown in Fig. 9(a), to extract its modulation with WFT. Figures 9(b) and 9(c) indicate that the WFT with the window size of 20 pixels (approximately equal to the wave length of the captured fringes) can obtain the more accurate modulation where the focusing areas and that the defocusing areas can be distinguished more clearly. Then we used WFT with a 20-pixel window size to estimate the depth map, and the results are displayed in Figs. 9(d) and 9(e). By comparing the results in Fig. 9 with that in Fig. 8, we can easily find that only a rough depth map can be obtained by using WFT, especially in the edge areas. Another point that should be noted is that the WFT is very time-consuming.

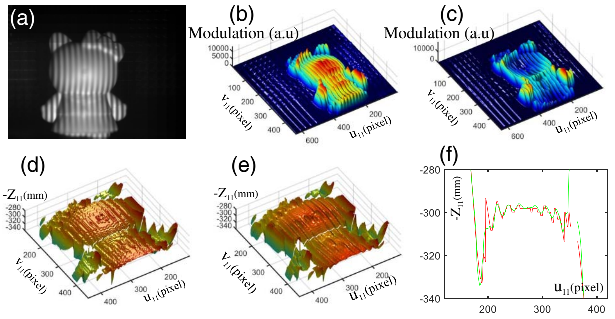

Fig. 9. Object to be analyzed by WFT. (b)–(c) Modulation maps obtained by WFT with the window size of 10 pixels and 20 pixels, respectively. (d) 3D result derived from WFT without interpolation. (e) 3D result derived from WFT with interpolation. (f) Profiles corresponding to the line labeled with the white color in (d)–(e); the red comes from (d), and the green comes from (e).

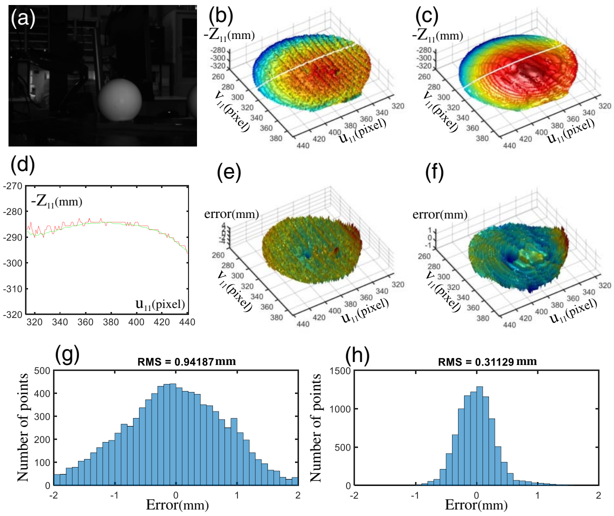

Fig. 10. Comparison between the 3D result without interpolation and with interpolation. (a) Measured object. (b) 3D result without interpolation. (c) 3D result with interpolation. (d) Profiles corresponding to the line labeled with the white color in (b)–(c), the red comes from (b), and the green comes from (c). (e)–(f) 3D error corresponding to (b) and (c), respectively. (g)–(h) Histograms corresponding to (e) and (f), respectively.

To further analyze the improvement of the interpolation, a ceramic ball with the radius of 25.4 mm and the nominal accuracy of ±5 μm were measured, as shown in Fig. 10(a). The depth estimation results are shown in Figs. 10(b) and 9(c). From Fig. 10(d), we can find the prominent improvement of the interpolation. Since the parameters of the ceramic ball have been known, the quantitative analysis is able to be implemented. We first calculated the X_c and Y_c coordinates using Z_{11}, then the ground truth of this ceramic ball was derived from the sphere fitting based on the measured 3D data. The differences between the ground truth and the measured data shown in Figs. 10(e) and 10(f) and Figs. 10(g) and 10(h) give the statistical information where we can easily find the quantitative improvement of the interpolation.

4. CONCLUSION

We have realized active depth estimation from defocus cues. The pixel-wise modulation can be easily extracted from three-step phase-shift fringes without any local analysis algorithms. Based on the regular relationship between modulation and defocus cues, we can accurately estimate and refine defocus cues as well as the depth without error propagation. However, the drawback of this method is that three fringe images are required to obtain the fringe modulation, which means the accuracy of this method is inferior to some other methods for dynamic scenes. Replacing the phase-shift algorithm with a Fourier transform [34] may be a feasible solution to further reduce the number of patterns.

Funding. National Key Technologies R&D Program of China (2017YFF0106403); National Natural Science Foundation of China (NSFC) (111574152, 61705105, 61722506); Final Assembly 13th Five-Year Plan Advanced Research Project of China (30102070102); Outstanding Youth Foundation of Jiangsu Province of China (BK20170034); National Defense Science and Technology Foundation of China (0106173); Six Talent Peaks project of Jiangsu Province, China (2015-DZXX-009); 333 Engineering research project of Jiangsu Province, China (BRA2016407); Fundamental Research Funds for the Central Universities (30916011322, 30917011204); Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (3091601410414); China Postdoctoral Science Foundation (2017M621747); Jiangsu Planned Projects for Postdoctoral Research Funds (1701038A).

REFERENCES

1. V. Vaish, G. Garg, E.-V. Talvala, E. Antunez, B. Wilburn, M. Horowitz, and M. Levoy, "Synthetic aperture focusing using a shear-warp factorization of the viewing transform," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (CVPR)* (IEEE, 2005).
2. 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2006), Vol. **2**, pp. 2331–2338.
3. K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "PiCam: an ultra-thin high performance monolithic camera array," ACM Trans. Graph. **32**, 166 (2013).

4. B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR* (IEEE, 2004), Vol. **2**, pp. II.
5. B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM Transactions on Graphics* (ACM, 2005), Vol. **24**, pp. 765–776.
6. Lytro, https://www.lytro.com/.
7. Raytrix, "3D light field camera technology," http://www.raytrix.de/.
8. R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. Rep. CTSR 2005-02 (Stanford, 2005), pp. 1–11.
9. C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, "Programmable aperture photography: multiplexed light field acquisition," in *ACM Transactions on Graphics (TOG)* (ACM, 2008), Vol. **27**, p. 55.
10. C.-K. Liang, G. Liu, and H. H. Chen, "Light field acquisition using programmable aperture camera," in *IEEE International Conference on Image Processing (ICIP)* (IEEE, 2007), Vol. **5**, pp. V-233.
11. H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. K. Nayar, "Programmable aperture camera using LCos," in *European Conference on Computer Vision* (Springer, 2010), pp. 337–350.
12. C. Zuo, J. Sun, S. Feng, M. Zhang, and Q. Chen, "Programmable aperture microscopy: a computational method for multi-modal phase contrast and light field imaging," Opt. Laser Eng. **80**, 24–31 (2016).
13. K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," ACM Trans. Graph. **32**, 46 (2013).
14. A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing," ACM Trans. Graph. **26**, 69 (2007).
15. E. H. Adelson and J. Y. Wang, "Single lens stereo with a plenoptic camera," IEEE Trans. Pattern Anal. Mach. Intell. **14**, 99–106 (1992).
16. G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: an overview," IEEE J. Sel. Top. Signal Process. **11**, 926–954 (2017).
17. R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: an approach to determining structure from motion," Int. J. Comput. Vis. **1**, 7–55 (1987).
18. A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," Comput. Vis. Image Underst. **97**, 51–85 (2005).
19. T. E. Bishop and P. Favaro, "The light field camera: extended depth of field, aliasing, and superresolution," IEEE Trans. Pattern Anal. Mach. Intell. **34**, 972–986 (2012).
20. H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1547–1555.
21. S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012), pp. 41–48.
22. H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3451–3459.
23. M. Watanabe and S. K. Nayar, "Rational filters for passive depth from defocus," Int. J. Comput. Vis. **27**, 203–225 (1998).
24. Y. Y. Schechner and N. Kiryati, "Depth from defocus vs. stereo: how different really are they?" Int. J. Comput. Vis. **39**, 141–162 (2000).
25. M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 673–680.
26. L. Kontsevich, A. Petrov, and I. Vergelskaya, "Reconstruction of shape from shading in color images," J. Opt. Soc. Am. A **11**, 1047–1052 (1994).
27. M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1940–1948.
28. R. Ng, "Fourier slice photography," in *ACM Transactions on Graphics (TOG)* (ACM, 2005), Vol. **24**, pp. 735–744.
29. R. I. Hartley, "Theory and practice of projective rectification," Int. J. Comput. Vis. **35**, 115–127 (1999).
30. A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," Mach. Vis. Appl. **12**, 16–22 (2000).
31. F. Kangni and R. Laganiere, "Projective rectification of image triplets from the fundamental matrix," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2006), Vol. **2**, p. II.
32. Y.-S. Kang, C. Lee, and Y.-S. Ho, "An efficient rectification algorithm for multi-view images in parallel camera array," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video* (IEEE, 2008), pp. 61–64.
33. J. Yang, Z. Ding, F. Guo, and H. Wang, "Multiview image rectification algorithm for parallel camera arrays," J. Electron. Imaging **23**, 033001 (2014).
34. X. Su, L. Su, W. Li, and L. Xiang, "New 3D profilometry based on modulation measurement," Proc. SPIE **3558**, 1–7 (1998).
35. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000).