Area-Based Depth Estimation for Monochromatic Feature-Sparse Orthographic Capture

Yongwei Li*, Gabriele Scrofani†, Mårten Sjöström* and M.Martinez-Corral†
*Department of Information Systems and Technology, Mid Sweden University
Sundsvall, Sweden SE-85170
†Department of Optics, University of Valencia, Burjassot, Spain E-46100

***Abstract*—With the rapid development of light field technology, depth estimation has been highlighted as one of the critical problems in the field, and a number of approaches have been proposed to extract the depth of the scene. However, depth estimation by stereo matching becomes difficult and unreliable when the captured images lack both color and feature information. In this paper, we propose a scheme that extracts robust depth from monochromatic, feature-sparse scenes recorded in orthographic sub-aperture images. Unlike approaches which rely on the rich color and texture information across the sub-aperture views, our approach is based on depth from focus techniques. First, we superimpose shifted sub-aperture images on top of an arbitrarily chosen central image. To focus on different depths, the shift amount is varied based on the micro-lens array properties. Next, an area-based depth estimation approach is applied to find the best match among the focal stack and generate the dense depth map. This process is repeated for each sub-aperture image. Finally, occlusions are handled by merging depth maps generated from different central images followed by a voting process. Results show that the proposed scheme is more suitable than conventional depth estimation approaches in the context of orthographic captures that have insufficient color and feature information, such as microscopic fluorescence imaging.**

***Index Terms*—Depth estimation, integral imaging, orthographic views, depth from focus.**

I. Introduction

Since integral photography (IP) was proposed by Lippmann based on the microlens array (MLA) [1], light field has become an active research topic and many studies on capturing the spatial-angular information of 3D scenes have been reported [2], [3]. Among post-processing techniques of multiple sub-aperture images captured by light field cameras, depth estimation is a fundamental problem due to its wide range of potential applications, such as image-based rendering [4] and refocusing [2]. Although various depth estimation schemes have been proposed during the past decade [5], [6], very little work has explicitly considered the challenging task of estimating depth from monochromatic images with scarce scene features, where either color or textural cues is insufficient for conventional stereo matching methods.

In this paper, we address the depth estimation problem for monochromatic feature-sparse images. Our main contributions are: 1) A focal stack generation scheme for orthographic captures by shifting and superimposing normalized elemental images (EIs). 2) An area-based depth estimation method by measuring normalized cross-correlation (NCC) across the original central image and its corresponding focal stack. 3) A depth candidate voting scheme designed for occlusion handling in sparse-feature space.

The paper is organized as follows: we first briefly revisit related depth from focus (DFF) and Fourier integral microscope (FIMic) techniques in Section II. Section III provides more detailed description of the proposed method, followed by experimental results and analysis presented in Section IV. Finally, this paper is concluded in Section V.

II. Related Work

A. Depth from Focus

DFF refers to methods that estimate the depth of the 3D scene from a focal stack [7]–[9]. A focal stack is defined as a set of images that are captured with varying focus settings, and it is widely used in various applications, such as light field microscopy [10], controlling depth of field (DoF) [11], and depth recovery [12]. In order to determine the depth of an object point by using the focal stack, a focus measure that describes the sharpness of an object is calculated in small regions for each captured image. The focus measure attains the maximum when the object is in focus, and the corresponding focus setting is assigned as the depth of the object. In this paper, we explore the focal stack generation from captured orthographic views.

B. Fourier Integral Microscope

The design of the light field microscope was first initiated by Jang and Javidi [13], who applied the concept of IP to the field of microscopy. Shortly after this, Levoy et al. [10] extended light field microscopy with plenoptic photography. However, the classical plenoptic camera design presents several drawbacks on light field microscopy, including diffraction, vignetting and reduced spatial resolution.

Recently, Scrofani et al. [14] proposed a FIMic scheme based on a telecentric architecture and Fourier-plane recording by placing an MLA at the aperture stop of a microscope objective. The FIMic captures a set of orthographic perspectives of a 3D specimen. Compared with conventional plenoptic microscopes, the FIMic demonstrates an extended DoF, an improved spatial resolution, and reduced influence

ISBN 978-90-827970-1-5 © EURASIP 2018

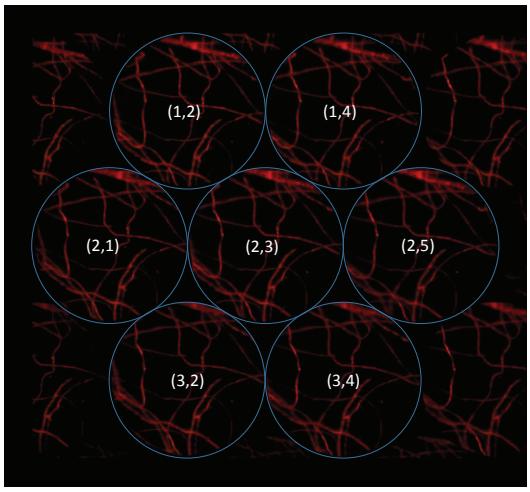


Fig. 1: Elemental images of fluorescent cotton fiber captured by the FIMic. Seven fully captured images are highlighted with circles and labeled with their corresponding row and column indices (r, c).

from optical aberrations such as point spread function (PSF) and vignetting.

III. Proposed Depth Estimation Method

In contrast to studies that are dedicated to depth estimation based on abundant color and feature information, the depth estimation method we propose is designed for monochromatic and feature-sparse images, such as fluorescence microscopy as shown in Fig. 1. In such images, conventional feature-based correspondence search methods become unreliable in detecting robust and accurate correspondences due to the shortage of color and texture information. Therefore, we propose a depth estimation scheme based on DFF and ray-tracing.

A. Focal Stack Generation

For integral images recorded by an orthographic capturing setup such as the FIMic, each micro-lens captures an orthographic sub-aperture image, or the EI, of the scene from a slightly different perspective. A disparity vector can be estimated for every pair of EIs if any spatial point of the scene is not occluded in either of the EIs. Therefore, $N-1$ disparity vectors can be estimated for N EIs with respect to a chosen central view. The magnitude and direction of such disparity vectors indicate the depth of the spatial point and relative position of two EIs respectively. Thus, we can synthesize virtual images that focus on any depth plane within the DoF for an arbitrary EI by superimposing the other EIs with a shift, where the shift amounts and directions are respectively determined by the magnitudes and directions of disparity vectors.

For the hexagonal structure shown in Fig. 1, the left-most column of EIs is assigned with column index $c = 1$. The next left-most column of EIs is assigned with $c = 2$ and so forth. Assignment of EI row index r is done in the same manner.

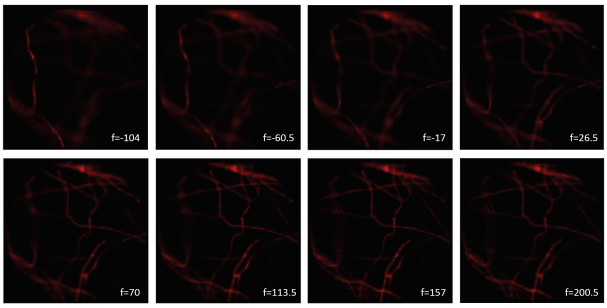


Fig. 2: Examples of the refocused images from the focal stack generated by shifting EIs towards central image EI=$(2, 3)$, focusing at indicated depths. Unit of length: μm.

Therefore, each EI is labeled with (r, c) to mark its specific position in the integral image. Shift amount for each EI with respect to the chosen central image can be calculated by the following equation:

$$(x\prime, y\prime) = (x + D_r \times \cos\frac{\pi}{6} \times s, y + D_c \times \sin\frac{\pi}{6} \times s) \quad (1)$$

where (x, y) is the original coordinate for a pixel in an EI before shift, $(x\prime, y\prime)$ is the superimposed position of (x, y) on the synthesized central image, focal depth is defined by the shifted amount s, and finally D_r and D_c are the differences between the EI and the chosen central image in r and c respectively. However, bilinear interpolation is required in order to have correct pixel positions. Finally, images of the focal stack with respect to the central image can be synthesized by averaging intensities from all the superimposed EIs:

$$I(x\prime, y\prime) = \frac{1}{N}\sum_{m=1}^{N} I_m(x\prime, y\prime) \quad (2)$$

where $I(x\prime, y\prime)$ is the intensity value of the refocused image at position $(x\prime, y\prime)$, $I_m(x\prime, y\prime)$ is the intensity contributed by image m after shift, and N is the number of EIs.

A focal stack composed of the synthesized images for the central view can be generated with the depth:

$$d_f = s \cdot \Delta d_f - \Delta F \quad (3)$$

where d_f is focal depth, ΔF is the focus offset of the capturing system and Δd_f is the difference between each focal depth as a consequence of the orthographic sub-aperture projections. Eight of the refocused images in the generated focal stack for the central image EI=$(2, 3)$ are shown in Fig. 2.

B. Area-Based Depth Estimation

Indeed, the accurate depth can be derived once the pixel-wise correspondence problem between an image pair is solved [3]. However, monochromatic samples, such as fluorescence and gray-scale images, do not provide sufficient and distinct information for unique pixel correspondence matching. Hence, in order to find robust and accurate depths for all pixels, the area-based approach is used, which considers an $n \times n$ window around each pixel to handle the ambiguities [15] caused by

ISBN 978-90-827970-1-5

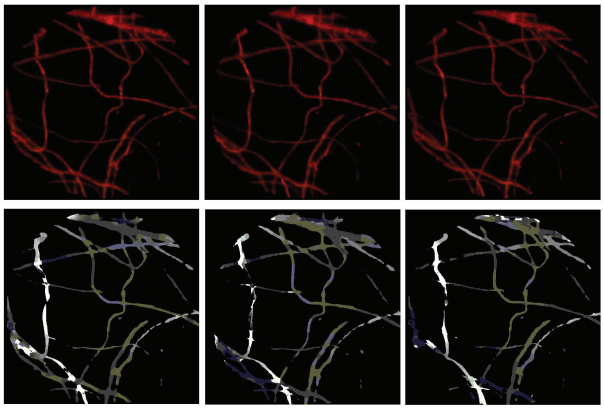
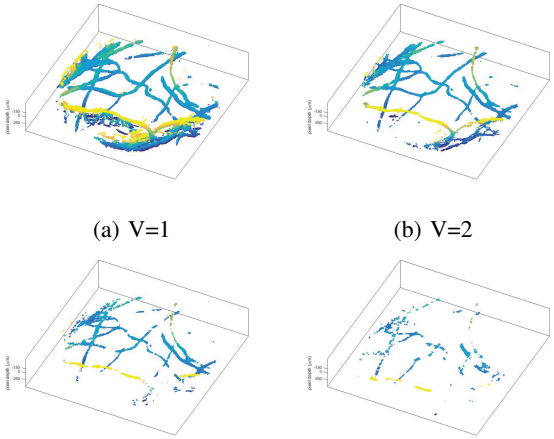


Fig. 3: Examples of depth maps (below) generated for the different EIs (above).



(a) V=1 (b) V=2

(c) V=3 (d) V=4

Fig. 4: 3D point clouds generated by merging all depth maps with respect to different voting threshold V.

irradiance noise and photometric differences among EIs of the capturing system.

The depth of a pixel at (x, y) in the original central image is found by first composing an $n \times n$ square window of neighboring pixels. Then the intensity pattern within the window is compared with identically positioned windows in the synthesized images in the focal stack. NCC is employed as the matching measure between windows, because it is less sensitive to noise than cross-correlation, and more robust to photometric differences among micro-lenses than distance-based measures [16], such as the sum of absolute differences (SAD) and the sum of squared differences (SSD). NCC for the $n \times n$ window centered at (x, y) between the original central image I_{c} and synthesized refocused image I_{r} is calculated as follows:

$$\rho_{I_{\mathrm{c}}I_{\mathrm{r}}}(x,y) = \frac{\sigma^2_{I_{\mathrm{c}}I_{\mathrm{r}}}(x,y)}{\sigma_{I_{\mathrm{c}}}(x,y) \cdot \sigma_{I_{\mathrm{r}}}(x,y)} \tag{4}$$

with

$$\mu_{I_\cdot}(x,y) = \frac{1}{W} \sum_{(i,j)} I_\cdot(x+j, y+i) \tag{5}$$

$$\sigma^2_{I_\cdot}(x,y) = \frac{1}{W} \sum_{(i,j)} \{I_\cdot(x+j, y+i) - \mu_{I_\cdot}\}^2 \tag{6}$$

$$\begin{aligned}\sigma^2_{I_{\mathrm{c}}I_{\mathrm{r}}}(x,y) = \frac{1}{W} \sum_{(i,j)} &\{I_{\mathrm{c}}(x+j, y+i) - \mu_{I_{\mathrm{c}}}(x,y)\} \\ &\cdot \{I_{\mathrm{r}}(x+j, y+i) - \mu_{I_{\mathrm{r}}}(x,y)\}\end{aligned} \tag{7}$$

where $i, j \in [-n, n]$, $W = (2n+1)^2$ refers to the window size, μ is the mean value of the $n \times n$ window and '$I_\cdot$' denotes either the original central image I_{c} or the synthesized refocused image I_{r}.

Thus, the maximum of NCC value is attained when the window is in focus, and the corresponding depth is assigned to the center pixel (x, y) of the window. In case multiple maximums are attained for a pixel, the best depth can be identified by verifying the maximum in different scales. By considering all the pixels (x, y) in the central image I_{c}, a dense depth map is generated. This process is repeated to generate a corresponding depth map for each EI, as shown in Fig. 3.

C. Occlusion Handling

An important benefit of integral photography is that multiple perspectives are available in a single shot, enabling us to handle the occlusion problem which cannot be tackled by a single-perspective capture. For spatial points that are not occluded in any of the EIs, corresponding pixels from different EIs converge to a single spatial point when focused at the correct depth, meaning that the intensities of these pixels are identical if the scene is Lambertian. If we then consider neighboring pixels in a small window, the intensities in such a window share the same distribution in all the EIs. However, this no longer holds when an occlusion occurs and corresponding pixels of EIs record the light emitted by different spatial points. Thus, enforcing stereo matching across the EIs on the occluded area often leads to erroneous depth, causing smooth depth variations between the occluder and the occluded objects [17].

One way of handling the occlusion problem is by inpainting the missing information [18]. However, such reasoning about the consistency of the scene may result in noticeable visual artifacts and fail to produce realistic imagery. Therefore, we handle the occlusion problem by collecting information from all the perspectives and merging them together, not only to achieve a better robustness, but also to avoid artifacts introduced during the inpainting process.

In order to handle the occlusions in the scene, we first use each EI as the central image and generate the corresponding depth map for it using the aforementioned approach III-B. After that we reproduce the scene by back-projecting pixels in all the generated depth maps with respect to the coordinate system of the point cloud via:

$$d(x\prime, y\prime) = d(x, y) \tag{8}$$

ISBN 978-90-827970-1-5 © EURASIP 2018

where $d(x\prime, y\prime)$ and $d(x, y)$ are the depth of generated point cloud P at $(x\prime, y\prime)$ and depth from the depth map at (x, y) respectively. The transition between depth map coordinates (x, y) and point cloud coordinates $(x\prime, y\prime)$ is given in Eq. 1.

When there is occlusion, the scene can be seen by a subset of all the EIs and only unoccluded pixels converge to a single spatial point, showing the same intensity. However, occluded pixels from the other EIs exhibit no photometric consistency as they do not converge to any spatial point. Based on the above reasoning, the accuracy and robustness of the point cloud can be further improved by introducing a voting scheme: if a spatial point is seen by at least V EIs, then it is registered as an object point in the point cloud P, as shown in Fig. 4. Note that, when $V = 1$, all the back-projected points from the depth maps are kept in the point cloud and the voting process is not involved. Thus, to avoid erroneous depths around the boundaries of occlusions, one need to assure $V > 1$ to adopt voting process.

IV. Experimental Results

Our proposed depth estimation scheme is tested on FIMic capture, where the MLA is arranged in a hexagonal structure. Only 7 fully captured EIs which are highlighted in Fig. 1 are used for depth estimation. The focal plane of the FIMic is indicated as $d_f = 0$. The pitch of each EI is 683 pixels, the focus offset $\Delta F = 70\mu m$, and one-pixel disparity corresponds to $\Delta d_f = 14.5\mu m$ depth change. An integral image depicting cotton fiber is used to evaluate the proposed method in various aspects. The cotton fiber specimen is stained with red fluorescent ink, therefore only intensity information stored in the red channel of the captured image is used for depth estimation. The specimen is illuminated by a laser of wavelength $\lambda = 532nm$.

A focal stack can be generated with sub-pixel shifts by applying the scheme proposed in section III-A. However, we observe that focus settings beyond the DoF of the capturing setup cannot be used to generate accurate and robust depth. For the sake of simplicity without compromising the performance of the proposed method, a focal stack composed of 25 different focus settings within the DoF of FIMic setup is generated by shifting one pixel each step towards an arbitrary central EI. Eight of the 25 refocused images from the focal stack with respect to central EI=$(2, 3)$ are shown in Fig. 2.

Point clouds (Fig. 4) are generated by merging all the depth maps and registering points which are seen by at least V depth maps. It is shown in Fig. 4a that there are many replicas for the same object when no voting process is involved. Such phenomenon is caused by registering inaccurate and unstable spatial points that can be seen by only one view. In order to achieve an accurate and robust point cloud, it is required that $V > 1$ in the voting process. However, the number of points in the point cloud decreases as V increases, which means that setting up an excessively high V will discard the essential information for scene reconstruction, as shown in Fig. 4d. Therefore, V should be chosen according to the specific capturing setup and application.

To illustrate the accuracy and robustness of the proposed method in estimating depths, comparisons have been conducted by applying the proposed method, shape from focus (SFF) [9] and improved SFF based on reliability measure (R-SFF) [19] with regard to the same central image EI=$(2, 3)$ and its focal stack. Although the ground truth depth of cotton fiber cannot be acquired at the present stage, a visually noticeable comparison can be seen from Fig. 5 that both SFF and R-SFF fail in recovering the depth information of the cotton fibers in two ways: 1) The generated point cloud shows that the estimated depths for all object points are extended from the surface to the bottom, which means that the correct depths are not estimated and the scene cannot be reconstructed afterwards. 2) The fiber structure on depth maps estimated by SFF and R-SFF is wider than the fiber structure in the captured EI. This implies that some spatial points are artificially put into the depth map, resulting in a false structure of the scene after reconstruction.

In contrast, the proposed method shows several significant advantages compared with the other SFF-based methods. First of all, fibers can be distinguished easily by observation, and the depths for each fiber section is constrained in a reasonable depth interval that matches the approximate thickness of the fibers, as shown in Fig. 5a. Secondly, the fiber structure in the estimated depth map highly matches with the scene recorded in EIs as shown in Fig. 3, meaning that the reconstructed 3D scene is of a higher fidelity compared with the other methods.

However, some failures occur for parts where the captured scene is not in focus at any of the focal planes of the focal stack, especially in the overlapping area of multiple blurry threads which are beyond the DoF of the capturing setup. In this case, all the refocused images from the focal stack cannot distinguish between the swarm of threads and reconstruct the thread structure. Such a failure results in the incorrect and unstable depths, which are visible in the top right corner of all the depth maps in Fig. 3.

V. Conclusions

In this paper, we have proposed a method that enables depth estimation in the event of insufficient color and feature information for orthographic views of the scene. This is achieved by processing orthographic EIs in three steps: initial focal stack generation, area-based depth estimation, and occlusion handling by a voting scheme. The experimental results have shown that the proposed method outperforms the other SFF-based depth estimation methods and estimates depths with high accuracy and robustness. In more compromised scenes (e.g. parts of scene are out of DoF), the proposed method gives inconsistent results and further improvements are necessary.

Acknowledgment

The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

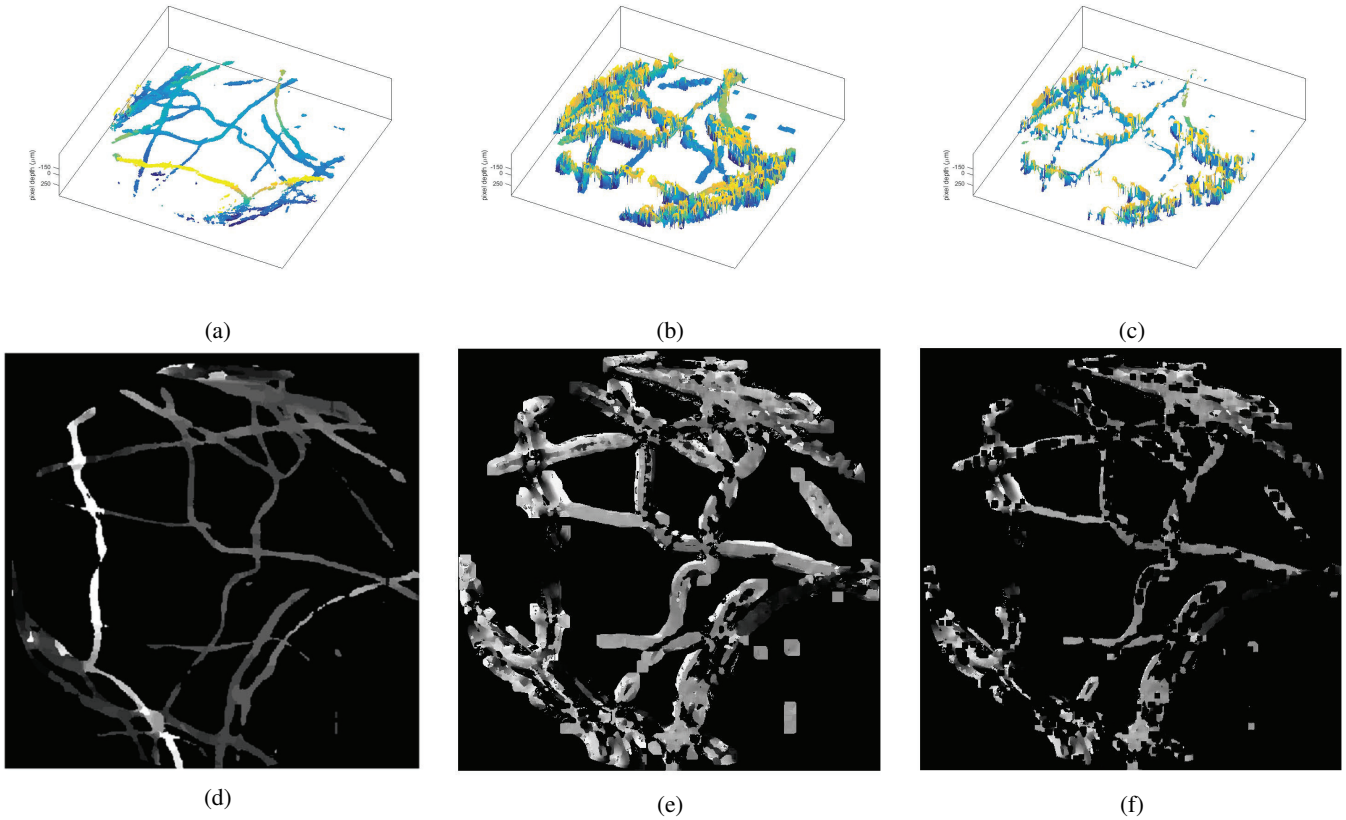


Fig. 5: Depth estimation performance evaluation of the proposed method, SFF, and reliability-based SFF. The top row shows the point clouds generated by (a) the proposed method with $V = 2$, (b) SFF and (c) R-SFF with $\alpha = 20dB$. The bottom row shows their corresponding estimated depth maps.

REFERENCES

[1] G. Lippmann, "Epreuves reversibles donnant la sensation du relief," *J. Phys. Theor. Appl.*, vol. 7, no. 1, pp. 821–825, 1908.

[2] 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*, vol. 2, no. 11, pp. 1–11, 2005.

[3] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*, vol. 8291. International Society for Optics and Photonics, 2012, p. 829108.

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.

[5] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2170–2181, 2016.

[6] C.-T. Huang, "Robust pseudo random fields for light-field stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 11–19.

[7] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 2, pp. 97–108, 1993.

[8] M. Subbarao and J.-K. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 864–870, 1998.

[9] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[10] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," in *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 924–934.

[11] D. E. Jacobs, J. Baek, and M. Levoy, "Focal stack compositing for depth of field control," *Stanford Computer Graphics Laboratory Technical Report*, vol. 1, no. 1, p. 2012, 2012.

[12] H. Lin, C. Chen, S. Bing 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.

[13] J.-S. Jang and B. Javidi, "Three-dimensional integral imaging of micro-objects," *Optics letters*, vol. 29, no. 11, pp. 1230–1232, 2004.

[14] G. Scrofani, J. Sola-Pikabea, A. Llavador, E. Sanchez-Ortiga, J. Barreiro, G. Saavedra, J. Garcia-Sucerquia, and M. Martínez-Corral, "Fimic: design for ultimate 3d-integral microscopy of in-vivo biological samples," *Biomedical Optics Express*, vol. 9, no. 1, pp. 335–346, 2018.

[15] H.-H. Nagel and W. Enkelmann, "An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 565–593, 1986.

[16] O. Schreer, P. Kauff, and T. Sikora, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*. John Wiley & Sons, 2005.

[17] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3487–3495.

[18] Z. Tauber, Z.-N. Li, and M. S. Drew, "Review and preview: Disocclusion by inpainting for image-based rendering," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 4, pp. 527–540, 2007.

[19] S. Pertuz, D. Puig, and M. A. Garcia, "Reliability measure for shape-from-focus," *Image and Vision Computing*, vol. 31, no. 10, pp. 725–734, 2013.

ISBN 978-90-827970-1-5 © EURASIP 2018