3D IMAGING SYSTEM USING MULTI-FOCUS PLENOPTIC CAMERA AND TENSOR DISPLAY

Mehrdad Teratani (Panahpourtehrani), Shu Fujita, Wenzhe Ouyang, Keita Takahashi, Toshiaki Fujii

Nagoya University, Japan

ABSTRACT

In this paper, we introduce a 3D imaging system using a multi-focus plenoptic camera and a tensor display. A multi-focus plenoptic camera is a powerful device that can capture the light field (LF), which is interpreted as a set of dense multi-view images. This camera has the potential to acquire a LF having high spatial/view resolutions and deep depth-of-field. A tensor display, which consists of a few light attenuating layers stacked in front of a backlight, can visualize a light field with high resolution. To display the captured data by a multi-focus plenoptic camera, we need to extract multi-view images. For this, we propose a sophisticated rendering process for the complicated optical system of multi-focus plenoptic cameras. The multi-view images are converted to a multi-layer representation for the multi-layer display, i.e. tensor display. To demonstrate the effectiveness of this system, we conducted both computer simulations and experiments on our prototype display.

Index Terms— Multi-focus plenoptic camera, lenslet, conversion, multi-view video, tensor display.

1. INTRODUCTION

Light field (LF) has both spatial and angular information, and provides a rich representation of real-world scenes, i.e. dense set of multi-view images. A LF display should be simultaneously capable of displaying a set of dense multi-view images. Here, we introduce a processing pipeline from capturing data with a multi-focus plenoptic camera to a display, as shown in Fig. 1. The processing pipeline converts data from the Raytrix camera into multiview video and then into a layered pattern for a tensor display.

A LF [1] is a 4D signal representation that describes all light rays traveling in 3D space. A LF is represented as a set of dense multi-view images, where 2D spatial and 2D view coordinates are used to describe the 4D space. LFs have been used for many computer vision and graphics applications such as depth estimation [2], digital refocusing [3], and 3D displays [4]. For capturing a LF, one of the commercially available powerful devices is a focused plenoptic camera (FPC) [5] [6]. The FPC consists of a main lens and a microlens array, and each microlens performs as a micro camera. This structure enables us to capture a LF with high spatial and view resolutions.

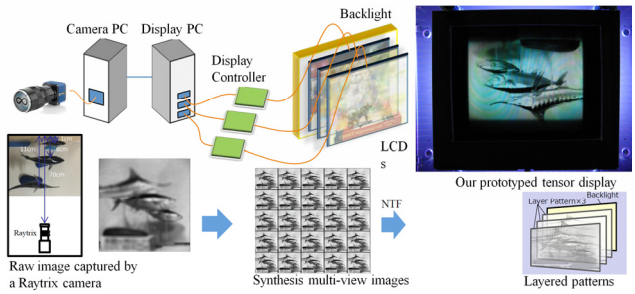

Fig. 1. 3D imaging system from acquisition to display.

Recently, multi-focus plenoptic cameras (MFPC) [7] [8] has been developed. It includes an array of interleaved microlenses with different focal lengths and can produce a deeper depth-of-field. The Raytrix cameras that utilize the MFPC's effectiveness have been commercialized. However, despite their effectiveness, MFPCs are not often utilized. We think that this is because they are difficult to handle due to their complicated optical system. Hence, we need a sophisticated rendering process to generate a LF.

A LF display can provide a better stereoscopic experience compared to traditional binocular displays, which display only two views, one for the left eye and one for the right eye. To develop such a display, several methods have been proposed, such as parallax barrier [9], lenticular lenses [10], and rear projection [11][12] displays. Among them, a tensor display [13] has been introduced as an emerging and promising technology. This display is composed of a few liquid crystal displays (LCDs) stacked in front of a backlight. By controlling the transmittance of each layer, this configuration can display a view-dependent image that corresponds to the target 3D object.

To display real-world scenes using a multi-layer display, i.e. a tensor display, a processing pipeline from acquisition to display has been demonstrated in [14], in which a Lytro Illum camera [15] and a multi-view camera were used to capture LF data. In this paper, we present a processing pipeline from the Raytrix camera [16, 17] to a tensor display. The Raytrix camera is also a plenoptic camera, but it has some structural difference compared to the Lytro Illum camera. Because of this difference, the Raytrix camera can capture LF with higher spatial resolution and deeper depth-of-field. The Raytrix camera can capture LF video.

978-1-5386-7590-8/18/$31.00 ©2018 IEEE

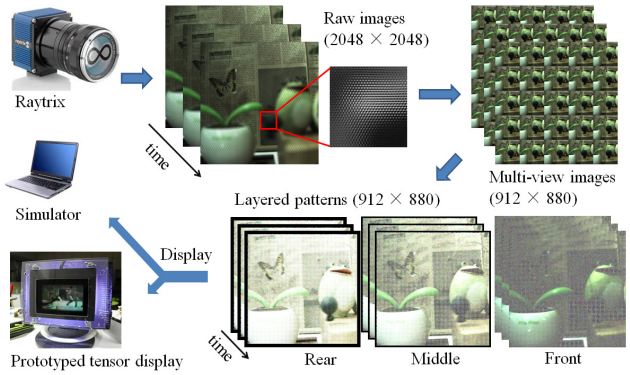

Fig. 2. Processing pipeline from capture to display.

For displaying, we used our developed tensor display [14]. We completed the pipeline that can convert input from the Raytrix camera into layered pattern for a tensor display. To demonstrate the effectiveness of our pipeline, we conducted computer simulations and experiments on our prototype tensor display.

2. 3D IMAGING SYSTEM

This section introduces the acquisition and display devices of our 3D imaging system, as shown in Fig. 2.

2.1. Acquisition using multi-focus plenoptic camera

As for the input device, we used the Raytrix R5-CGigE-F2.4 camera to capture LF video content at up to 180 fps. Specifically, this Raytrix camera has 2048x2048 pixels; the number of microlenses is 8787 and each covers a hexagonal region that is 23 pixels in diameter. The Raytrix is a type of focused plenoptic camera with microlenses having different foci. An Example of a captured image [18] by using the Raytrix R5-CGigE-F2.4 camera is shown in Fig. 3.

2.2. Display using tensor display

Recently, we have also developed a layered 3D display, i.e. tensor display [13], where the user can see a 3D scene with the naked eye. This display is a stacked layered display. It can emit a LF through a few light attenuating layers, e.g. LCD panels that are stacked in front of a backlight. This method can achieve remarkable performance in terms of efficiency because each pixel of a layer is shared by many views simultaneously.

Our developed prototype tensor display (input and output) and its hardware components are shown in Fig. 4. For the output device, we used our tensor display [14]. Our prototype consists of three semi-transparent LCD panels and a hand crafted backlight. Visualizing real-world 3D scenes with this display is a challenge in terms of data acquisition because a dense LF, i.e. set of multi-view images (typically dozens of images) with very small viewpoint intervals, is required as the input. Each layer of LCD has a resolution of

Fig. 3. Raytrix R5 and the captured lenslet image.

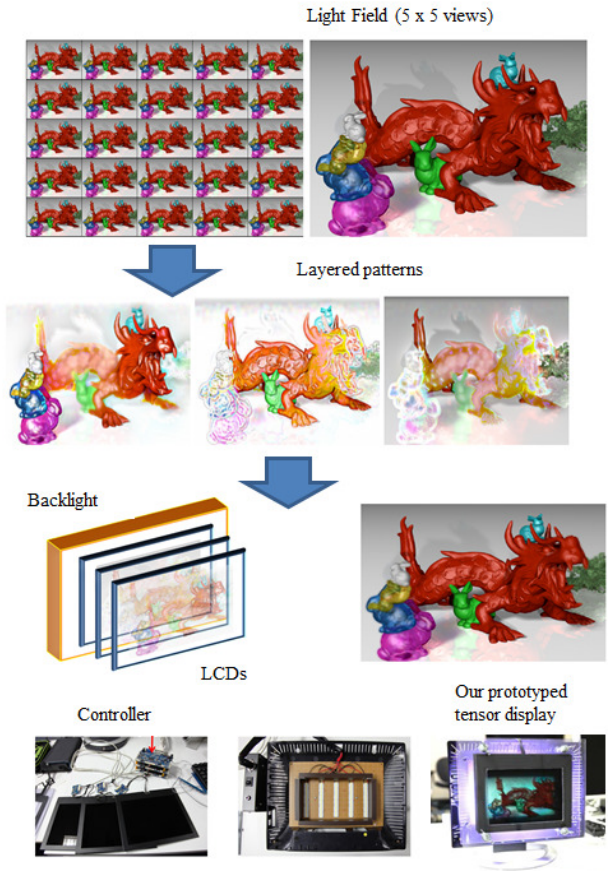

Fig. 4. Processing pipeline from multi-view images to layered patterns for tensor display.

1024x768 pixels on a 9.7 inch panel and a refresh rate of 60 or 75 Hz.

3. PROCESSING PIPELINE

The processing pipeline from capture to display consists of two important parts that are the conversion from the lenslet video captured by the Raytrix camera to multi-view video, and the conversion from the multi-view video to the multi-layer pattern videos, as shown in Fig. 2 and Fig. 4. The main focus of this section is on the rendering process for extracting multi-view video from the lenslet video captured by a multi-focus plenoptic camera.

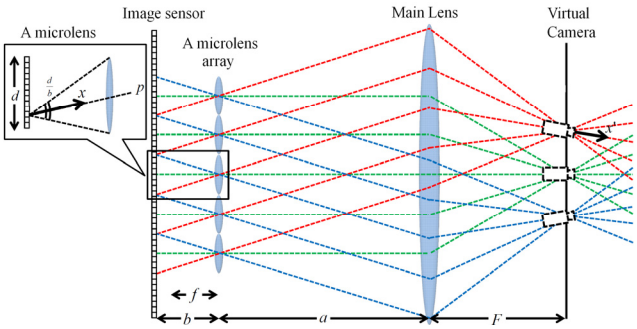

Fig. 5. Optical system of plenoptic camera, e.g. Lytro ($b=f$) and focused plenoptic camera, e.g. Raytrix ($b\neq f$).

3.1. Extracting multi-view images from multi-focus plenoptic camera lenslet image

The optical system of MFPC is more complicated compared to that of an ordinary plenoptic camera, e.g. the Lytro. Despite the effectiveness, MFPCs are not often utilized. We think, the main reason is that MFPCs are difficult to handle due to their complicated optical system. Fig. 5 shows the optical system in a FPC compared to a plenoptic camera.

The optical system of the plenoptic camera [3] has a microlens array inside the camera to acquire the LF. The imaging plane of the main lens is at the position of the microlens array. On the other hand, in a FPC such as the Raytrix camera, the imaging plane of the main lens is in front of the microlens array. This makes it possible to acquire high-density directional rays, and as a result, it is possible to acquire a LF with a higher spatial resolution [14] than the number of microlenses used. In Fig. 5, a is the distance from the main lens to the microlens, and b is the distance from the microlens to the sensor surface. F and f are the focal lengths of the main lens and microlens, respectively. For the plenoptic camera, $b=f$, and for the focused plenoptic camera, $b\neq f$.

The Raytrix camera, i.e. MFPC, combines three types of microlenses with different foci. Due to complicated optical system in the Raytrix, we need a sophisticated rendering process to generate a high-quality LF. The straightforward rendering method [5] is to extract not a pixel but a patch with a fixed size from each image of a microlens, as shown in the Fig. 6. However, using a fixed patch size causes severe artifacts as shown in Fig. 7. We need to adaptively adjust the patch size to solve the problem, but few studies have addressed this problem [19] [20]. In addition, to the best of our knowledge, the only available rendering method is the one provided by Wanner et al. [20], but this method is not robust to some of the camera configurations.

Under the circumstances, the methodology to render a LF is not sufficiently sophisticated, and few people can easily handle the MFPC. We therefore propose a rendering method [21] [22] with a better performance, and release the source code, for academic purposes. We explain our proposed conversion method in the following.

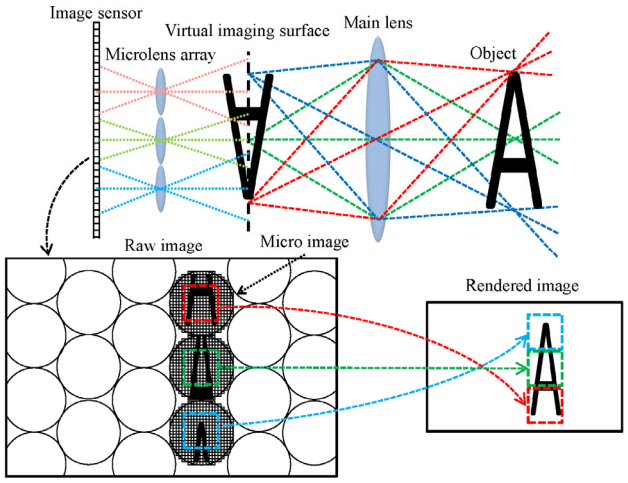

Fig. 6. Optical system of focused plenoptic camera (top) and diagram of rendering image from raw image captured by focused plenoptic camera (bottom).

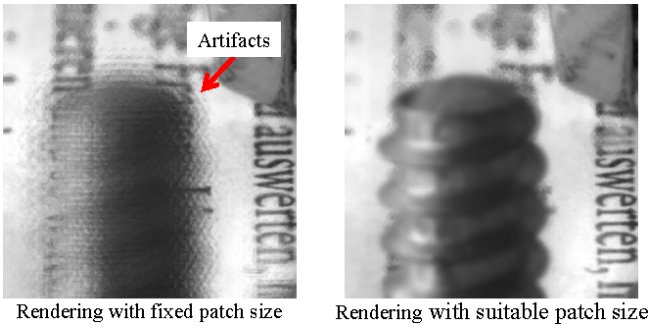

Fig. 7. Rendering artifacts.

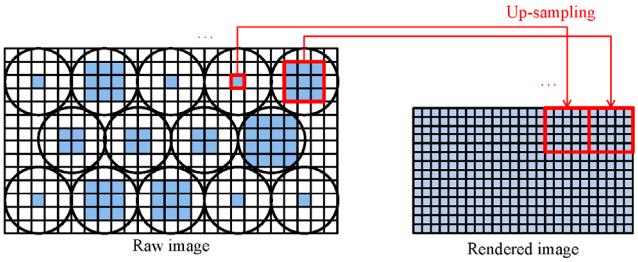

Fig. 8. Rendering image at particular point using suitable patch sizes.

Our method is inspired by the Laplacian-based method proposed by Wanner et al. [20]. We utilize all the viewpoints in the rendered multi-view video. We first consider the patch size estimation. Wanner et al. estimate the patch size by using only the information of the center viewpoint. In contrast, we compute the Laplacian of all viewpoint images. This is because a suitable patch size for a micro image is common for rendering of almost all the viewpoints. Hence, by computing the Laplacian of all viewpoint images, we are able to robustly estimate the patch

size. Fig. 8 illustrates how different patch sizes are selected and up-sampled for rendering a viewpoint.

We introduced a novel integrating method for three sets of multi-view images with different foci produced from the multi-focus plenoptic camera. The averaging method can efficiently reduce the rendering artifacts while simultaneously causing blurring of textures. To avoid this problem, we introduce a weighted averaging method.

There are three sets of multi-view images with different foci; each multi-view image is focused on a particular depth according to the focal length. Moreover, information of suitable patch size can also be interpreted as depth information. Therefore, we weight each multi-view image depending on the lens type.

The multi-view images have clear texture due to weighting, while the weighting causes artifacts when an unreliable patch size, e.g., patch sizes estimated from low-frequency regions, is used. We solve this problem by applying a blending process of two multi-view image sets integrated with averaging and weighted averaging. The patch size is unreliable when it is estimated from low textured or occluded regions. The details of our proposed method are explained in [21] [22].

3.2. Converting multi-view images to layered pattern representation for tensor display

For generating layered patterns using the method that has been explained in [14], we utilized the method proposed in [23], which introduced a generalized framework for LF factorization that can handle both the orthographic and perspective models. Here, our target is to display real-world 3D content, which is captured by a perspective camera; therefore we adopted the perspective model only.

After extracting multi-view video from lenslet video, we calculate layered patterns. We will have a set of layered patterns for every frame. A few light attenuating layers are stacked in front of a backlight. The transmittance of each layer pixel can be controlled individually in accordance with the content to be displayed. Depending on the viewing direction, these layers overlap with a different shift, so the displayed images are direction dependent.

More precisely, multi-view video is expected to be observed from different viewing directions. Having multi-view video as the input, the layered patterns are optimized so that it can reproduce the multi-view video as accurately as possible in the form of layered patterns. This optimization [14] [23] is conducted through a non-negative tensor factorization (NTF), where the transmittance values of layers is alternately updated. Those layered patterns are input of our simulated and our prototype tensor displays.

4. EXPERIMENTS

To demonstrate the effectiveness of our pipeline, we conducted the following two sets of experiments.

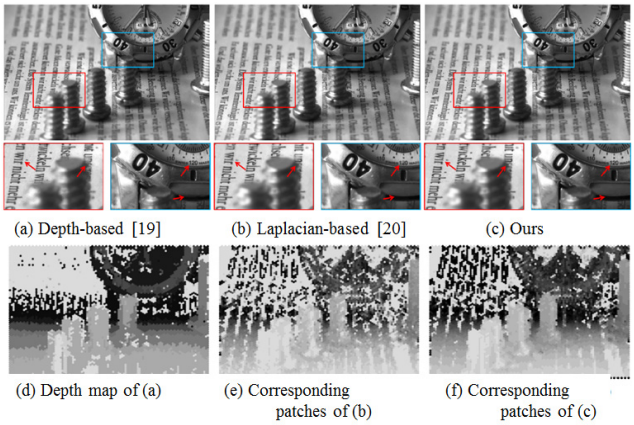

Fig. 9: Rendering using different patch size estimations.

4.1. Evaluation of the conversion tool from lenslet image captured by the Raytrix to multi-view video

We verify our contribution by comparing it with Wanner et al. [20]. We use raw datasets captured by a multi-focus plenoptic camera for this experiment. For the raw lenslet video data, we use a dataset that has been provided in [20], two datasets that were provided by Palamieri et al. [25], and we used a raw image that we have captured using the Raytrix R5 camera only for verification of our proposed conversion. We implemented the rendering methods using C/C++ on Linux platforms. The implementation is released in our website for academic purposes. In our experiment, due to the complexity of the optical system in multi-focus plenoptic cameras, it is difficult to generate ground truth using computer simulation. Therefore, we can only visually evaluate results of the different methods for converting lenslet video to multi-view video.

Fig. 9 shows the rendering results and corresponding patch maps estimated by three methods: the conventional depth-based method [7], the conventional Laplacian-based method [20], and our method. For fair comparison, we have implemented all methods, i.e. depth-based [19], Laplacian-based [20] and our proposed method. For the same reason, we have also integrated the multi-view video extracted from the three microlens types, using only averaging. We can confirm that the depth-based result has a clear texture on flat regions from Fig. 9a, with significant artifacts around occlusions. The conventional Laplacian-based method can suppress these artifacts, but some regions are blurred as shown in Fig. 9b. In contrast, our improved method can suppress both, artifacts and blurs. Moreover, the Laplacian-based patch maps are noisy, and the depth-based patch map seems to be good. However, as shown in Fig. 9, the accuracy of depth information does not always contribute to the image quality.

Fig. 10 shows the rendering results using different integration methods for three microlens types. Here, we fixed the patch estimation for our proposed method. The

weighted-averaging integration achieved a clearer texture for high-frequency regions than the averaging integration, which produced blurs. Meanwhile, the blurriness effectively works to suppress artifacts in regions that have an unreliable patch size such as extremely blurred regions in the background of Fig. 10. Blending the two methods enables us to obtain multi-view images that have the clearest texture and the least artifacts as shown in Fig. 10c.

Finally, we compare the results of our implementation with [20]. Fig. 11 shows the comparison results. The top row is a dataset provided in [14], the second row is one of the dataset provided by Palamieri et al. [16], and the bottom row is a dataset we captured only for this evaluation. Note that color images are not supported in the implementation provided for the method in [20]. Although this implementation works well with their dataset as shown in the top row of Fig. 11, our implementation can produce sharper results than in [20]. With the datasets in [25], the difference in performance between the conversion method by Wanner et al. [20] and ours is even more obvious. Additionally, using our dataset as the input, the implementation in [20] failed to render multi-view images. Therefore, Wanner et al.'s implementation is not robust to variation of configurations in the dataset, whereas ours can robustly produce good multi-view images captured by a multi-focus plenoptic camera.

4.2. Evaluation of 3D imaging system from capture to display

We examine the 3D system performance by subjectively evaluating the displayed quality from a set of captured lenslet videos.We captured several datasets with the Raytrix R5-CGigE-F2.4 camera. These datasets consist of three different capturing configurations, as shown in Table 1. The target scene with configuration 3 is shown in Fig. 12. Note that these three configurations look similar, while the contents are not synchronized. Every dataset has 100 frames with 30 fps. We extracted 5x5 multi-view images from the raw data of every frame and used them to calculate layered patterns using the method in [23].

We used the tensor display simulator that we have previously developed [24]. With this simulator, we can directly change the viewpoint. It can change several parameters that affect the displayed image 3D effect, such as the distance between the observer's viewpoint and the display, the intervals between the layers, and the distance among the virtual viewpoints which input images are assigned to. We modified this simulator to display several frames in a sequence. The simulator can display LF videos at up to 60 fps with an Intel Quad-core processor and 16GB RAM. The simulation results with configurations 1 and 2 at the 10th frame are shown in Fig. 13 and Fig. 14, where the left, middle, and right columns correspond to the (View A), (View B) and (View C) marked in Fig. 12.

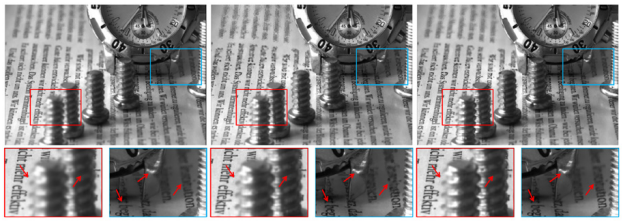

Fig. 10: Rendering using different integration methods of three microlens types.

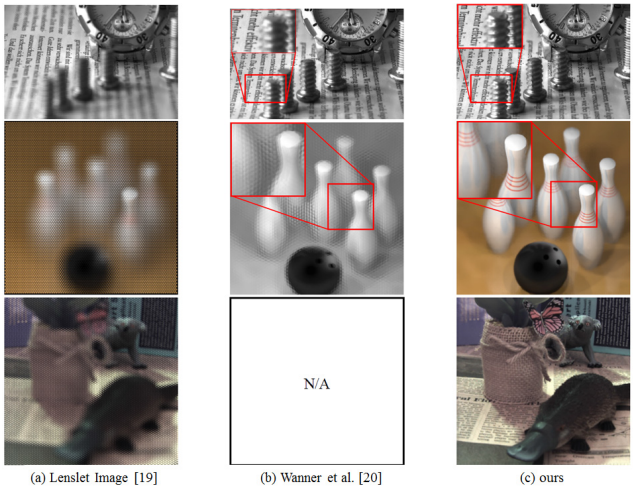

Fig. 11. Comparison of our implementation and that of [14, 17].

Table 1. Capturing configuration used in the experiments.

Object	Three datasets with three configurations		
	Configuration-1	Configuration-2	Configuration-3
Tadpole	150 mm	225 mm	300 mm
Leaves	200 mm	300 mm	400 mm
Butterfly	300 mm	450 mm	600 mm

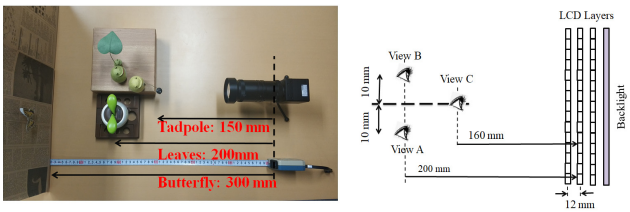

Fig. 12. Capturing condition in configuration-3 (left), and simulation setting of configuration-1 (right), in Table 1.

By comparing the results of (View A) and (View B), we can observe stereoscopic effects, especially in configuration-1, which has the maximum range of disparities among the three configurations. We also conducted experiments on our prototype of tensor display [14]. The experimental results with configuration-1 at the 10th frame are shown in Fig. 15. We perceived sensation 3D (depth) when we observed the display from 50 cm away.

5. CONCLUSIONS

In this paper, we presented a 3D system consisting of the Raytrix camera and a tensor display. In our system, the processing pipeline has two steps that are the conversion from the captured lenslet video by the Raytrix camera to multi-view video, followed by the conversion from the multi-view video to a layered pattern representation for a tensor display. To show the effectiveness of our system, we subjectively evaluated the displayed video using both, our simulated and prototype tensor displays.

We proposed a sophisticated rendering method for generating multi-view images from a multi-focus plenoptic camera. Specifically, we improved the patch size estimation for rendering images for each microlens type and introduced a method to integrate the three microlens types. Our conversion tool outperforms existing methods. As a result, we can successfully generate multi-view images that have sharp edges and less artifacts. The results showed that our implementation can robustly work in various configurations. However, this conversion cannot yet be performed in real-time due to its complexity. Currently, in our processing pipeline, only conversion from a multi-view video to a layered pattern for the tensor display is possible to be performed in real-time.

In our future work, we will focus on improving the conversion tool from lenslet to multi-view images, direct conversion from lenslet to layered pattern video, and adding compression in the chain from capture to display. We release the tool for converting lenslet to multiview video, for academic purposes. We believe that our tool can contribute to increasing the opportunities for applications using multi-focus plenoptic cameras.

6. ACKNOWLEDGEMENT

This work is partially supported by Grants-in-Aid for Scientific Research (C) registered number 16K06349.

7. REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *ACM SIGGRAPH*, pp. 31–42, 1996.

[2] M. W. Tao and S. Hadap and J. Malik and R. Ramamoorthi, "Depth from Combining Defocus and Correspondence Using Light-Field Cameras," In *Proc. of ICCV,* 673—680, 2013.

[3] R. Ng, M. Levoy, M.,Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light-field photography with a handheld plenoptic camera," In *Stanford University Computer Science Tech Report CSTR*, 2005.

[4] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multi-layer displays with directional backlighting," In *ACM Transactions on Graphics* 31(4), 1-11, 2012.

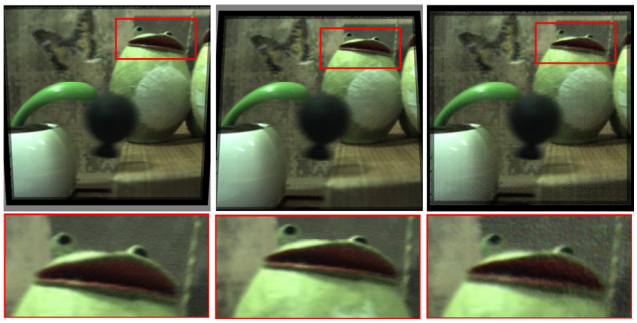

Fig. 13. Displayed results on our "simulated tensor display" for (a) Configuration 1 (left, middle, and right correspond to View A, B, and C in Fig. 12, respectively).

Fig. 14. Displayed results on our "simulated tensor display" for Configuration 2 (left, middle, and right correspond to View A, B, and C in Fig. 12, respectively).

Fig. 15. Displayed results on our "prototype tensor display" for configuration 1 (observed from the left, center, and right direction, respectively).

[5] A. Lumsdaine, and T. G. Georgiev, "The focused plenoptic camera," In *IEEE International Conference on Computational Photography*, 1-8, 2009.

[6] A. Lumsdaine, and T. G. Georgiev, "Focused plenoptic camera and rendering," In *Journal of Electronic Imaging* 19(2), 021106:1-021106:11, 2010.

[7] A. Lumsdaine, and T. G. Georgiev, "The multi-focus plenoptic camera," In *Proceedings of SPIE, The International Society for Optical Engineering*, 8299:7-8299:17, 2012.

[8] C. Perwaβ, C. and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," In *Proceedings of SPIE, Human Vision and Electronic Imaging* XVII, 8291:1-8291:15, 2012.

[9] F. E. Ives, "Parallax stereogram and process of making same," *U.S. Patent US72556A*, 1903.

[10] G. Lippmann, "Épreuves réversibles donnant la sensation du relief," *Journal de Physique Theorique et Appliquee*, vol. 7, no. 1, pp. 821–825, 1908.

[11] W. Matusik and H. Pfister, "3D TV: A Scalable System for Real-Time Acquisition, Transmission and Autostereoscopic Display of Dynamic Scenes, " *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 814–824, 2004.

[12] Y. Takaki and N. Nago, "Multi-projection of lenticular displays to construct a 256-view super multi-view display," *Optics express*, vol. 18, no. 9, pp. 8824–8835, 2010.

[13] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–11, 2012.

[14] Y. Kobayashi, S. Kondo, K. Takahashi, and T. Fuji, "A 3D Display Pipeline: Capture, Factorize, and Display the Light Field of a Real 3D Scene," *ITE Transactions on Media Technology and Applications*, vol. 5, no. 3, pp. 88–95, 2017.

[15] “Lytro, "https://www.lytro.jp/".

[16] “Raytrix, "https://www.raytrix.de/.

[17] M. Panahpour Tehrani, S. Fujita, S. Mikawa, Y. Kobayashi, K. Takahashi, and T. Fujii, "[MPEG-I Visual] Development of a 3D Imaging System Using Llight Field Camera and Tensor display," *ISO/IEC JTC1/SC29/WG11*, M41244, 2017.

[18] M. Panahpour Tehrani Fujita, S. Mikawa, K. Takahashi, and T. Fujii, "[MPEG-I Visual] Introduction to A New Test Sequence Tunnel_Train 2 Captured by Light Field Video Camera," *ISO/IEC JTC1/SC29/WG11*, M41787, 2018.

[19] T. G. Georgiev and A. Lumsdaine, "Reducing plenoptic camera artifacts," *Computer Graphics Forum*, vol. 29, no. 6, pp. 1955–1968, 2010.

[20] S. Wanner, J. Fehr, and B. Jahne, "Generating epi representations of 4D light fields with a single lens focused plenoptic camera," in *International Conference on Advances in Visual Computing*, pp. 90–101, 2011.

[21] M. Teratani (Panahpourtehrani), S. Fujita, S. Mikawa, Y. Kobayashi, K. Takahashi, and T. Fujii, "[MPEG-I Visual] Development of a 3D Imaging System Using Light Field Camera and Tensor display, " *ISO/IEC JTC1/SC29/WG11*, M44731, 2018.

[22] S. Fujita, S. Mikawa, M. Teratani (Panahpourtehrani), K. Takahashi, and T. Fujii, "Extracting Multi-view Images from Multi-focus Plenoptic Camera," To be in *International Workshop on Advanced Image Technology*, IWAIT, 2019.

[23] S. Kondo, Y. Kobayashi, K. Takahashi, and T. Fujii, "Physically-Correct Light-Field Factorization for Perspective Images, " *IEICE Transactions on Information and Systems*, vol. E100.D, no. 9, pp. 2052–2055, 2017.

[24] S. Mikawa, K. Takahashi, and T. Fujii, "Simulating a Stacked-Layer Light-Field Display Using Orthographic/Perspective View Models," *International Workshop on Advanced Image Technology*, 2018.

[25] Palmieri, L., Veld, R. O. H., and Koch, R., "The plenoptic toolbox 2.0 aims to help promote research using focused plenoptic cameras," in *IEEE International Conference on Image Processing*, 2018.