Light Field Acquisition from Focal Stack via a Deep CNN

Yasutaka Inagaki, Keita Takahashi, Toshiaki Fuji

Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan

Keywords: light-field, coded aperture camera, convolutional neural network

ABSTRACT

We succeeded in acquiring a dense light field from a focal stack, i.e, only a few images with different focused depth, by using a deep convolutional neural network (CNN) trained for this purpose. We validated our method through both simulative and real-camera experiments.

1. INTRODUCTION

A light field describes all light rays traveling in a free three-dimensional space. It is often represented as a set of dense multi-view images. The light field representation has been used in various applications, such as 3D displays [1], view synthesis, and depth estimation.

Acquiring a light field is a challenging task due to the fact that it consists of dozens of images. The most straightforward approach is to use a moving camera gantry or multiple cameras, which is costly in terms of the hardware or the time required to capture the entire light field. In another approach, lens-array based cameras can obtain the entire light field from a single acquired image, but the spatial resolution of each image is in a trade-off relationship with the number of viewpoints.

In our method, we focus on a different approach based on compressive sensing, where the entire light field is computationally reconstructed from less observation data without sacrificing the spatial resolution of each image. Compressive sensing can be implemented with several methods. One popular method is to insert a semi-transparent coded pattern (coded aperture or coded mask) into the optical path of a camera [2]. Another promising method is to take several images with different focused depths (a focal stack) [3].

In our previous work [4], we achieved a state-of-the-art performance on light-field acquisition using a coded-aperture camera; we successfully reconstructed 5×5 and 8×8 light-field images only from two acquired images by using the powerful framework of convolutional neural networks (CNNs). In the present paper, we extend the network developed in the previous work [4] to the light-field reconstruction from a focal stack. We demonstrate that only two images that are focused at different depths are sufficient for light field reconstruction. We also compare the coded aperture (CA) and focal stack (FS) methods through both simulative experiments and experiments using real cameras shown in Fig. 1.

Fig. 1: Cameras used for FS (left) and CA (right) methods.

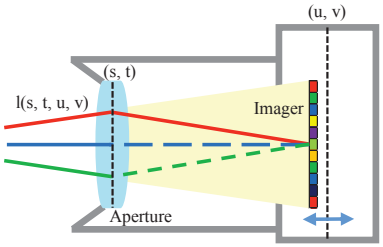


Fig. 2: Light field definition inside a camera.

2. PROPOSED METHOD

2.1 Light field, CA, FS, and compressive acquisition

A light field is defined over a 4D space (s,t,u,v), and the intensity of a light ray is described as $l(s,t,u,v)$. In Fig. 2, (s,t) and (u,v) denote the intersections of a light ray with the aperture and imager planes, respectively. More accurately, the position of the imager plane changes as the focused depth changes, but the (u,v) coordinate is fixed at a default position. The light field is equivalently described as a set of rectified multi-view images called "sub-aperture images", $\{x_{s,t}(u,v) = l(s,t,u,v)\}$. Here, (s,t) corresponds to the viewpoint defined on the aperture ($(s,t) \in \mathscr{A}$), where $\mathscr{A}$ is a set of small areas on the aperture plane and has M elements ($|\mathscr{A}| = M$).

In the case of the CA method, we can design an observation model by changing the semi-transparent code pattern located on the aperture plane. Let $a_n(s,t)$ be the transmittance at position (s,t) for the n-th acquisition ($n = 1,\ldots,N$). The observed image $y_n(u,v)$ is formed as

$$y_n(u,v) = \sum_{(s,t)\in\mathscr{A}} a_n(s,t) x_{s,t}(u,v) \qquad (1)$$

In the case of the FS method, we can control the focused depth for each acquisition. The observed image is

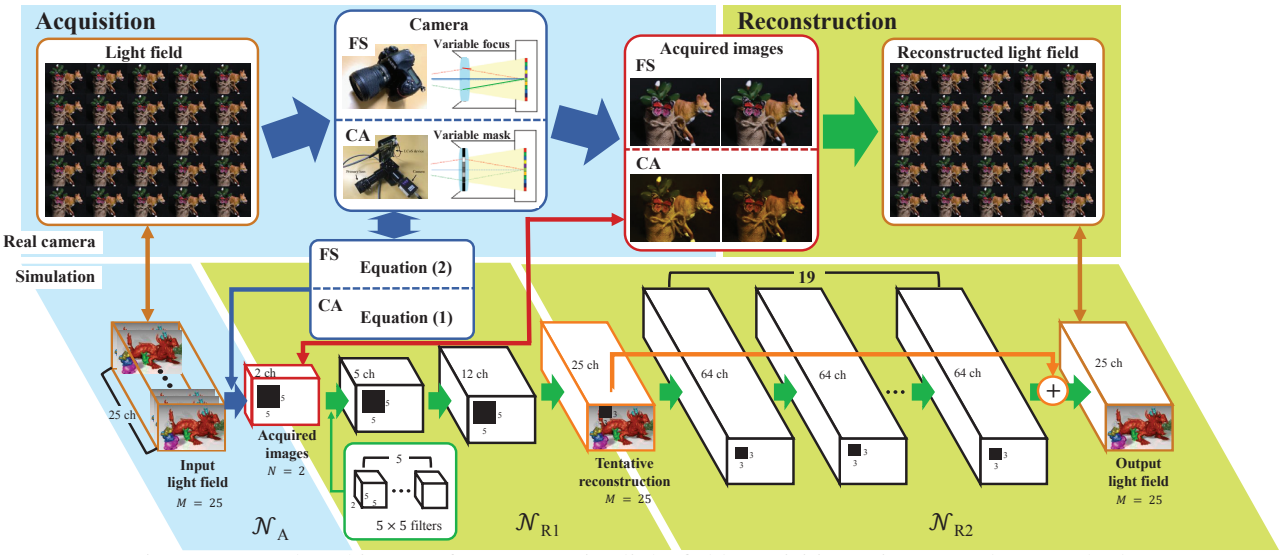


Fig. 3: Network architecture for compressive light-field acquisition using FS and CA methods.

Table. 1: Channel transitions for reconstruction networks ($\mathcal{N}_{R1}$, $\mathcal{N}_{R2}$) and disparities for FS method (d_n).

N	$\mathcal{N}_{R1}$	$\mathcal{N}_{R2}$	d_n
1	$1\to2\to5\to12\to25$	$25\to64\to64$, ..., $64\to25$	0
2	$2\to5\to12\to25$		-1, 1
3	$3\to6\to12\to25$		-1, 0, 1

modeled by shear-and-add operations as

$$y_n(u,v) = \sum_{(s,t)\in\mathscr{A}} x_{s,t}(u+d_n(s_c-s), v+d_n(t_c-t)). \quad (2)$$

where (s_c,t_c) is the central viewpoint and d_n is the amount of shear for the n-th acquisition, which corresponds to the focused depth.

Reconstructing the light field is equivalent to getting M sub-aperture images $\hat{x}_{s,t}(u,v)$ from the N given observations $y_n(u,v)$, where $\hat{x}_{s,t}(u,v)$ is an estimation of $x_{s,t}(u,v)$. In particular, we are interested in the case of $N \ll M$, where the entire light field could be reconstructed from only a few observed images.

2.2 Formulating Compressive Acquisition Using CNN

The formulation we present here is an extension from our previous work [4] that was limited to the CA method.

Optimization of compressive observation and reconstruction can be regarded as a problem of auto-encoder; M sub-aperture images $x_{s,t}(u,v)$ are once encoded to N acquired images $y_n(u,v)$, and then, they are decoded to M sub-aperture images $\hat{x}_{s,t}(u,v)$. Specifically, the observation (encoder) and reconstruction (decoder) processes can be represented as mappings

$$f : \mathscr{X} \to \mathscr{Y}, \quad g : \mathscr{Y} \to \hat{\mathscr{X}}, \quad (3)$$

where $\mathscr{X}$ represents a tensor that contains all the pixels of $x_{s,t}(u,v)$ for all $(s,t)\in\mathscr{A}$. Similarly, $\mathscr{Y}$ and $\hat{\mathscr{X}}$ correspond to $y_n(u,v)$ and $\hat{x}_{s,t}(u,v)$, respectively. The composite mapping $h = g\circ f$ should be as close to the identity as possible, under the condition that $N \ll M$. The goal of optimization is formulated with the squared error loss as

$$\arg\min_h |\mathscr{X} - \hat{\mathscr{X}}|^2 = \arg\min_h \sum_{s,t,u,v} |x_{s,t}(u,v) - \hat{x}_{s,t}(u,v)|^2. \quad (4)$$

We optimized the h mapping using a collection of training samples. In the training stage, training samples pass through the entire network. However, in a real application, the f mapping is conducted by the physical imaging process of a camera, and the acquired images are fed to the network corresponding to g, by which we can computationally reconstruct the target light field.

We implemented the composite mapping $h = g\circ f$ as a stack of 2D convolutional layers. An example with $M = 25$ and $N = 2$ is illustrated in Fig. 3. The f mapping corresponds to the network $\mathcal{N}_A$. The g mapping is decomposed into two networks, $\mathcal{N}_{R1}$ and $\mathcal{N}_{R2}$. The former network $\mathcal{N}_{R1}$ reconstructs the target light field tentatively, and the latter network $\mathcal{N}_{R2}$ refines the output of $\mathcal{N}_{R1}$. Throughout the networks, the size of images is unchanged; only the number of channels is changed. The channel for the input to $\mathcal{N}_A$ corresponds to the viewpoints M. Meanwhile, the channel for the output from $\mathcal{N}_A$ corresponds to the number of the acquired images N. Finally, the channels for the outputs from $\mathcal{N}_{R1}$ and $\mathcal{N}_{R2}$ correspond again to the viewpoints M. To better simulate the physical imaging process, Gaussian noise was added to the acquired images $y_n(u,v)$. We consider several network configurations for different values of N, which are summarized in Table 1 (left).

In the case of the CA method, the f mapping should be equivalent to Eq. (1). We implemented this using a single 2D convolutional layer with 1×1 convolution kernels, where the filter weights are limited within the range $[0,1]$. These weights correspond to the transmittance values $a_n(s,t)$, which are jointly optimized with the reconstruction networks in the training stage. Meanwhile, in the case of the FS method, the f mapping should corresponds to Eq. (2). We used a single 2D convolutional layer with 5×5 convolution kernels, where the filter weights take binary values in accordance with Eq. (2). Here, the focused depths d_n might be trainable, but in this paper we fixed them as summarized in Table 1 (right). Therefore, only the reconstruction networks are optimized in the training stage.

3. EXPERIMENTS

Figure 4 presents quantitative reconstruction quality of the FS and CA methods for different numbers of acquisitions and different datasets. We used five datasets that were not included in the training datasets. The difference is not significant, so we conclude that both methods yield satisfactory quality. Several resulting images with two acquired images are shown in Fig. 5, where top and bottom rows correspond to the FS and CA methods, respectively. From left to right on each row, two acquired images, the central view of the reconstructed light field, and the difference from the ground truth (magnified by 5) are shown. To see the disparities of the reconstructed light field, epipolar-plane images (EPIs) corresponding to green and blue lines are also presented.

Finally, we conducted experiments using real cameras that are shown in Fig. 1. For the FS method, we used a Nikon D850 camera and a AF-S Micro NIKKOR 60mm f/2.8G ED lens. The exposure time and the F number were set to 50 msec and 5.6, respectively. For the CA method, we used a camera developed in a previous work [5], where arbitrary aperture pattern can be generated using a LCoS display that was inserted in the optical path, and the exposure time was set to 40 msec. We acquired two images for both methods. Shown in Fig. 6 (a) is the acquisition setup. The light field obtained with the FS method is shown in (b). For reference, we captured the same scene with a narrow aperture (the F number was 32) as shown in (c). Shown in (d) are two acquired images and a central view of the reconstructed light field using the FS method. Shown in (e) are the counterparts using the CA method. Some close-up images are shown in (f). The output light field from the CA method was multiplied by three because it was too dark due to the limited light transmittance of the optical system. The optical system also affected the color rendition. Meanwhile, some details and disparities were better reconstructed with the CA method. We believe the overall quality of the light fields obtained with both methods was of satisfactory level.

4. CONCLUSION

In this paper, we studied two methods for efficient light field acquisition: one using a coded aperture (CA) and another using a focal stack (FS). We developed CNNs that can accept the input from both CA and FS methods and reconstruct the entire light field. Our experimental results demonstrate that we can acquire a light field with sufficient quality from only a few acquired images by either of these methods .

However, in real situations, each of the methods has its own difficulties. The CA method is weak to noise due to the intentional blocking of light rays at the aperture plane. Moreover, due to the optical system implementation using a LCoS display, the light transmittance is much lower than in theory, and the color is also distorted. Meanwhile, in the FS method, we need a mechanical movement for changing the focused depth and post-processing for correcting the zooming ratio of the acquired images. These issues should be considered in designing an efficient high-quality acquisition system.

References

[1] Toyohiro Saito, Yuto Kobayashi, Keita Takahashi, and Toshiaki Fujii: "Displaying real-world light fields with stacked multiplicative layers: Requirement and data conversion for input multiview images," Journal of Display Technology, vol. 12, no. 11, pp. 1290–1300, 2016.

[2] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," ACM Transactions on Graphics (TOG), vol. 26, no. 3, pp. 69, 2007.

[3] Anat Levin: "Linear view synthesis using a dimensionality gap light field prior," Proc. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pp. 1831–1838, 2010.

[4] Yasutaka Inagaki, Yuto Kobayashi, Keita Takahashi, Toshiaki Fujii, and Hajime Nagahara, "Learning to capture light fields through a coded aperture camera," in The European Conference on Computer Vision (ECCV), September 2018, pp. 431–448.

[5] Toshiki Sonoda, Hajime Nagahara, and Rin-ichiro Taniguchi, "Motion-invariant coding using a programmable aperture camera," IPSJ Transactions on Computer Vision and Applications, vol. 6, pp. 25–33, 6 2014.

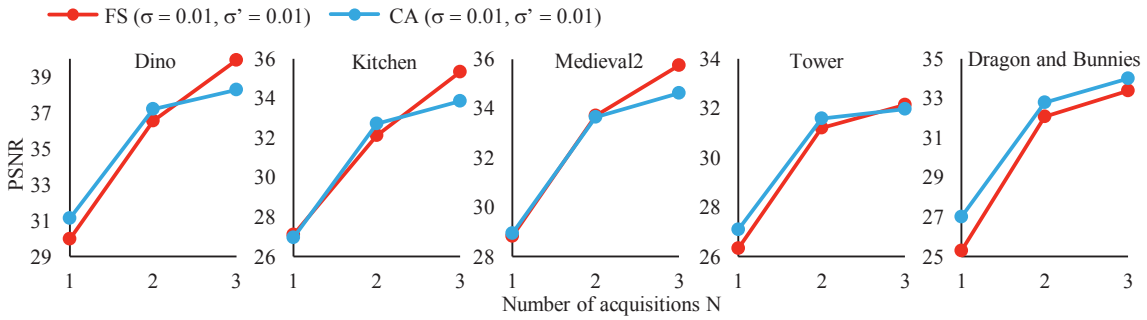


Fig. 4: Quantitative reconstruction quality of FS and CA methods for different numbers of acquisitions N and different datasets.

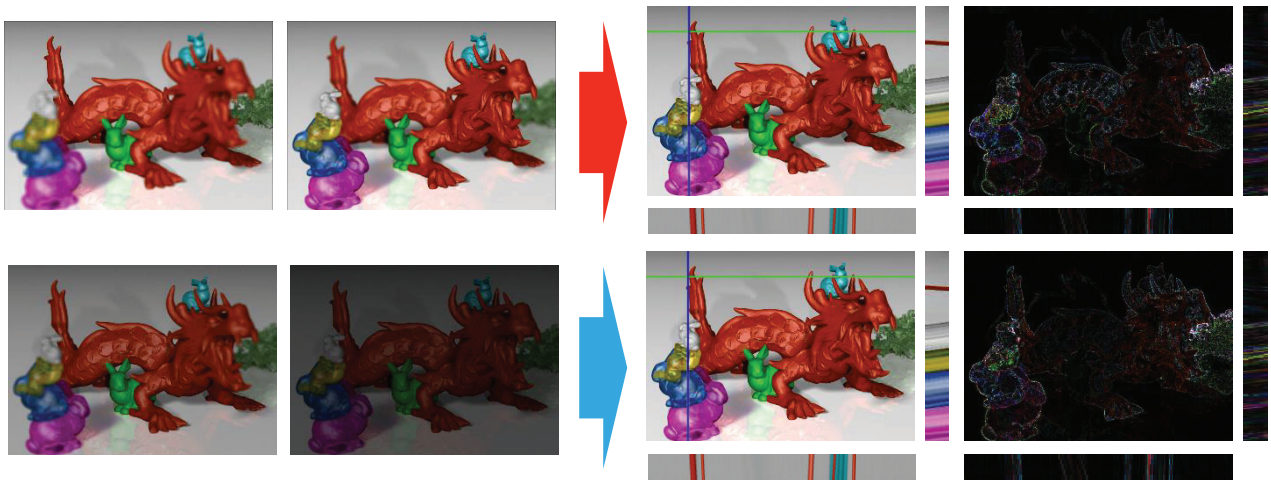
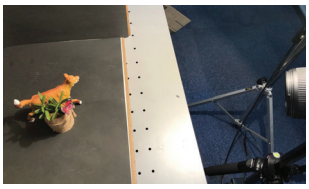


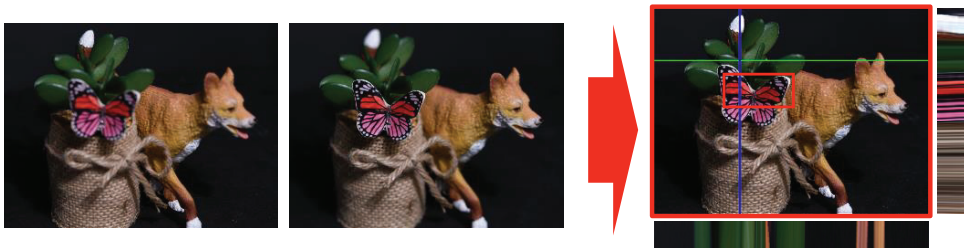
Fig. 5: Simulative experiments with FS (top) and CA (bottom) methods. Left: acquired images. Right: central images of reconstructed light fields and differences from the ground truth (magnified by 5) with EPIs corresponding to the blue and green lines.



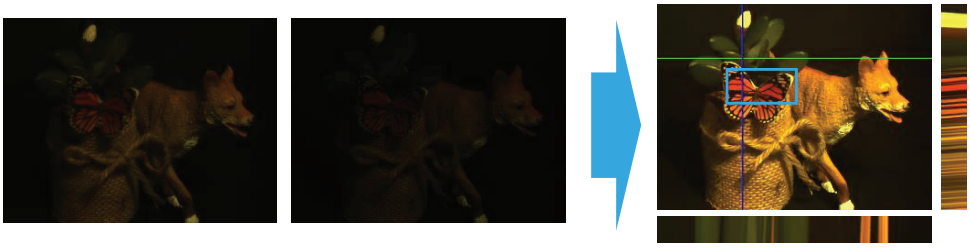
(a) Experimental setup.

(b) Reconstructed light field (FS).

(c) Center view with narrow aperture (for reference).



(d) FS method: acquired images (left) and reconstructed central view with EPIs (right).



(e) CA method: acquired images (left) and reconstructed central view with EPIs (right).

(f) Close-ups in the reference (left), FS (center), and CA (right).

Fig. 6: Experiments using real cameras.