# Light-Field Image Processing Using Deep Neural Network

**Toshiaki Fujii**

Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, JAPAN

Keywords: 3D image processing, Light field, Deep neural network

**ABSTRACT**

*In this paper, we report results of our experiments where deep neural networks (DNNs) are adopted to perform the light-field image processing. Experimental results show that we can successfully reduce the computation cost by using DNN with almost the same performance of conventional methods.*

## 1 INTRODUCTION

Three-dimensional image processing has a long history and many researches have been conducted on this topic. Among them, the capturing and displaying of light fields used to be a straightforward process and the amount of computation was not a serious problem so far. However, new technologies have emerged in this fields, such as a computational camera and a computational display, and these technologies require high computation. On the other hand, deep neural network (DNN) has been introduced and applied to various research fields and it showed high performance in various problems. The DNN could provide a solution for the problem of high computation cost for image processing researches. In this paper, we first introduce the light field concept, and then we review the state-of-the-art light field acquisition and display systems and point out that these systems require huge computation. Finally, we introduce some examples where DNN helps to solve the problem.

## 2 DEFINITION OF LIGHT FIELD

We see a 3D scene by our eyes that have the same mechanism as an optical camera, which is a device that records light rays from a 3D scene. This means that we obtain our visual information from a collection of light rays from the scene. Therefore, if we can represent a collection of light rays, we can represent 3D visual information of the scene. Ray space [1] and Light field [2] are proposed based on this notion. Figure 1 shows one of the parameterization methods of a light field [1], which is called ray space. In the light field concept, it can be seen that capturing and displaying of a 3D scene is equivalent to capturing and displaying a light field. In this paper, we deal with the problem of light field acquisition and display from this perspective.

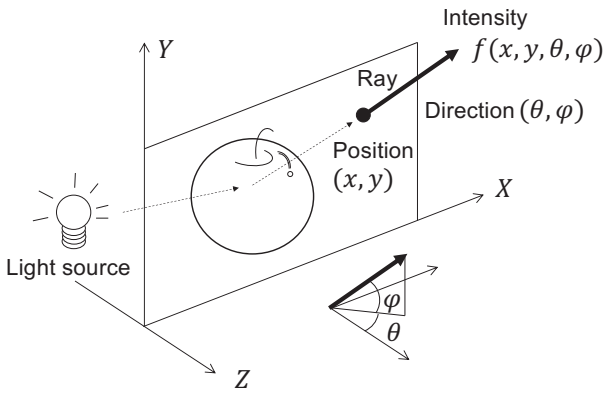


Fig. 1 Definition of ray space.

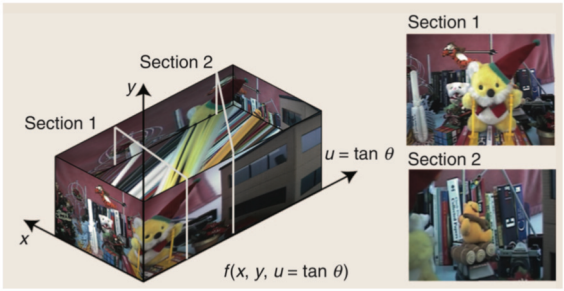


Fig. 2 Example of ray space.

## 3 DIMENSIONALITY REDUCTION OF LIGHT FIELD

A light field is equivalent to a large number of multiview images taken with very dense camera spacing. It means that a light field consists of a bunch of images of a scene taken from slightly different viewing directions. Therefore, a light field has much redundancy in both spatial and view axis directions. This characteristic has been utilized in various light-field processing, such as view interpolation, data compression, compressed sensing, and compressed display.

Based on the redundancy of a light field, we can reduce the dimension of the light field using DNN as shown in Fig. 3. The input light field is the original light field, e.g. $M = 5 \times 5$ multiview images captured by a 2D camera array. The input light field is reduced by a mapping $f$ to an intermediate representation which is composed of just a few ($N$) images. The original light field

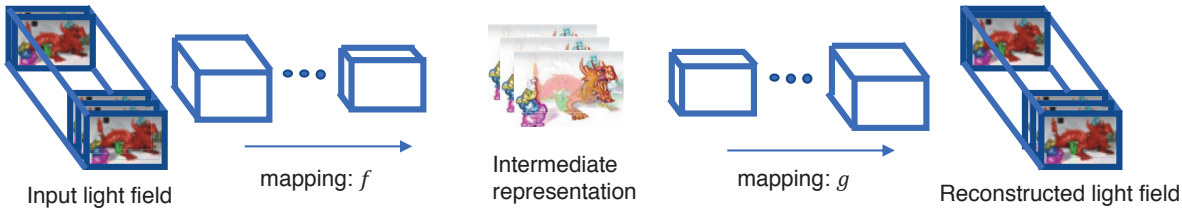


Fig. 3 Dimensionality reduction of light field.

is reconstructed from the intermediate representation by a mapping $g$ using DNN. This formulation can be viewed as an autoencoder, which is known as a type of artificial neural network used to learn efficient data coding in an unsupervised manner. The composite mapping $h = g \circ f$ should be as close to the identity as possible, under the condition that $N \ll M$.

This formulation indicates the following two things: (1) a light filed can be reduced by using a DNN to a couple of images, which include all the information on the original light field data, (2) a light field can be reconstructed by using a DNN from a few images which are captured by a specific optical system in a compressive manner. From this viewpoint, we apply a DNN computation to light field acquisition and display.

## 4 LIGHT FIELD ACQUISITION

Light field acquisition in a compressive way is formulated as follows. Input light field is captured through a specific optical system such as a coded aperture camera or a focal stack, which corresponds to the $N$ observations. This is a physical system and can be viewed as the mapping $f$ in Fig. 3. Then the original light field is reconstructed through the computation. Note that the number of observations $N$ is greatly smaller than the number of original images. In the following we take two examples in which heavy computation is required for the naïve implementation and try to reduce it by adopting DNN for the reconstruction computation of the light field data.

### 4.1 Coded Aperture Camera and Focal Stack

Here, we introduce two methods to acquire a light field in a compressive way. The first method is to use a coded aperture camera, which is equipped with a semi-transparent coded pattern (coded aperture: CA) in the optical path of a camera. The second method is to capture a focal stack (FS), which is composed of several images taken with different focused depth.

To formulate the above-mentioned acquisition methods, we introduce a coordinate system shown in Fig. 4. In the figure, a 4-D light field $l(s,t,u,v)$ is defined, where $(s,t)$ denote the viewpoint of a sub-aperture image and $(u,v)$ denotes image coordinates. In this coordinate system, we can describe a sub-aperture image as $x_{s,t}(u,v) = l(s,t,u,v)$.

In the CA case, the observed image $y_n(u,v)$ is represented as

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

where $a_n(s,t)$ is the transmission at position $(s,t)$. On the other hand, FS is represented as

$$y_{n(u,v)} = \sum_{s,t} x_{s,t}(u + d_n s, v + d_n t), \qquad (2)$$

where $d_n$ is the focused depth.

Reconstructing the light field is equivalent to reconstructing $M$ sub-aperture images $\hat{x}_{s,t}(u,v)$ from the $N$ observations $y_n(u,v)$, where $\hat{x}_{s,t}(u,v)$ is an estimation of $x_{s,t}(u,v)$. In the previous work, we formulated the reconstruction problem from the perspective of principal component analysis (PCA) and non-negative matrix factorization (NMF) [6]. From this formulation, we derived optimal non-negative aperture patterns and a straight-forward reconstruction algorithm.

### 4.2 Light Field Reconstruction using DNN

In the DNN method, we define the loss function as follows:

$$\underset{h=g\circ f}{\operatorname{argmin}} \left| x_{s,t}(u,v) - \hat{x}_{s,t}(u,v) \right|^2, \qquad (3)$$

and train the network and obtain the mapping $g$. Note that in a real application, the mapping $f$ 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. 5.

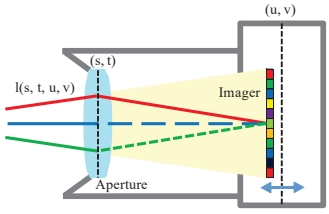


Fig. 4 Light field definition inside a camera.

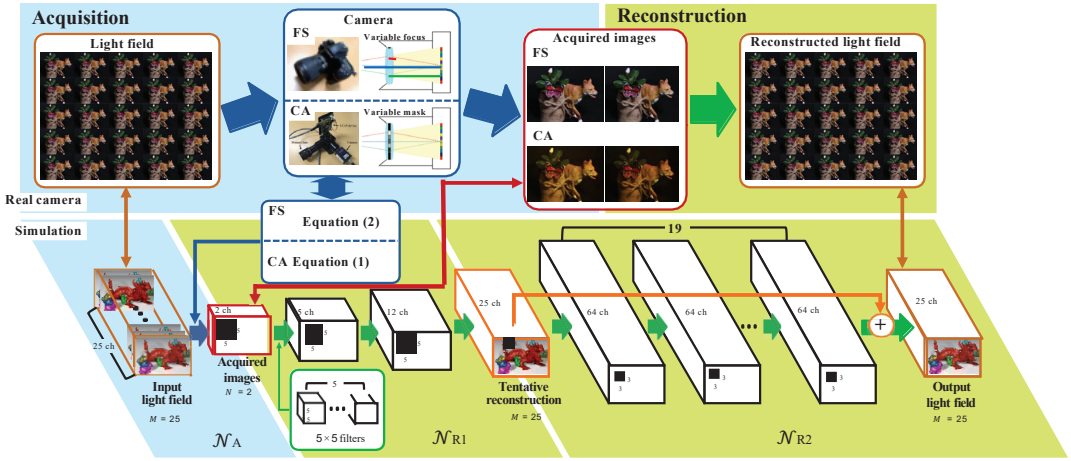


Fig. 5 Network architecture for compressive light field acquisition [9]

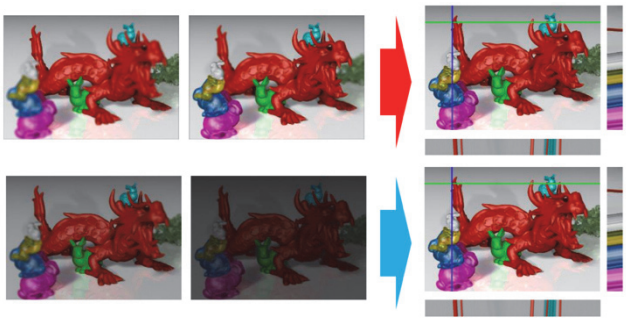

Fig. 6 Experimental results with focal stack (top) and coded aperture (bottom) methods [9].

Figure 6 shows the results of our simulation, where left images are acquired images and right images are central images of reconstructed light fields with EPIs corresponding to the blue and green lines. We can see the entire light field is well reconstructed from only $N = 2$ observations in both CA and FS cases. For more details please refer to [9].

## 5 LIGHT FIELD DISPLAY

There are various kinds of light field display including a lenslet-based display (e.g. Integral photography) and a barrier-based display. Here we focus on a stacked layer type display shown in Fig. 7. In the following, we describe how the stacked layer light field display works and show that it requires huge computation to calculate the layer patterns.

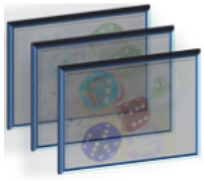

Fig. 7 Layered light field display.

### 5.1 Layered Light Field Display

In a layered light field display, each pixel of the light attenuating layers has an individual transmittance and the layer patterns are optimized so as to reproduce the light field as faithfully as possible. Let $T_n(x, y)$ denote the transmittance of the $n$-th layer and a backlight luminance $L$ be attenuated by $T_n(x, y)$. A light ray $l(s, t, x, y)$ is expressed as

$$l(s,t,x,y) = L \prod_{n=-1}^{1} T_n(x + ns, y + nt). \qquad (4)$$

The optimal $T_n(x, y)$ is derived through the following optimization:

$$\arg\min_{T_n} \sum_{s,t,x,y} \left| I_{s,t}(x,y) - \prod_{n=-1}^{1} T_n(x + ns, y + nt) \right|^2, \qquad (5)$$

where $0 \le T_n(x, y) \le 1$. This optimization is conducted through non-negative tensor factorization (NTF). Please refer to the original paper [3] for descriptions of the optimization method and the extension to time multiplexing. Since the transmittance values are alternately updated layer by layer, it requires heavy computations.

### 5.2 Calculation of Layer Patterns Using CNN

We conducted an experiment where we adopted CNN for the calculation instead of the iterative updates. Here, the mapping $g$ in Fig. 3 is implemented as the CNN. The process flow from capturing multi-view images to displaying the light field is illustrated in Fig. 8. The input to the network is the patches of $I_{s,t}(x, y)$ and output is the patches of $T_n(x, y)$. This one directional computation reduces computations and increases the calculation speed, avoiding computationally heavy iterations. As a numerical example, the calculation of layer patterns using NTF took about 12 seconds with 50 iterations,

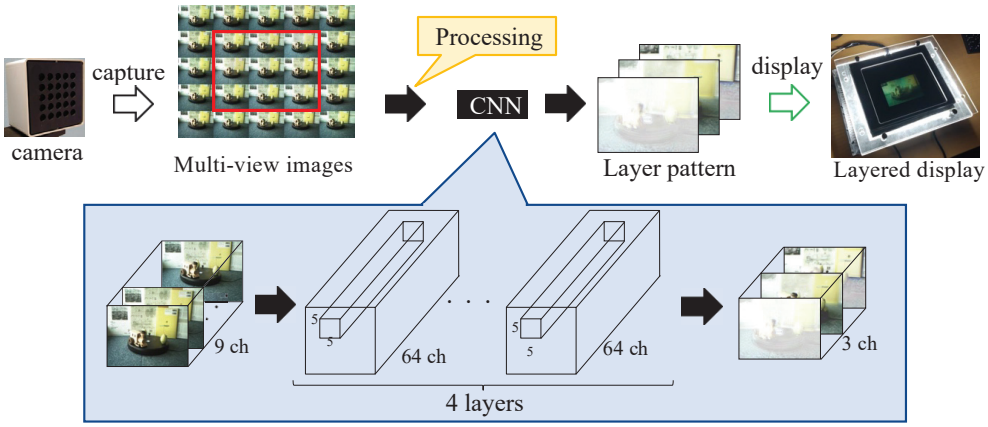


Fig. 8 Process flow from capture to display.

while CNN took 0.03 seconds in the same condition with the comparative quality (PSNR = -3 [dB]). Based on this result, we developed a full pipeline system from capture to display of the light field of a real 3D scene. In this prototype system, we used a multi-view camera (ViewPLUS ProFUSIOIN 25) to capture a light field. The $5 \times 5$ multi-view images captured by the ProFUSION25 are transferred to the PC. Rectification of the Multiview images and adjustment of convergence plane are performed on the central $3 \times 3$ views, and the appropriate layer pattern is calculated from the processed multi-view images using the CNN and fed to the layered display. For more details, please refer to [10].

## 6 SUMMARY

In this paper, we gave an overview of our light field image processing researches using deep neural networks. First, light field acquisition and display is formulated as an autoencoder. It is viewed as the composite mapping $h = g \circ f$, where $f$ is the encoder and $g$ is the decoder. For the light field acquisition, $f$ is a physical process and is implemented as a specific optical device. In this case, the mapping $g$ is implemented by using DNN. From this viewpoint, we introduced two examples, where $f$ is a case of a coded aperture camera and a focal stack. For the display, we introduced a layered light field display, where $g$ is a physical process and DNN is used to implement the mapping $f$. We developed a real-time full chain system from multiview image capturing to display. In future works, we will develop a full DNN-based system from capture to display based on these results.

## ACKNOWLEDGEMENT

This overview is based on research activities in our laboratory. We would like to thank Dr. Keita Takahashi, Dr. Ryutaroh Matsumoto, and other laboratory members.

## REFERENCES

[1] T. Fujii and M. Tanimoto, “Ray Space Coding for 3D Visual Communication,” PCS '96, 2, pp. 447-451 (1996).

[2] M. Levoy and P. Hanrahan, “Light field rendering,” SIGGRAPH ’96, pp. 31-42 (1996).

[3] G. Wetzstein, D. Lanman, M. Hirsch, and Ramesh Raskar, “Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting,” ACM Transactions on Graphics (TOG), Vol. 31, Issue 4, pp. 1-11 (2012).

[4] T. Saito, Y. Kobayashi, K. Takahashi, and T. Fujii, “Displaying Real-World Light-Fields with Stacked Multiplicative Layers: Requirement and Data Conversion for Input Multi-view Images,” IEEE/OSA JDT, Vol. 12, Issue 11, pp. 1290-1300 (2016).

[5] K. Takahashi, Y. Kobayashi, and T. Fujii, “From Focal Stack to Tensor Light-Field Display,” IEEE Trans. IP, Vol. 27, Issue 9, pp. 4571-4584 (2018).

[6] Y. Yagi, K. Takahashi, T. Fujii, T. Sonoda, and H. Nagahara, “Designing Coded Aperture Camera Based on PCA and NMF for Light Field Acquisition,” IEICE Trans. on Information and Systems, Vol. E101-D, No.9, pp.2190-2200 (2018).

[7] Y. Inagaki, Y. Kobayashi, K. Takahashi, and T. Fujii, “Learning to Capture Light Fields through A Coded Aperture Camera,” ECCV 2018, P-2B-14 (2018).

[8] Y. Kobayashi, S. kondo, K. Takahashi, and T. Fujii, “A 3-D Display Pipeline: Capture, Factorize, and Display the Light Field of a Real 3-D Scene,” ITE-MTA, Vol. 5, No. 3, pp. 88-95 (2017).

[9] Y. Inagaki, K. Takahashi, and T. Fujii, “Light Field Acquisition from Focal Stack via a Deep CNN,” IDW 2019 (2019).

[10] Y. Ota, K. Maruyama, R. Matsumoto, K. Takahashi, and T. Fujii, “Displaying Live 3-D Video from a Multi-view Camera on a Layered Display,” IDW 2019 (2019).