Displaying Light Fields from Various Input Modalities

Keita Takahashi[1]

keita.takahashi@nagoya-u.jp
[1]Nagoya University, Furo-cho, Chikusa, Nagoya, 464-8603, Japan
Keywords: Light-field display, Multi-view images, Focal stack, Coded-aperture camera

ABSTRACT

This paper summarizes our recent works on displaying dense light fields from real scenes on a light-field display. We developed processing workflows for various input modalities, including multi-view cameras, plenoptic cameras, focal stacks, and coded-aperture cameras. We also indicate a direction toward a unified framework for various modalities.

1 INTRODUCTION

True 3-D experience can be achieved by reproducing a dense light field of a target scene, which provides not only binocular depth perception but also natural motion parallax in response to the head motion. A light field is usually represented as a set of dense multi-view images, and thus, a light-field display should be capable of emitting many views into the corresponding directions simultaneously. To develop such displays, several architectures using parallax barriers, lenticular lenses, and rear projections have been proposed. Among them, we focus on a newly emerging architecture using a stack of a few semi-transparent layers [1-3]. This architecture enables us to display many views simultaneously without the resolution for each view being sacrificed. In other words, many views can be represented as a set of several layers in a compressive manner. Specifically, we followed the design of "tensor display" [3] shown in Fig. 1, and developed a prototype using three liquid crystal display (LCD) panels and a backlight.

Compared to the architecture design, less attention has been focused on the issue of how 3-D contents for such displays are created. In particular, when we aim to display a real scene, we need to prepare a light field of that scene in a suitable format for the display's architecture. This paper summarizes our recent works [4-8] on this issue. As shown in Fig. 2, we developed processing workflows for our prototype light-field display from various input modalities, including multi-view cameras, plenoptic cameras, focal stacks, and coded-aperture cameras. This paper provides an overview on the difference among these input modalities, and indicates a direction toward a unified framework for different modalities that is constructed on deep neural networks.

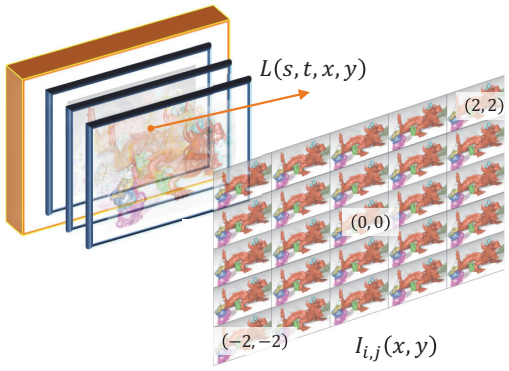

Figure 1: Light-field display with a stack of three semi-transparent layers. Multi-view images are associated with corresponding light rays.

2 Layered Light-Field Display

As shown in Fig. 1, two or three light attenuating layers such as liquid crystal display (LCD) panels are stacked in front of a backlight. With three layers, each of the outgoing light rays can be described as

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

Here, $T_n(x,y) \in \mathcal{T}$ denotes the transmittance of the *n*-th layer where *n=-1,0,1* are assigned to the rear, middle, and front layers, respectively, and L is the luminance of a uniform backlight. The outgoing direction is represented by (s,t).

To display a desired 3-D content on this display, we first need to prepare light field data consisting of a set of multi-view images, $I_{i,j}(x,y)$, where (i,j) denote the horizontal and vertical position of a viewpoint. We set the central viewpoint to $(0,0)$ without loss of generality. These images are associated with the views observed from different directions. More specifically, each image corresponds to the target light field as

$$L(i,j,x,y) \approx I_{i,j}(x,y). \quad (2)$$

The transmittance patterns for the layers $T_n(x,y) \in \mathcal{T}$ are optimized so as to satisfy the above condition as much as possible. This optimization is described as

Figure 2: Process workflow from various input modalities to layered light-field display.

$$\mathcal{T}^* = \operatorname{argmin} E(\mathcal{T}) \quad (3)$$

$$E(\mathcal{T}) = \sum_{i,j,x,y} \left\| I_{i,j}(x,y) - \prod_{n=-1}^{1} T_n(x+ni, y+nj) \right\|^2 \quad (4)$$

This optimization is formulated as a problem of non-negative factorization (NTF), and was solved by analytical methods based on iteration-based multiplicative update rules. More recently, we developed learning-based solutions using convolutional neural networks (CNNs) [7,8] as will be mentioned later.

3 Workflows for Various Input Modalities

As shown in Fig. 2, we established processing workflows for our prototype light-field display from various input modalities, including multi-view cameras, plenoptic cameras, focal stacks, and coded-aperture cameras.

3.1 Multi-view camera [4,5]

As a straight-forward approach, we first used a multi-view camera (ViewPLUS ProFUSION 25) as an input modality for our prototype display [4,5]. The camera had 25 viewpoints compactly arranged in a 5 x 5 array. The viewpoint intervals were 12 mm, which turned out to be too large for our purpose in most practical setups; we needed a denser light field to display the content with high quality. Specifically, disparities among the neighboring viewpoints should be limited within ± 1 pixels. To satisfy this requirement, we used a view interpolation (virtual view synthesis) method. For example, we interpolated (densify) the original 5 x 5 viewpoints into 17 x 17 viewpoints in accordance with the disparity range of a target scene. Then, the interpolated multi-view images were fed to the iteration-based algorithm to produce the optimized layer patterns, $T_n(x,y) \in \mathcal{T}$. We verified that the light field interpolated in this manner led to better visual quality than the case of using the original light field without interpolation.

3.2 Plenoptic camera [5]

A plenoptic camera, such as a Lytro-Illum, can also be used as an input modality [5]. Due to a micro-lens array attached on the imaging sensor, this camera can capture a light field (consisting of, e.g., 15 x 15 viewpoints) with a single shot, but the spatial resolution for each viewpoint image is reduced in return. The viewpoint intervals were sufficiently small due to the small aperture of the camera device. Therefore, no viewpoint interpolation was necessary; the obtained multi-view images are directly fed to the iterative algorithm to produce the optimized layer patterns, $T_n(x,y) \in \mathcal{T}$. On the contrary, the disparity range was more likely to be too small. The amount of disparities was directly related to the strength of 3-D sensation we perceived from the light-field display. Therefore, we needed to carefully configure the target

scene so as to produce sufficiently large disparities among the viewpoints. By doing so, we achieved visually-compelling results on our prototype display using inputs from a Lytro Illum camera.

3.3 Focal Stack [6]

The modalities mentioned so far requires specialized hardware devices for light field acquisition. We also investigated another modality that can be implemented with an ordinary camera: a focal stack, which is a set of differently focused images taken from the same viewpoint [6]. The idea of replacing a light field with a focal stack aligns with the fact that a focal stack can contain most of the information of the original light field in a compressive manner. Specifically, we need three images for the three layer-patterns, each of which focused at the depth of each layer. An image focused at the *n*-th layer is approximately represented as

$$J_n(x,y) = \sum_{i,j} I_{i,j}(x - ni, y - nj) \quad n \in \{-1,0,1\}. \qquad (5)$$

The layer patterns, $T_n(x,y) \in \mathcal{T}$, should be derived from these three images. We modified the iteration-based optimization algorithm through approximations. More specifically, the original algorithm requires each view of the original light field, $I_{i,j}(x,y)$. In contrast, our modified algorithm refers only to the focal stack consisting of $J_n(x,y)$. We evaluated the light fields displayed with our modified algorithm, and verified that the quality was almost comparable to the case with the original algorithm.

Using a focal stack as the input eliminates the need of specialized hardware. It also leads to a significant reduction of the amount of input data; we need only three differently-focused images instead of dozens of images taken from different viewpoints. Moreover, it can be considered as a desirable process flow: a direct conversion from a compressed representation (a focal stack) to another compressive representation (a set of layer patterns).

3.4 Coded-Aperture Camera [7]

The final modality mentioned in this paper is a coded-aperture camera. This type of cameras is used for depth estimation, focus synthesis, and compressive light-field acquisition. For the purpose of compressive light-field acquisition, each viewpoint is associated with a point on the aperture plane, which is coded using a semi-transparent mask pattern. Using an aperture pattern $a_n(i,j)$ for the n-th acquisition, an image acquired from this camera is written as

$$J_n(x,y) = \sum_{i,j} a_n(i,j) I_{i,j}(x,y). \qquad (5)$$

From several images, consisting of $J_n(x,y)$ acquired with different coding patterns $a_n(i,j)$, the original light field, consisting of $I_{i,j}(x,y)$, is computationally reconstructed. As shown in [9], we need only 2 to 4 acquired images to reconstruct the original light field with 5 x 5 or 8 x 8 views, where substantial compression was achieved in light-field acquisition. The key to this success was the use of deep neural networks. More specifically, we trained a convolutional neural network (CNN) using a massive amount of light field data so as to reconstruct the original light field ($I_{i,j}(x,y)$) from several acquired images ($J_n(x,y)$).

Extending the work in [9], we established a process flow from a coded aperture camera to a layered light-field display [7]. This extension is simple; the network's output changed from a light field to a set of layer patterns. The extended network can be trained in the same manner as the original, because using Eq. (1), the layer patterns are converted into the corresponding light field. In other words, the left-hand side of Eq. (4) was used as the loss function for each light-field sample during the training procedure. Once the training is finished, only a single forward inference process on the network is necessary to obtain a set of layer patterns from a set of acquired images. We validated this process flow using a real coded-aperture camera and our prototype display.

4 Towards Unified Framework [8]

The process flow mentioned in Section 3.4 includes an important suggestion; deep neural networks are useful for obtaining a set of layer patterns for a target light field. Aligned with this suggestion, we revealed that CNN-based methods can be used as the substitutes for the analytical iteration-based methods [8].

A CNN-based method is easily composed as shown in Fig. 3; we can use arbitrary architecture as far as the input and output are a light field and the corresponding set of layer patterns, respectively. In the training procedure, the left-hand side of Eq. (4) was used as the loss function for each training sample. Once the training is finished, a set of layer patterns is obtained by giving a light field as an input to the trained network.

We demonstrated that CNN-based methods yield comparable quality to the analytical iteration-based methods. An obvious advantage of CNN-based methods is the computational speed as shown in Fig. 4; although the training procedure takes significant time, a forward inference process on the trained network is usually very fast (e.g., several hundreds of milliseconds for typical setups). In contrast, analytical methods gradually optimize a set of layer patterns for a given light field in an iterative manner, which requires several seconds until convergence.

Moreover, CNN-based methods have the potential of further generalization to various input modalities. As mentioned so far, the input to the network can be either

a light field (as in [8]) or a set of images acquired using a coded-aperture camera (as in [7]). Furthermore, the input can be of any modality; it can be a set of sparse multi-view images (as in [4]) or a focal stack with arbitrary number of differently-focused images (a more general setup than [6]). In any case, the output is a set of layer patterns used for a layered light-field display, and the training loss is defined to evaluate the accuracy of the light field reconstructed from (displayed with) the layer patterns. The network is trained to learn the mapping between the input data given in a specific modality and a set of layer patterns that can accurately reproduce the target light field. We can use any network architecture, but not so much modification would be required to handle the difference of the input modalities. To conclude, it is expected that a unified framework for various input modalities can be constructed on deep neural networks.

5 Conclusions

This paper described our recent works on displaying dense light fields from real scenes on a layered light-field display. We developed several processing workflows corresponding to various input modalities, including multi-view cameras, plenoptic cameras, focal stacks, and coded-aperture cameras. We also demonstrated that CNN-based methods can be used to derive a set of layer patterns for the target light field, as the substitutes for analytical iteration-based methods. We finally indicate a direction toward a unified framework for various input modalities that is constructed on deep neural networks. Please visit our website[1] for more results and software.

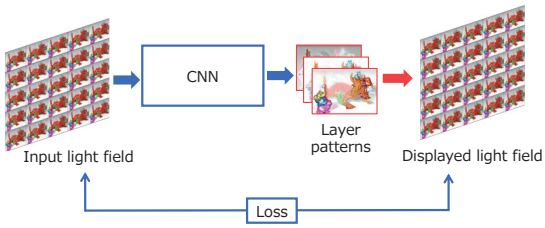

Figure 3: CNN-based method for obtaining layer patterns from light field

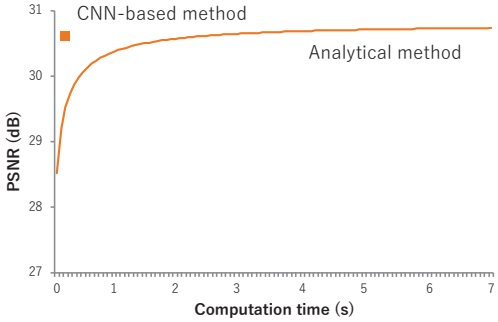

Figure 4: Reconstruction quality (in PSNR) against computational time for typical setup

REFERENCES

[1] G. Wetzstein, D. Lanman, W. Heidrich, R. Raskar: “Layered 3D: Tomographic Image Synthesis for Attenuation-based Light Field and High Dynamic Range Displays,” SIGGRAPH 2011 papers, Article No. 95 (2011).

[2] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, R. Raskar: “Polarization Fields: Dynamic Light Field Display using Multi-Layer LCDs,” SIGGRAPH Asia 2011 papers, Article No. 186, 2011.

[3] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar: “Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting,” SIGGRAPH 2012 papers, Article No. 80 (2012).

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

[5] Y. Kobayashi, S. Kondo, K. Takahashi, T. Fujii: "A 3-D Display Pipeline: Capture, Factorize, and Display the Light Field of a Real 3-D Scene", ITE Transactions on Media Technology and Applications Vol. 5, No. 3, pp. 88—95, (2017).

[6] K. Takahashi, Y. Kobayashi, T. Fujii: "From Focal Stack to Tensor Light-Field Display", IEEE Transactions on Image Processing, Vol. 27, No. 9, pp. 4571—4584 (2018).

[7] K. Maruyama, Y. Inagaki, K. Takahashi, T. Fujii, H. Nagahara: "A 3-D Display Pipeline from Coded-Aperture Camera to Tensor Light-Field Display through CNN", IEEE International Conference on Image Processing (ICIP) (2019).

[8] K. Maruyama, K. Takahashi, T. Fujii: "Comparison of Layer Operations and Optimization Methods for Light Field Display", IEEE Access, DOI: 10.1109/ACCESS.2020.2975209 (2020).

[9] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, H. Nagahara: "Learning to Capture Light Fields through a Coded Aperture Camera", European Conference on Computer Vision (ECCV) (2019)

[1] https://www.fujii.nuee.nagoya-u.ac.jp/Research/LFDisplay