From Focal Stack to Tensor Light-Field Display

Keita Takahashi, *Member, IEEE*, Yuto Kobayashi, *Student Member, IEEE*, and Toshiaki Fujii, *Member, IEEE*

***Abstract*— We propose a method of using a focal stack, i.e., a set of differently focused images, as the input for a novel light field display called a "tensor display." Although this display consists of only a few light attenuating layers located in front of a backlight, it can be viewed from many directions (angles) simultaneously without the resolution of each viewing direction being sacrificed. Conventionally, a transmittance pattern is calculated for each layer from a light field, namely, a set of dense multi-view images (typically dozens) that are to be observed from different directions. However, preparing such a massive amount of images is often cumbersome for real objects. We developed a method that does not require a complete light field as the input; instead, a focal stack composed of only a few differently focused images is directly transformed into layer patterns. Our method greatly reduces the cost of acquiring data while also maintaining the quality of the output light field. We validated the method with experiments using synthetic light field data sets and a focal stack acquired by an ordinary camera.**

***Index Terms*— Light field, 3D display, focus.**

I. Introduction

3-D DISPLAYS have been the subject of study for several years [1]–[5]. These displays can be categorized on the basis of several criteria, such as the necessity of wearing glasses and the number of supported viewing directions. Glasses-free (naked-eye) displays have attracted attention because they enable a more natural viewing experience than glasses-based ones. Multi-view displays have more potential than the conventional stereo-only displays because they not only provide depth perception by showing different images to the left and right eyes but also present natural motion parallax along with the movement of observers.

To develop glasses-free multi-view displays, researchers have devised several methods, including those that use parallax barriers [1], [6]–[8], specially designed lenses (lenticular screens or integral photography lenses) [2], [3], [9]–[11], and stacked layers [12]–[16]. In this paper, we focus on the third method, which is based on a few light-attenuating layers [13], [16]. This type of display, called a "tensor display," can be viewed from many directions (angles) simultaneously without the resolution of each viewing direction being sacrificed, which is deemed as one of the desirable properties for glasses-free multi-view displays.

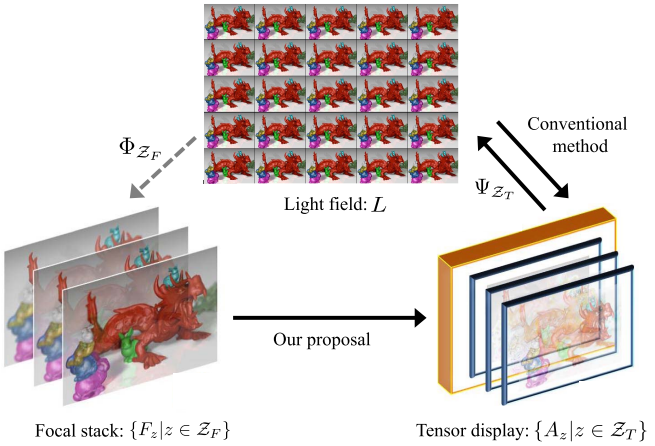


Fig. 1. Relationship between light field, focal stack, and tensor display. Operators $\Phi_{\mathcal{Z}_F}$ and $\Psi_{\mathcal{Z}_T}$ denote transforms among them. Conventional method uses entire light field for obtaining layer patterns for tensor display. Our method requires only focal stack for this purpose.

The structure of a typical tensor display is illustrated on the right in Fig. 1. A few light attenuating layers, where each pixel on each layer has an individual transmittance, are stacked in front of a backlight. Depending on the viewing direction, these layers overlap with different degrees in how much the layers overlap each other, so the displayed images are direction-dependent. For an object to be displayed in 3-D with this structure, the transmittance patterns of layers should be designed so as to make the direction-dependent views consistent with the 3-D appearance of the object. More precisely, many images or a light field [17], [18] (on the top in Fig. 1), which are expected to be observed from different viewing directions, are given as the input, and then, the layer patterns are optimized so as to reproduce the light field as faithfully as possible. This optimization is conducted through non-negative tensor factorization (NTF), where the transmittance values are alternately updated layer by layer. Although there are only a few layers, the optimized layers can reproduce the original light field with reasonable quality. This means that these layer patterns, few in number, contain information that is approximately equivalent to the original light field, which consists of many images. Therefore, this type of display is also called a "compressive display." The same structure has also been adopted for projection-based or near-eye displays [19], [20].

Visualizing real world 3-D scenes with a tensor display presents a challenge in terms of acquiring data because a dense light field, i.e., a set of multi-view images (typically dozens of images) with very small viewpoint (or viewing-direction) intervals, is required as the input [21]. However, given the fact

Manuscript received August 15, 2017; revised March 12, 2018 and April 19, 2018; accepted May 15, 2018. Date of publication May 22, 2018; date of current version June 15, 2018. This work was supported by JSPS Kakenhi under Grant 15H05314. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick Le Callet. *(Corresponding author: Keita Takahashi.)*

The authors are with the Graduate School of Engineering, Nagoya University, Nagoya 464-8603, Japan (e-mail: keita.takahashi@nagoya-u.jp).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes videos that show several experimental results. The total size of the videos is 9.98 MB. Contact keita.takahashi@nagoya-u.jp for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2839263

that this massive amount of data will ultimately be compressed into only a few layer patterns, preparing a complete light field in the first place seems to be redundant. In this paper, we demonstrate that a focal stack, which is composed of only a few differently focused images, as shown on the left in Fig. 1, can be used as the input to this display instead of a complete light field. More specifically, if we use three semi-transparent layers for the display, we only need three images, where each image is focused on each layer. While greatly reducing the cost of data acquisition, our method can maintain the quality of the output light field, as will be experimentally demonstrated by using several synthetic light field datasets. We also applied our method to a focal stack acquired by an ordinary camera and confirmed that a sufficient amount of motion parallax can be reproduced only from the focal stack. The resulting light field was displayed on a real prototype display we developed [22], [23] to confirm that natural 3-D perception is possible with our method.

In previous pieces of work, focal stacks were used for light field representation and depth estimation [24]–[29]. However, to our knowledge, our proposal of using a focal stack directly as the input to this type of light-field display is a novel and original contribution. A preliminary version of this paper was presented at a conference [30]. A more complete description, thorough discussions, and additional experimental results are included in the present paper. Moreover, we made our software public on our website [31] along with supplementary videos to encourage prospective research in this field.

II. Parameters and Models

The parameters and models that are used to explain our proposal are given in this section.

A. Light Field Parameterization

A light field is defined as a 4-D function to describe all the light rays that travel straight in a free space [17], [18]. In this paper, we adopt a plane + angle parameterization, as shown in Fig. 2. A reference plane ($z = 0$) is defined, and a light ray is parameterized by the point of intersection with the reference plane [(u, v)] and the outgoing direction with respect to the z axis [(θ, ϕ)]. The luminance of each light ray is described as $L(u, v, s, t)$ with $s = \tan(\theta)$ and $t = \tan(\phi)$. We assume that all of the elements of $L(u, v, s, t)$ take non-negative values because the light intensity is non-negative.

The reference plane can be located anywhere theoretically, but for convenience, it is located in parallel with the layers of a tensor display. Specifically, we place the reference plane at the central layer when we use three layers. We use the same reference plane for modeling a focal stack.

B. Modeling a Focal Stack

Next, we can introduce a process called *refocusing*, which is a re-parameterization of the light field using another plane located at depth z as the new reference plane [24]. The refocused light field $L(u, v, s, t; z)$ is given as

$$L(u, v, s, t; z) = L(u - zs, v - zt, s, t). \tag{1}$$

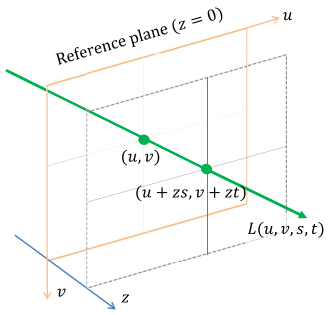


Fig. 2. Light field parameterization. Each light ray is parameterized by intersection with reference plane (u, v) and outgoing direction (s, t). It intersects with plane located at depth z (shown by dotted lines) on $(u + zs, v + zt)$.

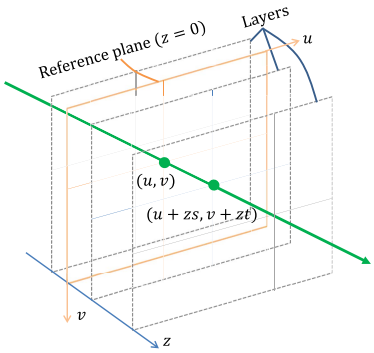


Fig. 3. Configuration of tensor display, where light attenuating layers are evenly spaced in parallel to reference plane.

Using this re-parameterization, an image focused at depth z is simply represented as

$$F_z(u, v) = \iint_{\mathcal{S}\times\mathcal{T}} L(u, v, s, t; z) ds dt, \tag{2}$$

where $\mathcal{S}$ and $\mathcal{T}$ are the effective ranges of s and t. A physical interpretation of this equation is as follows. All light rays that pass through a single point (u, v) at depth z and go into various directions (s, t) gather over a finite aperture range $\mathcal{S} \times \mathcal{T}$ to produce a single pixel value $F_z(u, v)$. This serves as a model of a focused image that is taken by a camera aimed at the reference plane perpendicularly. A focal stack is represented as a series of focused images described as $\{F_z(u, v) | z \in \mathcal{Z}_F\}$, where $\mathcal{Z}_F$ denotes a set of focus depths. Given a set of depths $z \in \mathcal{Z}_F$, generating a focal stack from a light field $L(u, v, s, t)$ is a deterministic process and can be written by using an operator $\Phi_{\mathcal{Z}_F}$ as follows.

$$\{F_z | z \in \mathcal{Z}_F\} = \Phi_{\mathcal{Z}_F}(L) \tag{3}$$

C. Modeling a Tensor Display

The structure of a tensor display [13], [16] is illustrated in Fig. 3, where a few light attenuating layers are stacked with consistently spaced intervals in front of a backlight. Let us consider a light ray passing through point (u, v) on the reference plane and going in the direction of (s, t). We can see that the intersection of this light ray with a layer located at depth z is $(u + zs, v + zt)$. Therefore, we can describe the

light rays emitted from the display as

$$L(u, v, s, t) = \prod_{z \in \mathcal{Z}_T} A_z\,(u + zs, v + zt)\, L_0, \tag{4}$$

where $A_z(u, v)$ denotes the transmittance of a layer located at depth z, $\mathcal{Z}_T$ denotes a set of depths where the layers are located, and L_0 is the luminance of the backlight and can be omitted under the assumption that the light intensity is normalized. Generating a light field L from given layer patterns $\{A_z(u, v) | z \in \mathcal{Z}_T\}$ is a deterministic process and is described with an operator Ψ as

$$L = \Psi(\{A_z | z \in \mathcal{Z}_T\}). \tag{5}$$

Inversely, to obtain the layer patterns $\{A_z(u, v) | z \in \mathcal{Z}_T\}$ from a given light field L, one needs to solve the least squares problem given as

$$\begin{aligned} &\arg\min_{\{A_z | z \in \mathcal{Z}_T\}} \iiiint_{\mathcal{U}\times\mathcal{V}\times\mathcal{S}\times\mathcal{T}} |L - \Psi(\{A_z | z \in \mathcal{Z}_T\})|^2 du\,dv\,ds\,dt \\ &\text{s.t. } 0 \le A_z(u, v) \le 1, \end{aligned} \tag{6}$$

where $\mathcal{U}$, $\mathcal{V}$, $\mathcal{S}$, and $\mathcal{T}$ are the effective ranges of u, v, s, and t, respectively. In the discretized domain, this minimization can be formulated as non-negative tensor factorization [13], [16], where each layer pattern is alternatively optimized through multiplicative update rules.

As analyzed in [16], increasing the number of layers will in theory lead to better quality of the displayed light field. However, in practice, the number of layers is typically assumed to be 2 or 3 due to implementation issues.[1] Another way to improve the visual quality is time-multiplexing [13], [16], where different sets of layer patterns, $\{A_{z,m}(u, v) | z \in \mathcal{Z}_T\}$, are alternatively displayed so that the human eyes can perceive their average over time. With M folds of time-multiplexing, Eq. (4) is rewritten as

$$L(u, v, s, t) = \frac{1}{M} \sum_{m=1}^{M} \prod_{z \in \mathcal{Z}_T} A_{z,m}\,(u + zs, v + zt)\, L_0. \tag{7}$$

As reported in [13] and [16], using a large M, e.g., 12, significantly improves the visual quality. However, to implement time-multiplexing on real hardware, we have to use layer devices with a very fast refresh rate and synchronize all of them; otherwise, flickering artifacts over time are perceived by human eyes. Due to this hardware requirement, M is practically limited to 2 or 3. Considering these hardware issues, we believe tensor displays without time-multiplexing still have value despite the limited visual quality.[2] In the remainder of this paper, we assume that no time-multiplexing is used unless otherwise noted.

A more advanced configuration using directional backlighting was also reported in [16], where further improvement to visual quality can be achieved with the help of time-multiplexing. A directional backlight can also be regarded as a lower resolution light field display by itself. Therefore, placing it behind the layers is equivalent to synthesizing two light field displays into one. In this paper, we focus on the basic setup described as Eq. (4) to keep the discussion simple and leave this advanced configuration for future work.

D. Discrete Coordinate System

Although most of the formulations are given in continuous forms in this paper, the data we actually process are discrete. Typically, what is treated as the light field is a set of multi-view images $\{I_{i,j}(x, y)\}$, where two integers (i, j) denote the 2-D index of the images, which takes $(0, 0)$ for the central image, and the other two integers (x, y) denote the discrete pixels on each image. Assuming that these images are (or can be approximated to be) captured by orthographic cameras[3] and that (i, j) corresponds to the viewing directions arranged in regular intervals, we simply associate the images with the light field of the tensor display as

$$L(\Delta_s x, \Delta_s y, \Delta_d i, \Delta_d j) = I_{i,j}(x, y), \tag{8}$$

where Δ_s and Δ_d denote the sampling intervals for the spatial and directional domains, respectively. The spatial sampling interval Δ_s is determined by the pixel size on the layers. Meanwhile, Δ_d is configured so as to describe all light rays that pass through the layers on the integer pixel positions. Therefore, we have the relation

$$\Delta_d = \frac{\Delta_s}{\Delta_z}, \tag{9}$$

where Δ_z denotes the interval among the layers.

Using this discretization, Eqs. (1) is rewritten as

$$\begin{aligned} &L(\Delta_s x, \Delta_s y, \Delta_d i, \Delta_d j; z) \\ &\quad = L(\Delta_s(x - (z/\Delta_z)i), \Delta_s(y - (z/\Delta_z)j), \Delta_d i, \Delta_d j) \\ &\quad = I_{i,j}(x - (z/\Delta_z)i, y - (z/\Delta_z)j), \end{aligned} \tag{10}$$

with which a focal stack is defined. Similarly, for a tensor display, Eq. (4) is rewritten as

$$\begin{aligned} &L(\Delta_s x, \Delta_s y, \Delta_d i, \Delta_d j) \\ &\quad = \prod_{z \in \mathcal{Z}_T} A_z\,(\Delta_s(x + (z/\Delta_z)i), \Delta_s(y + (z/\Delta_z)j))\, L_0. \end{aligned} \tag{11}$$

It should be noted that, in both Eqs. (10) and (11), the depth z is used in the normalized form z/Δ_z. In fact, the depth z always appears in this form throughout the paper if the equations are rewritten in the discrete coordinate system mentioned above. Therefore, the depth values are represented in this normalized form when we mention experimental setups.

[1] One issue is the computational cost. Another and more crucial issue is the brightness. Semi-transparent layers are typically implemented with LCD panels, whose maximum transmittance is much less than 50%. Therefore, stacking many layers results in a further loss of brightness, making the display system impractical.

[2] Our prototype display mentioned later supports only 60 or 75 Hz, with which time-multiplexing is impractical.

[3] Orthographic cameras are more suitable for this display than projective cameras because, with orthographic cameras, we need to handle only the light rays that pass through integer pixel positions on the display's layers. An extension to projective cameras has been reported in [32] with limited results.

III. Algorithm

We first derive an algorithm for obtaining the layer patterns of a tensor display from a given light field, which is essentially equivalent to the multiplicative update rule in [13] and [16]. Then, by modifying the process of iterative updates, we propose a novel algorithm that requires only a focal stack as the input.

A. From a Light Field

Given a light field that should be emitted from the display, $L(u, v, s, t)$, the goal of optimizing layer patterns is described as Eq. (6), or equivalently rewritten as

$$\begin{aligned}\arg\min_{\{A_z|z\in\mathcal{Z}_T\}} &\iiiint_{\mathcal{U}\times\mathcal{V}\times\mathcal{S}\times\mathcal{T}} |L(u,v,s,t)-\bar{L}(u,v,s,t)|^2 dudvdsdt\\ &\bar{L}(u,v,s,t)=\prod_{z\in\mathcal{Z}_T} A_z\left(u+zs, v+zt\right)\\ \text{s.t. }\ &0\le A_z(u,v)\le 1,\end{aligned} \tag{12}$$

where the light field generated by the display is denoted as $\bar{L}(u, v, s, t)$. This optimization is non-convex and cannot be solved in a closed form. Therefore, an alternative approach is adopted to solve it.

Suppose that we want to obtain the pattern for a specific layer $A_z(u, v)$ under the assumption that the other layer patterns $A_{z'}(u, v)$ ($z' \in \mathcal{Z}_T\backslash\{z\}$) are known. By refocusing $L(u, v, s, t)$ and $\bar{L}(u, v, s, t)$ on depth z, we can transform Eq. (12) into

$$\begin{aligned}\arg\min_{A_z} &\iiiint_{\mathcal{U}\times\mathcal{V}\times\mathcal{S}\times\mathcal{T}} |L(u,v,s,t;z)-\bar{L}(u,v,s,t;z)|^2 dudvdsdt\\ &\bar{L}(u,v,s,t;z)=\tilde{A}_z(u,v,s,t)A_z(u,v)\\ &\tilde{A}_z(u,v,s,t)=\prod_{z'\in\mathcal{Z}_T\backslash\{z\}} A_{z'}(u+(z'-z)s, v+(z'-z)t),\end{aligned} \tag{13}$$

where only $A_z(u, v)$ is unknown and the other known layer patterns are gathered into a single term $\tilde{A}_z(u, v, s, t)$. Equation (13) can be solved for each pixel (u, v) in a closed form as

$$A_z(u,v)=\frac{\iint_{\mathcal{S}\times\mathcal{T}} L(u,v,s,t;z)\tilde{A}_z(u,v,s,t)dsdt}{\iint_{\mathcal{S}\times\mathcal{T}} |\tilde{A}_z(u,v,s,t)|^2 dsdt}. \tag{14}$$

When all of the elements in $A_{z'}(u, v)$ ($z' \in \mathcal{Z}_T\backslash\{z\}$) are non-negative, the left hand-side of Eq. (14) is always non-negative. After Eq. (14) is applied, all of the elements of $A_z(u, v)$ are clipped to $[\epsilon, 1]$ (ϵ is a sufficiently small positive number) to satisfy $0 \le A_z(u, v) \le 1$.

On the basis of the above, we can derive an algorithm (Algorithm 1) that optimizes the layer patterns for a given light field $L(u, v, s, t)$. Although it looks different at first glance, this algorithm is completely the same as the multiplicative update rule used in [16] when time-multiplexing is disabled. See Appendix A for more details.

Algorithm 1 Obtain Layer Patterns From Given Light Field

```
Input: L(u, v, s, t)
Output: A_z(u, v) (z ∈ Z_T)
Initialize A_z(u, v) (z ∈ Z_T) with random numbers in [0, 1]
Do until convergence
    For each z ∈ Z_T
        Update A_z(u, v) using Eq. (14)
    End
End
```

B. From a Focal Stack

Here, we propose a novel method for optimizing layer patterns for a tensor display that uses a focal stack as the only input. Our method is derived by simplifying Eq. (14). First, for the denominator of Eq. (14), there must be a $K(u, v)$ for each (u, v) that satisfies

$$\iint_{\mathcal{S}\times\mathcal{T}} |\tilde{A}_z(u,v,s,t)|^2 dsdt = K(u,v)\iint_{\mathcal{S}\times\mathcal{T}} \tilde{A}_z(u,v,s,t)dsdt, \tag{15}$$

because both hand-sides are given by the definite integrals of non-negative functions. At a convergence point, the light field given as the input should closely be approximated by the light field generated by the display. This condition is described as follows.

$$L(u,v,s,t;z) \simeq \tilde{A}_z(u,v,s,t)A_z(u,v) \tag{16}$$

Using this relation and Eq. (15), the numerator of Eq. (14) is rewritten as follows.

$$\begin{aligned}&\iint_{\mathcal{S}\times\mathcal{T}} L(u,v,s,t;z)\tilde{A}_z(u,v,s,t)dsdt\\ &\quad\simeq A_z(u,v)\iint_{\mathcal{S}\times\mathcal{T}} |\tilde{A}_z(u,v,s,t)|^2 dsdt\\ &\quad= K(u,v)\iint_{\mathcal{S}\times\mathcal{T}} A_z(u,v)\tilde{A}_z(u,v,s,t)dsdt\\ &\quad\simeq K(u,v)\iint_{\mathcal{S}\times\mathcal{T}} L(u,v,s,t;z)dsdt\end{aligned} \tag{17}$$

Combining Eqs. (15) and (17), we can finally approximate Eq. (14) as

$$\begin{aligned}A_z(u,v) &\simeq \frac{\iint_{\mathcal{S}\times\mathcal{T}} L(u,v,s,t;z)dsdt}{\iint_{\mathcal{S}\times\mathcal{T}} \tilde{A}_z(u,v,s,t)dsdt}\\ &= \frac{F_z(u,v)}{\iint_{\mathcal{S}\times\mathcal{T}} \tilde{A}_z(u,v,s,t)dsdt},\end{aligned} \tag{18}$$

where the numerator is replaced with the image focused on depth z in accordance with Eq. (2). Similarly to Eq. (14), we can see that when all of the elements of $A'_z(u, v)$ ($z' \in \mathcal{Z}_T\backslash\{z\}$) are non-negative, $A_z(u, v)$ obtained with Eq. (18) never becomes negative. After Eq. (18) is applied, all of the elements of $A_z(u, v)$ are clipped to $[\epsilon, 1]$ (ϵ is a sufficiently small positive number) to satisfy $0 \le A_z(u, v) \le 1$.

Algorithm 2 Obtain Layer Pattern From Focal Stack

Input: $F_z(u,v)$ $(z \in \mathcal{Z}_T)$
Output: $A_z(u,v)$ $(z \in \mathcal{Z}_T)$
Initialize $A_z(u,v)$ $(z \in \mathcal{Z}_T)$ with random numbers in $[0,1]$
Do until convergence
  For each $z \in \mathcal{Z}_T$
    Update $A_z(u,v)$ using Eq. (18)
  End
End

Equation (18) indicates that the original light field $L(u,v,s,t)$ is no longer necessary—only the image $F_z(u,v)$ is required to obtain the numerator of Eq. (18).

Accordingly, we can modify the previous algorithm into Algorithm 2, which requires a focal stack $F_z(u,v)$ $(z \in \mathcal{Z}_T)$ as the input instead of a light field $L(u,v,s,t)$. This significantly reduces the cost of data acquisition because a focal stack consists of only a few images (the same number as the display layers), while a light field typically consists of dozens of images. Moreover, Algorithm 2 requires less computational cost than Algorithm 1 for each iteration.

Equation (18) can be interpreted in another way. From Eq. (13), the ideal condition that should be satisfied at a convergence point is described as

$$L(u,v,s,t;z) = \bar{L}(u,v,s,t;z), \tag{19}$$

which means that the input light field (the left hand-side) should be equivalent to that reproduced by the display (the right hand-side). Integrating both hand-sides over $(s,t) \in \mathcal{S} \times \mathcal{T}$, we can derive a necessary condition of Eq. (19) as follows.

$$\iint_{\mathcal{S}\times\mathcal{T}} L(u,v,s,t;z)dsdt = \iint_{\mathcal{S}\times\mathcal{T}} \bar{L}(u,v,s,t;z)dsdt \tag{20}$$

The left hand-side is equivalent to the focused image $F_z(u,v)$, which is given as the input. The right hand-side is a virtually refocused image, which is synthesized from the light field emitted from the display. Using the relation shown in Eq. (13), Eq. (20) can be rewritten as

$$F_z(u,v) = A_z(u,v) \iint_{\mathcal{S}\times\mathcal{T}} \tilde{A}_z(u,v,s,t)dsdt, \tag{21}$$

which is equivalent to Eq. (18).

C. Depth Range Compression

Our method introduced in Section III-B requires a focal stack in which each image is focused on each of the layers of a tensor display. If we have a light field, an appropriate focal stack can easily be computed using Eq. (2); all we need is to use the same set of depths as those for the tensor display. However, if we want to use a focal stack captured by a real camera, the situation is a little tricky. It is often the case that, when capturing a focal stack, we want to use a wider range of focus than the physical thickness of the tensor display. Specifically, in our experiment, the distance between the focus depths was on the order of 10 cm, while the interval between the layers was only 8 mm. Even in this case, our method treated data as if each image in the focal stack was focused on each of the layers, and thus, the depth range was virtually compressed in the display space. Such compression is commonly used in the field of 3-D displays, both to compensate for the limited capabilities of the display hardware in terms of depth range and to prevent visual fatigue caused by too large of a parallax.

IV. Discussion and Analysis

The diagram in Fig. 1 describes the relation between the light field L, focal stack $\{F_z | z \in \mathcal{Z}_F\}$, and tensor display $\{A_z | z \in \mathcal{Z}_T\}$. The light field is a complete description of the light rays that concern our configuration. In previous pieces of work [13], [16], the layer representation for a tensor display is calculated from the light field. Meanwhile, our method provides a direct path from the focal stack to the tensor display, under the condition that the two representations share the same depth sets, i.e., $\mathcal{Z}_F = \mathcal{Z}_T$, which eliminates the need for the complete light field. Our method might seem somewhat infeasible at first glance because the focal stack does not contain the complete information of the light field. However, as will be demonstrated in Section V, our method can achieve reasonable image qualities that are comparable to those that were derived directly from the original light field.

The key to understanding this result is the *depth selectivity* of the focal stack and tensor display. The information preserved in the focal stack $\{F_z | z \in \mathcal{Z}_F\}$ is *depth-selective* because, in each image, the details of objects at the focus depth are well preserved, but they are gradually lost as the objects' depth diverges from the focus depth. The information presented by the tensor display is also *depth-selective* because the objects located near one of the layers are clearly visualized, while they become blurry as they diverge from the layers. If the focal stack $\{F_z | z \in \mathcal{Z}_F\}$ and the tensor display $\{A_z | z \in \mathcal{Z}_T\}$ share the same set of depths, i.e., $\mathcal{Z}_F = \mathcal{Z}_T$, they also share essentially equivalent information, and thus, direct conversion from the focal stack to the tensor display is not only reasonable but also ideal from the perspective of minimizing the data required as the input.

To make the above statement more quantitative, we will discuss depth selectivity in the frequency domain. We then present analytical experiments to validate the statements.

A. Depth in the Frequency Domain

We analyze a light field that is generated by a Lambertian planar surface located at depth z. The texture on the surface is described as $o_z(u,v)$, and the light field emitted from the surface is denoted as $L_z(u,v,s,t)$. We can easily derive a relation written as

$$L_z(u,v,s,t) = o_z(u+zs, v+zt). \tag{22}$$

The Fourier transform of $L_z(u,v,s,t)$ is given as

$$\hat{L}_z(\omega_u,\omega_v,\omega_s,\omega_t) = \hat{o}_z(\omega_u,\omega_v)\delta(\omega_s - z\omega_u, \omega_t - z\omega_v), \tag{23}$$

where $\hat{o}_z(\omega_u,\omega_v)$ is the Fourier transform of $o_z(u,v)$ and $\delta(\cdot,\cdot)$ is the Dirac delta function. See Appendix B for the

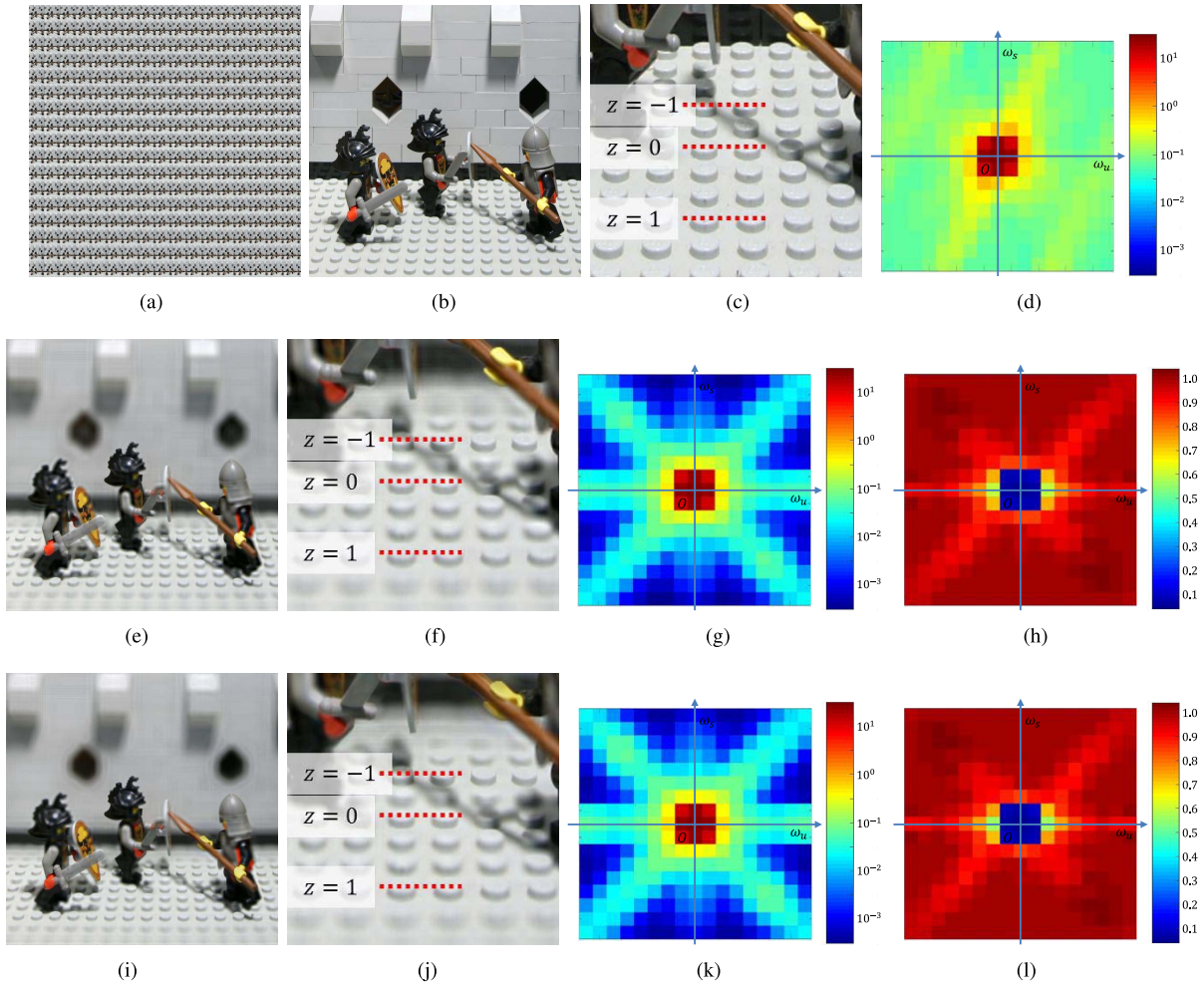


Fig. 4. Analytical experiment showing what information can be preserved/presented by focal stack and tensor display. Refer to text for details. (a) Original light field L. (b) Center view of (a). (c) Close-up of (b). (d) L in freq. domain. (e) Center view of $\bar{L}_F$. (f) Close-up of (e). (g)$\bar{L}_F$ in freq. domain. (h) err_F in freq. domain. (i) Center view of $\bar{L}_T$. (j) Close-up of (i). (k) $\bar{L}_T$ in freq. domain. (l) err_T in freq. domain.

derivation. Equation (23) means that the spectral support is bounded to the subspace that satisfies $\omega_s = z\omega_u$ and $\omega_t = z\omega_v$ in the 4-D frequency space $(\omega_u, \omega_v, \omega_s, \omega_t)$. This relation indicates how objects' depth z corresponds to the spectral information of the 4-D light field.

Similar analyses on light fields have already been presented in different contexts [16], [25], [27], [33], but we use this analysis for describing the correspondence between the focal stack and tensor display. Using the relation above, we can rewrite the *depth selectivity* of the focal stack and tensor display as follows.

- The focal stack $\{F_z | z \in \mathcal{Z}_F\}$ contains the information *mainly* along the subspace satisfying $\omega_s = z\omega_u$ and $\omega_t = z\omega_v$ for $z \in \mathcal{Z}_F$.
- The tensor display $\{F_z | z \in \mathcal{Z}_T\}$ can present the information *mainly* along the subspace satisfying $\omega_s = z\omega_u$ and $\omega_t = z\omega_v$ for $z \in \mathcal{Z}_T$.

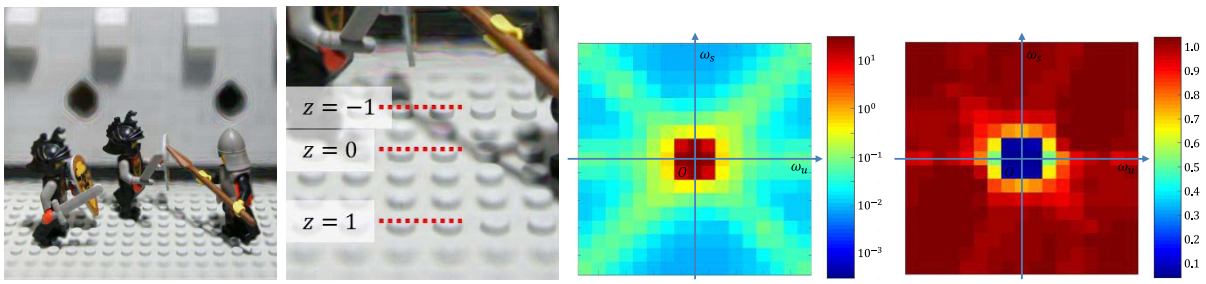
Here, we use the term *mainly* because the spectral supports are not strictly bounded in these two representations; each image in the focal stack still contains some information at out-of-focus depths, and the tensor display can also present some information at off-layer depths.

B. Analytical Experiments

To validate the discussions on *depth selectivity*, we performed analytical experiments. From a given light field L, we first generated a focal stack: $\{F_z | z \in \mathcal{Z}_F\} = \Phi_{\mathcal{Z}_F}(L)$. Then, we reconstructed a light field $\bar{L}_F$ by solving the least squares problem given as

$$\bar{L}_F = \arg\min_{L^*} \sum_{z \in \mathcal{Z}_F} \iint_{\mathcal{U}\times\mathcal{V}} |F_z(u,v) - F_z(u,v;L^*)|^2 du dv, \quad (24)$$

where $F_z(u,v;L^*)$ denotes an image that is generated from L^* using Eq. (2) and to be focused at z. More details are given in Appendix C. We finally analyzed $\bar{L}_F$ and the error $err_F = \bar{L}_F - L$, which will suggest what information is preserved in the focal stack. As for the tensor display, we followed the same procedure. From the given light field L, we obtained the layer representation $\{A_z | z \in \mathcal{Z}_T\}$ by solving Eq. (6). Then, using the layer patterns, we reconstructed the light field: $\bar{L}_T = \Psi(\{A_z | z \in \mathcal{Z}_T\})$. Finally, we analyzed $\bar{L}_T$ and the error $err_T = \bar{L}_T - L$ to see what information can be presented by the tensor display.



(a) (b) (c) (d)

Fig. 5. Analysis of tensor display with four-fold time-multiplexing. Refer to text for details. (a) Center view. (b) Close-up of (a). (c) $\bar{L}_T$ in freq. domain. (d) err_T in freq. domain.

As the data to analyze, we used a dataset, "Lego Knights" [34], which consists of 17 × 17 multi-view images, and each has 512 × 512 pixels. We used the same set of depths, $\mathcal{Z} = \{1, 0, -1\}$, for the focal stack and tensor display. Here, the depth values were normalized by the interval among the layers. The original light field L, the reconstructed light fields, $\bar{L}_F$ and $\bar{L}_T$, and the reconstruction errors, err_F and err_T, were analyzed in the frequency domain. More specifically, we took 261,120 blocks with a size of 17 × 17 in (u, s) space, applied DFT with the Hanning (raised cosine) window function, and observed the average amplitude over 261,120 samples. The error spectra for err_F and err_T were divided by the original power spectrum of L for better visualization.

The results are summarized in Fig. 4. The original light field L, the central view, and a close-up of it are shown in (a), (b), and (c), respectively. Depicted in (d) is the average amplitude of the original light field L in the frequency domain. The light fields reconstructed via the focal stack and tensor display are presented in (e)–(h) and (i)–(l), respectively. For either case, we can observe the depth selectivity in the reconstructed central image [(e) or (i)] and a close-up of it [(f) and (j)]. As indicated by the theory, the spectral information contained in the focal stack and tensor display are *depth-selective*. The spectral powers of $\bar{L}_F$ and $\bar{L}_T$ concentrated on the lines along $\omega_s = z\omega_u$ for $z \in \mathcal{Z}$, as shown in (g) and (k). The reconstruction errors err_F and err_T tended to be small along the lines $\omega_s = z\omega_u$ for $z \in \mathcal{Z}$, as shown in (h) and (l). These observations support the statement that the two representations share essentially equivalent information, and thus, direct conversion from the focal stack to the tensor display is reasonable and even ideal in terms of minimizing the data required as the input.

The analysis presented above was conducted under the assumption that no time-multiplexing was used for the tensor display. We also analyzed what information can be reconstructed with four-fold time-multiplexing in the same manner, the result of which is presented in Fig. 5. The effect of time-multiplexing is observed in the image details. Compared with the case without time-multiplexing [(i) and (j) in Fig. 4], the reconstructed image shown in Figs. 5(a) and (b) seems clearer in some parts thanks to the richer information reconstructed with time-multiplexing. The same effect can also be observed in the frequency domain. Compared with the case without time-multiplexing [(k) and (l) in Fig. 4], the spectral power of $\bar{L}_T$ covered a wider area, as shown in Fig. 5(c), and the reconstruction error was reduced, especially in the low frequency components, as shown in Fig. 5(d). These low frequency components are important because the original light field had much spectral power in these components, as shown in Fig. 4(d). This result indicates that a tensor display with time-multiplexing can represent richer information than a focal stack that is composed of only three focused images [(e) – (h) in Fig. 4], and thus, conversion from the focal stack to the tensor display is no longer feasible when time-multiplexing is enabled for the display.

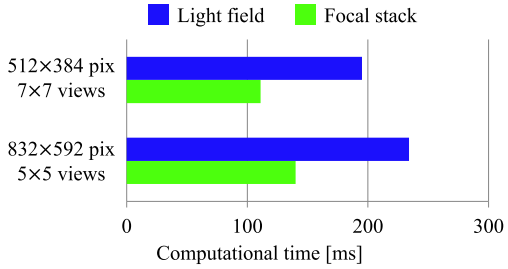


Fig. 6. Computational time for conventional (light field) and proposed (focal stack) methods.

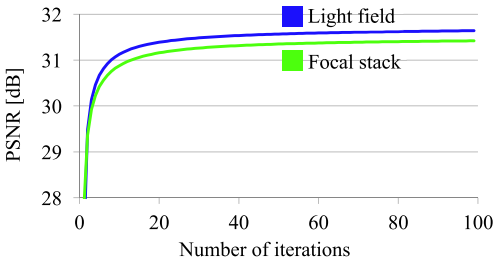
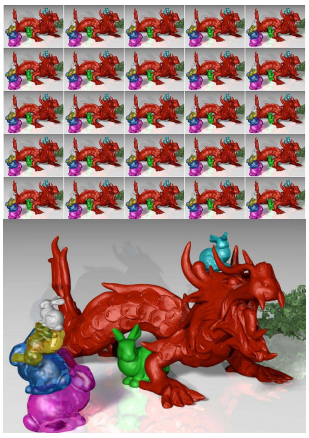


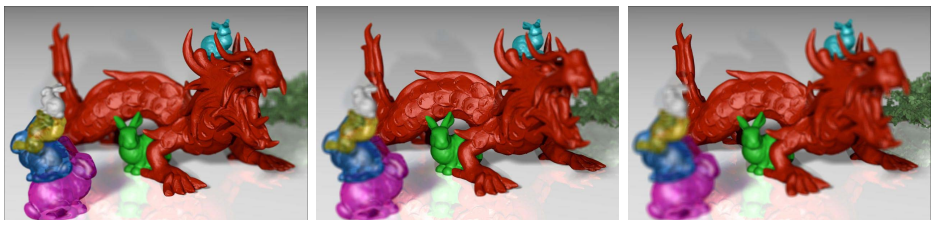
Fig. 7. Number of iterations and approximation errors obtained with conventional (light field) and proposed (focal stack) methods.

V. Experiments

It is clear that our method using a focal stack (Algorithm 2) can significantly reduce the cost of data acquisition compared with the conventional method using the original light field (Algorithm 1). In this section, we experimentally demonstrate that our method can achieve reasonable quality in reproducing the light field compared with the conventional method. We also evaluate the effect of the depth discrepancy between the



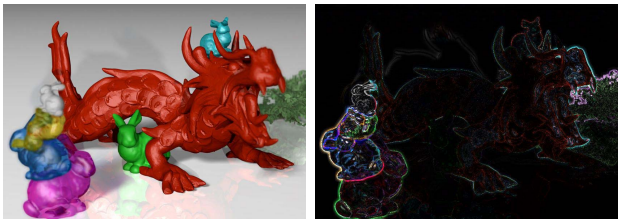
(a) Light field (5 × 5 views) and central view



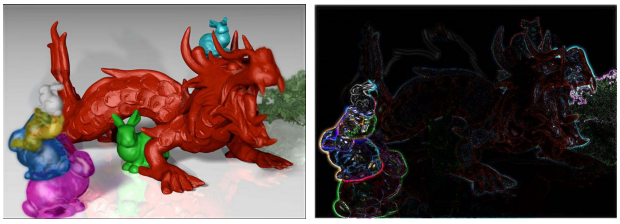
(b) Focal stack with 3 depths generated from (a)

(c) Layer patterns obtained from light field (a)

(d) Layer patterns obtained from focal stack (b)



(e) Simulated output using layer patterns (c) and its error



(f) Simulated output using layer patterns (d) and its error

Fig. 8. Overview of experiment with synthetic light field dataset. From light field dataset (a), we generated focal stack (b). Layer patterns obtained from the light field and focal stack are shown in (c) and (d), respectively. Simulated outputs, which were computed from layer patterns in (c) and (d), are visualized in (e) and (f), respectively.

focal stack and tensor display. Throughout the experiments, the number of display layers was set to 3, and their depths were set to 1, 0, and −1, which were the values normalized by the interval among the layers.

We implemented Algorithm 1 and Algorithm 2 both on a CPU and GPU. The CPU versions of the algorithms are available from our website [31]. In the GPU versions, both methods were parallelized for each layer pixel by using CUDA. Figure 6 shows the average computational times obtained with the GPU versions, where the number of iterations was fixed to 50. We used a PC that had an Intel Core i7 CPU with 16 GB of RAM and a NVIDIA GeForce GTX 1080 video card. The computational time was reduced to about 60% by our method (focal stack) thanks to the simplified calculation that uses only three images instead of the complete light field. Figure 7 shows the relationship between the number of iterations and reproduction errors (errors between the original and reconstructed light fields) that were measured with the "DragonsAndBunnies" dataset [35]. We observed that the numbers of iterations until convergence were not that different between the conventional method (light field) and our method (focal stack).

The first experiment was designed to evaluate the accuracy of light fields reproduced by the tensor display. We used several light field datasets obtained from the "Synthetic Light Field Archive" [35], the specifications of which are listed in Table I. An overview of this experiment is shown in Fig. 8. As shown in (a), the original light field consisted of 5 × 5 multi-view images. From this light field, we generated three refocused images, each of which was focused on each layer of the display, as shown in (b). We assumed that the focal stack was generated from 5 × 5 multi-view images in accordance with Eq. (2) and the discretization mentioned in Section II-D:

$$F_{\Delta_z z}(\Delta_s x, \Delta_s y) = \sum_{j=-2}^{2} \sum_{i=-2}^{2} I_{i,j}(x - zi, y - zj), \quad (25)$$

where z denotes the normalized depth. Shown in (c) and (d) are the optimized layer patterns obtained by the conventional and

TABLE I
LIGHT FIELD DATASETS

	Dataset	Views	Resolution
(a)	DragonsAndBunnies	5×5	840×593
(b)	Fishi		
(c)	Messerschmitt		
(d)	Dice	7×7	512×384
(e)	GreenDragon		
(f)	RedDragon		

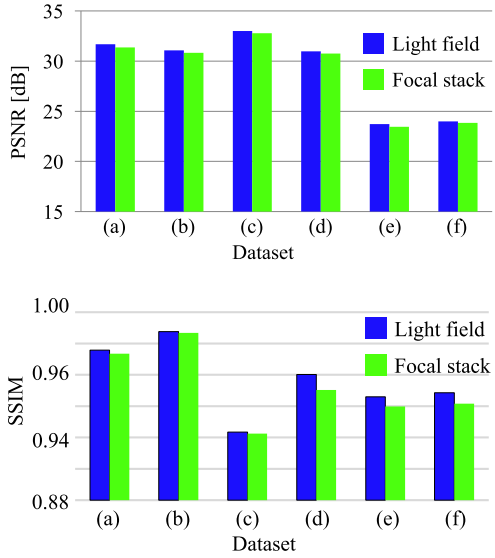


Fig. 9. Quantitative quality of output light fields over 6 datasets measured by (top) Y-PSNR and (bottom) SSIM. Proposed method (focal stack) achieved quality comparable to conventional method (light field).

proposed methods, respectively. The former was calculated directly from the original light field in (a), while the latter was from the focal stack in (b). Shown in (e) and (f) are the simulated images that could be observed when three stacked layers were seen from the central viewing direction. These images were generated computationally from the layer patterns shown in (c) and (d), respectively. Their errors from the ground truth are also presented (magnified by 5 for visualization). The errors were caused mainly around the object edges in both (e) and (f). These errors lead to halo artifacts around the edges, and these artifacts became more visible when we continuously changed the viewing direction. This problem should be addressed in future work.

As far as can be seen from Fig. 8, the conventional method [(c) and (e)] and proposed method [(d) and (f)] yielded similar results. To evaluate this quantitatively, we measured the accuracy of the reproduced light fields over six datasets and present the results in Fig. 9. Here, the number of iterations was set to 100 for both the conventional and proposed methods. The stacked layers were observed by simulation from the same viewing directions as those of the original light field dataset, and the reproduction quality was measured by Y-PSNR (peak signal-to-noise ratio) and SSIM (structural similarity) against the original light field. A Y-PSNR value was obtained from the mean squared error over all 5×5 multi-view images. This metric is suitable for seeing whether the algorithms worked

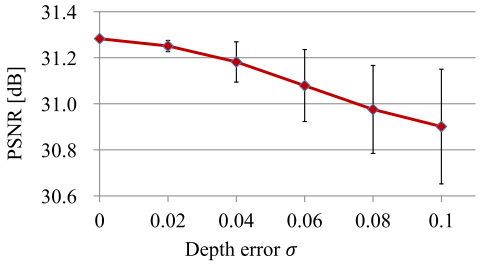


Fig. 10. Effect of depth errors in focal stack on output light field. Depth is normalized by the interval among layers.

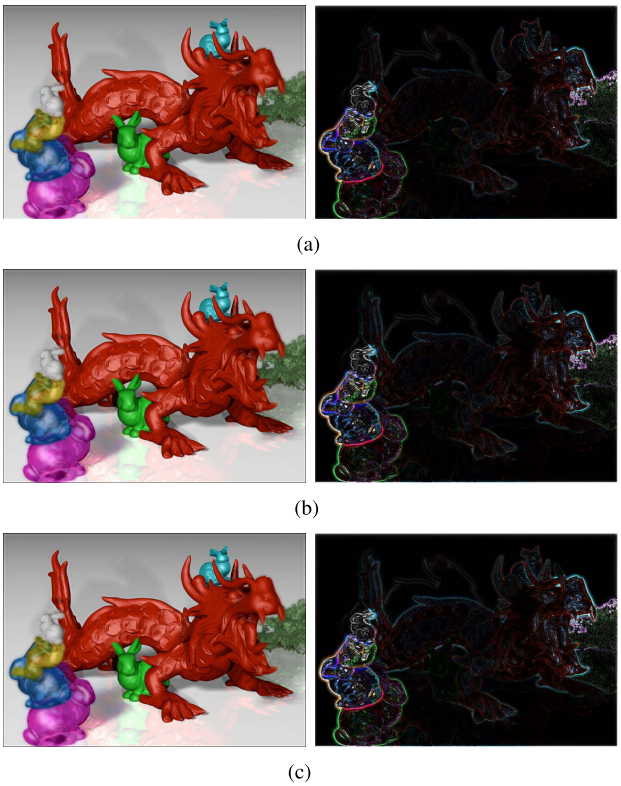


Fig. 11. Simulated output images and differences from ground truth in presence of depth errors between focal stack and layers' depths. (a) $\sigma = 0$, $\mathcal{Z} = \{1.00, 0.00, -1.00\}$, PSNR = 31.28 dB. (b) $\sigma = 0.06$, $\mathcal{Z} = \{0.97, 0.07, -1.02\}$, PSNR = 31.02 dB. (c) $\sigma = 0.1$, $\mathcal{Z} = \{0.96, 0.12, -1.03\}$, PSNR = 30.78 dB.

correctly because both methods try to minimize the squared error between the ground truth and reconstructed light fields, as shown by Eq. (6). The SSIM values reported here are the averages over all 5×5 multi-view images. SSIM is known as a perceptual metric that is closely correlated with the scores of subjective assessment.[4] For both metrics, our method (focal stack) achieved very close scores to those of the conventional method (light field), which indicates that our method (focal stack) achieved quality comparable to the conventional method (light field).

Next, we evaluated the effect of depth errors between the focal stack and tensor display. This experiment was designed

[4]Perceptual quality of light fields is more complex than that of conventional 2-D images, and several studies are being conducted on this topic [36]–[38].

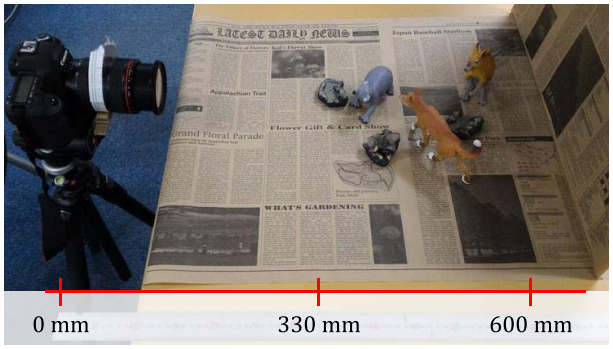


Fig. 12. Shooting set-up. Refer to text for details.

because, while the display's layers can be accurately located, the focus depth of a camera cannot be controlled accurately in practice.[5] In this experiment, the configuration of the tensor display was fixed; the layers were located at 1, 0, and −1. The focus depths of the input focal stack should ideally be set to 1, 0, and −1, but we added stochastic errors to the focus depths. More specifically, we added zero-mean Gaussian noise with the standard deviation σ to the focus depths when creating a focal stack from the original light field. We used bicubic interpolation to obtain sub-pixel values. Regardless of the noise, our method handled the focal stack as if it were focused at 1, 0, and −1. We took the average and standard deviation over 100 trials for each value of σ. As shown in Fig. 10, the reconstruction error increased as the depth error increased. However, as shown in Fig. 11, the visual quality was not significantly degraded; it seems that the human visual system is tolerant to small shape distortions caused by depth errors.

Finally, we tested our method with a real 3-D scene. This time, the input to our method was a focal stack acquired with an ordinary camera. We used a Canon EOS 5D Mark II and a zoom lens, EF24-105 F4L IS USM, and the shooting set-up is shown in Fig. 12. Three images focused at different depths are presented in Fig. 13(a).[6] The focus depths were arranged to cover the target depth range (from the koala to the kangaroo in this case) of the scene with approximately constant depth intervals. From this focal stack, we calculated the layer patterns by using our method assuming that the focal stack was generated by the same model as Eq. (25). Although we did not know the exact focus depths in the focal stack, we treated the stack as if it were captured at depths 1, 0, and −1. The resulting layer patterns are shown in Fig. 13(b). Using these patterns, we could simulate output images that would be observed from different directions. To check the subjective quality of 3-D visualization closely, please use the software published on our website [31]. The software includes a simulator of the display with which one can smoothly change

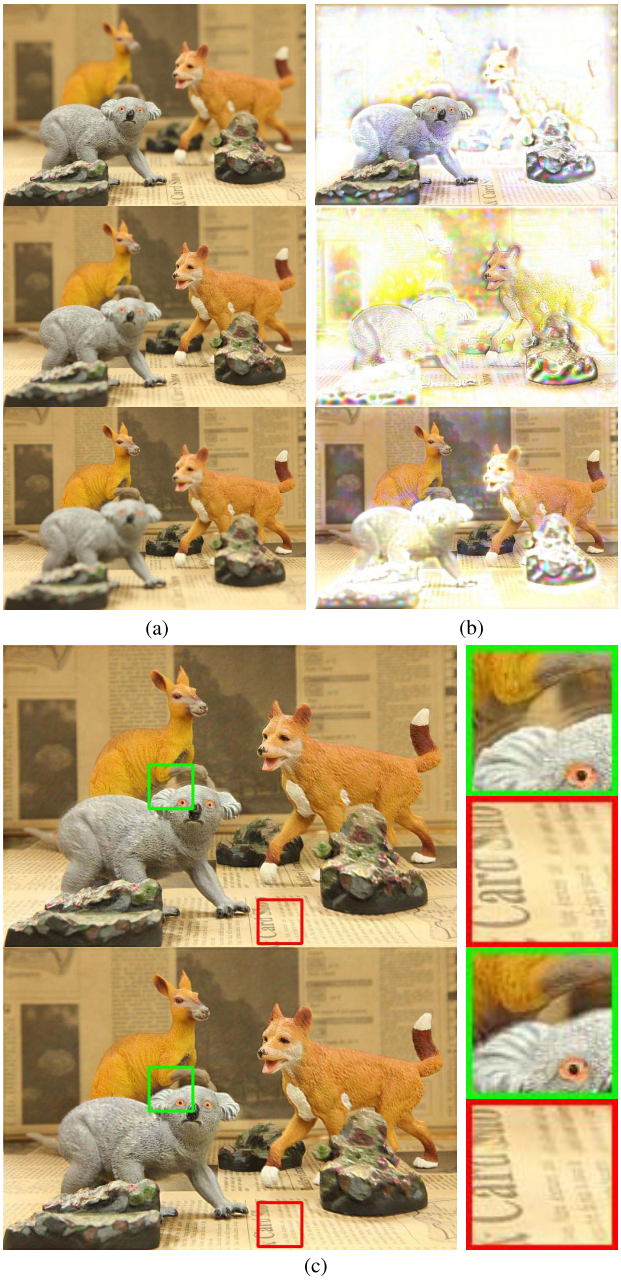


Fig. 13. Experiment with real scene. (a) Focal stack used as input, (b) three layer patterns obtained with proposed method, and (c) most top-left and bottom-right images computed from layer patterns with close-ups.

the viewing directions while observing the appearance of the display. In Fig. 13(c), we present the simulated images, which were computed from the layer patterns in (b), for the most top-left $[(i, j) = (-2, -2)]$ and bottom-right $[(i, j) = (2, 2)]$ viewing directions. Close-ups for the same areas are also presented to show the difference between the two images. Here, we analyzed the range in disparity among these images. As indicated by Eq. (25), an object located at z has a disparity of z pixels among the neighboring views because $I_{i,j}(x - zi, y - zj) = I_{0,0}(x, y)$ should be satisfied for a point (x, y) to be clearly focused in F_z. Accordingly, the range between these views should approximately be from 4 (the

[5]We usually control the focus depth of a camera by manually adjusting the focus ring or using the built-in auto-focus function, without knowing the exact focus depth.

[6]The images were originally taken at 5616 × 3744 pixels. Then, we manually modified the size of the images because, with our camera, the zoom slightly changes along with the focus depth. If the configuration were calibrated beforehand, this resizing process could be automated. All images were finally resized to 768 × 512 pixels to be fed to our algorithm.

Fig. 14. Our prototype display (top) and displayed result (bottom). Prototype was previously reported in [22] and [23]. Refer to text for details.

Fig. 15. Visual comparison of displayed results between conventional and proposed methods. See also supplemental video [31]. (a) From light field. (b) From focal stack.

koala) to -4 (the kangaroo) pixels because z ranges from 1 to -1. This range almost equaled that obtained by comparing the two images in Fig. 13(c).[7]

We also displayed these layer patterns on our prototype display [22], [23], as shown in Fig. 14. The prototype used three color LCD panels (manufactured by METASIGN Co., Ltd., 9.7 inches, 1024 $\times$ 768 pixels, 60/75 Hz) stacked at intervals of 8 mm. Therefore, the sampling intervals Δ_s and Δ_d were respectively 0.191 mm and 0.0239 (corresponding to approximately 1.4 degrees). The backlight was hand-made by using bright LED lamps because off-the-shelf backlights were insufficient in brightness. To prevent visually unappealing moiré patterns, diffuser sheets were inserted between the layers. We observed the display from a distance of 70 –100 cm, and natural 3D perception was possible despite the fact that the depth was virtually compressed in the display space. However, the visual quality obtained with this prototype was degraded compared with the results of the simulation due to several hardware-related factors such as the color filter matrix and polarizer sheets attached on each LCD panel, the diffuser sheets inserted between the layers, and the non-uniformity of the backlight. These hardware issues should be addressed in future work.

More results are presented in the supplemental videos [31], where one can clearly see that some amount of motion parallax is reproduced only from a focal stack. One of the videos includes visual comparisons of the displayed results between the conventional (using a light field as the input) and proposed (using a focal stack as the input) methods, like the ones shown in Fig. 15.

[7]The disparity values were measured by finding corresponding keypoints from the two images and comparing their coordinates.

VI. Conclusion

A method for visualizing a light field on a tensor display using a focal stack was proposed. Our method can greatly reduce the cost of data acquisition over the conventional method because only a focal stack (a few images) is necessary as the input, while a complete light field (many multi-view images) was required previously. In spite of the significant reduction in input data, our method can still reproduce high-quality light fields that are comparable to those obtained with the conventional method. We also presented frequency domain analyses to explain why a focal stack is suitable as the input for a tensor display from the perspective of *depth selectivity*. We finally demonstrated that, with our method, a natural 3-D scene captured with an ordinary camera as a focal stack can be reproduced in 3-D on our real prototype display.

As future work, we will develop an end-to-end video system where the entire process from capturing to displaying can be conducted in real-time. We are also interested in how our method can be extended to large-scale 3-D scenes other than indoor laboratory scenes, which may require a larger scale image acquisition system to be developed. Moreover, a more general framework for conversion from a focal stack to a tensor display will be necessary to handle more advanced display configurations such as those with time-multiplexing and directional backlighting [13], [16]. Specifically, our method can use only the same number of focused images as the number of layers of the tensor display, e.g., three focused images for three layers. However, if the tensor display can support time-multiplexing, we will need a method that can use more than three focused images to provide richer information for the display.

Light field display technology is still progressing. The capability of time-multiplexing and directional backlighting will be enhanced if faster display panels become available with

accurate synchronization schemes. It was recently reported in [39] that using display panels that have finer resolutions than the displayed images will significantly enhance the visual quality of a tensor display. Other researchers [40] have developed an additive layer display, where the intensity of a light ray is modulated additively by the layers. In keeping up with these developments, background technologies for light field acquisition and processing should also progress. In turn, this progress will encourage further development of better display hardware. We believe this mutual feedback will lead to truly immersive 3-D television systems in the future.

Appendix A
Multiplicative Update Rule

We first describe the update rule in a form given in the previous literature [13], [16]. For simplification, we consider only the 2-D subspace (s, u) of the original 4-D light field, and we set the number of layers to 2. Let $\mathbf{a}, \mathbf{b} \in \mathcal{R}^N$ be the two layer patterns, each of which consists of N pixels. An outgoing light ray is parameterized by the points of intersection with the two layers and is described as $X_{i,j} = a_i b_j$, or equivalently, in a matrix form as $\mathbf{X} = \mathbf{a}\mathbf{b}^T$. When a light field that should be emitted from the display is given as $\mathbf{X}$, one should solve the least squares problem given by

$$\arg\min_{\mathbf{a},\mathbf{b}} \|\mathbf{W} \circ (\mathbf{X} - \mathbf{a}\mathbf{b}^T)\|^2 \quad s.t.\ 0 \leq \mathbf{a}, \mathbf{b} \leq 1, \tag{26}$$

where $\mathbf{W}$ is the binary matrix that determines the effective angular range; usually, the angular range is defined as being symmetric, and thus, $W_{i,j}$ takes 1.0 for $|j - i| < \kappa$ and 0 otherwise. Symbol $\circ$ denotes the element-wise product. To solve this problem, $\mathbf{a}$ and $\mathbf{b}$ are randomly initialized with non-negative numbers, and the following multiplicative update rule is applied alternatively.

$$\mathbf{a} \leftarrow \mathbf{a} \circ ((\mathbf{W} \circ \mathbf{X})\mathbf{b})/((\mathbf{W} \circ \mathbf{a}\mathbf{b}^T)\mathbf{b}) \tag{27}$$

$$\mathbf{b} \leftarrow \mathbf{b} \circ ((\mathbf{W} \circ \mathbf{X})^T\mathbf{a})/((\mathbf{W} \circ \mathbf{a}\mathbf{b}^T)^T\mathbf{a}), \tag{28}$$

where $\circ$ and $/$ are element-wise operations.

Let us consider the first update equation closely. The i-th element of $\mathbf{a}$ is updated as

$$a_i \leftarrow a_i \frac{\sum_j X_{i,j} b_j}{\sum_j a_i b_j b_j} = \frac{\sum_j X_{i,j} b_j}{\sum_j b_j^2}, \tag{29}$$

where the range of j is limited to $|j - i| < \kappa$. To make the parameters consistent with the body text, we replace them with

$$a_i \rightarrow A_z(i) \tag{30}$$

$$X_{i,j} \rightarrow L(i, j - i; z) \tag{31}$$

$$b_j \rightarrow \tilde{A}_z(i, j - i), \tag{32}$$

where i denotes a discrete position on the layer and $j - i$ is a discrete angle. Substituting these into Eq. (29) and replacing $(j - i)$ with k $(|k| < \kappa)$, we finally derive

$$A_z(i) = \frac{\sum_k L(i, k; z)\tilde{A}_z(i, k)}{\sum_k |\tilde{A}_z(i, k)|^2}, \tag{33}$$

which is equivalent to the discretized version of Eq. (14). The same conclusion can be drawn for the second update equation, Eq. (28).

It is straight-forward to extend the above discussion to the full 4-D light field and to more than two layers.

Appendix B
Derivation of Eq. (23)

The Fourier transform of $L_z(u, v, s, t)$ is derived as follows.

$$\begin{aligned}
&\hat{L}_z(\omega_u, \omega_v, \omega_s, \omega_t) \\
&= \int\int\int\int_{-\infty}^{\infty} L_z(u, v, s, t) e^{-j(\omega_u u + \omega_v v + \omega_s s + \omega_t t)} du dv ds dt \\
&= \int\int\int\int_{-\infty}^{\infty} o_z(u + zs, v + zt) \\
&\quad \times e^{-j(\omega_u u + \omega_v v + \omega_s s + \omega_t t)} du dv ds dt \\
&= \hat{o}_z(\omega_u, \omega_v) \iint_{-\infty}^{\infty} e^{-j((\omega_s - z\omega_u)s + (\omega_t - z\omega_v)t)} ds dt \\
&= \hat{o}_z(\omega_u, \omega_v)\delta(\omega_s - z\omega_u, \omega_t - z\omega_v)
\end{aligned} \tag{34}$$

VII. Reconstructing a Light Field from a Focal Stack

In the discrete space mentioned in Section II-D, Eq. (24) can be rewritten in a matrix-vector form as

$$\arg\min_{\mathbf{l}} \|\mathbf{f} - \mathbf{A}\mathbf{l}\|^2, \tag{35}$$

where all the pixels in the focal stack and all the elements of the light field are reshaped into the column vectors $\mathbf{f} \in \mathcal{R}^M$ and $\mathbf{l} \in \mathcal{R}^N$, respectively. Each element in $\mathbf{f}$ can be represented by linearly combining several elements in $\mathbf{l}$. This relation is described by the matrix $\mathbf{A} \in \mathcal{R}^{M \times N}$. More specifically, $\mathbf{A}$ is determined to meet the condition that $\mathbf{f} = \mathbf{A}\mathbf{l}$ is equivalent to

$$f_z(x, y) = \sum_{i,j} I_{i,j}(x - zi, y - zj), \tag{36}$$

where $f_z(x, y)$ is an image focused at the normalized depth z, and $I_{i,j}(x, y)$ denotes an image for a discrete viewing direction (i, j). Since $\mathbf{A}$ is a huge matrix, the matrix inversion involved in the closed form solution of Eq. (35) would be computationally impractical. Therefore, Eq. (35) was minimized by using a gradient decent method, in which, with $\mathbf{l}$ initialized as $\mathbf{l}^{(0)}$, $\mathbf{l}$ is updated as

$$\mathbf{l}^{(t+1)} \leftarrow \mathbf{l}^{(t)} + \alpha \mathbf{A}^T(\mathbf{f} - \mathbf{A}\mathbf{l}^{(t)}) \tag{37}$$

$$\alpha = \frac{\|\mathbf{A}^T(\mathbf{f} - \mathbf{A}\mathbf{l}^{(t)})\|^2}{\|\mathbf{A}\mathbf{A}^T(\mathbf{f} - \mathbf{A}\mathbf{l}^{(t)})\|^2}, \tag{38}$$

until it converges. Moreover, to avoid keeping the huge matrices explicitly on memory, we implemented equivalent operations in the multiplications with $\mathbf{A}$ and $\mathbf{A}^T$. The multiplication with $\mathbf{A}$ is expressed by Eq. (36). Similarly, $\mathbf{l} = \mathbf{A}^T\mathbf{f}$ is equivalent to

$$I_{i,j}(x, y) = \sum_z f_z(x + zi, y + zj). \tag{39}$$

This can be confirmed by finding all the elements in $f_z(x, y)$ to which a pixel in $I_{i,j}(x, y)$ has a contribution in Eq. (36).

Both Eqs. (36) and (39) were implemented as simple image processing operations. Obtaining a feasible solution with this simple iteration was sufficient for the purpose of our analytical experiment.

REFERENCES

[1] F. E. Ives, "Parallax stereogram and process of making same," U.S. Patent 725 567 A, Apr. 14, 1903.

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

[3] T. Okoshi, *Three-Dimensional Imaging Techniques*. New York, NY, USA: Academic, 1976.

[4] S. Pastoor and M. Wöpking, "3-D displays: A review of current technologies," *Displays*, vol. 17, no. 2, pp. 100–110, 1997.

[5] B. Javidi and F. Okano, Eds., *Three-Dimensional Television, Video, and Display Technologies*. Berlin, Germany: Springer, 2002.

[6] H. Isono, M. Yasuda, and H. Sasazawa, "Autostereoscopic 3-D display using LCD-generated parallax barrier," in *Proc. Jpn. Display*, 1992, pp. 303–306.

[7] K. Sakamoto and T. Morii, "Multiview 3D display using parallax barrier combined with polarizer," *Proc. SPIE*, vol. 6399, 2006, doi: 10.1117/12.688879.

[8] T. Peterka, R. L. Kooima, D. J. Sandin, A. Johnson, J. Leigh, and T. A. DeFanti, "Advances in the dynallax solid-state dynamic parallax barrier autostereoscopic visualization display system," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 3, pp. 487–499, May/Jun. 2008.

[9] R. Börner, "Autostereoscopic 3D-imaging by front and rear projection and on flat panel displays," *Displays*, vol. 14, no. 1, pp. 39–46, 1993.

[10] M. McCormick, "Integral 3D imaging for broadcast," in *Proc. Int. Display Workshop*, vol. 3, 1995, pp. 77–80.

[11] J. Arai *et al.*, "Integral three-dimensional television using a 33-megapixel imaging system," *J. Display Technol.*, vol. 6, no. 10, pp. 422–430, 2010.

[12] S. Suyama, H. Takada, and S. Ohtsuka, "A direct-vision 3-D display using a new depth-fusing perceptual phenomenon in 2-D displays with different depths," *IEICE Trans. Electron.*, vol. E85-C, no. 11, pp. 1911–1915, 2002.

[13] D. Lanman, M. Hirsch, Y. Kim, and R. Raskar, "Content-adaptive parallax barriers: Optimizing dual-layer 3D displays using low-rank light field factorization," *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. no. 163.

[14] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar, "Polarization fields: Dynamic light field display using multi-layer LCDs," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. no. 186.

[15] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar, "Beyond parallax barriers: Applying formal optimization methods to multilayer automultiscopic displays," *Proc. SPIE*, vol. 8288, 2012, doi: 10.1117/12.907146.

[16] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting," *ACM Trans. Graph.*, vol. 31, no. 4, 2012, Art. no. 80.

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

[18] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. ACM SIGGRAPH*, 1996, pp. 43–54.

[19] M. Hirsch, G. Wetzstein, and R. Raskar, "A compressive light field projection system," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 58:1–58:12, Jul. 2014.

[20] R. Konrad, N. Padmanaban, K. Molner, E. A. Cooper, and G. Wetzstein, "Accommodation-invariant computational near-eye displays," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 88:1–88:12, 2017.

[21] T. Saito, Y. Kobayashi, K. Takahashi, and T. Fujii, "Displaying real-world light fields with stacked multiplicative layers: Requirement and data conversion for input multiview images," *OSA/IEEE J. Display Technol.*, vol. 12, no. 11, pp. 1290–1300, Nov. 2016.

[22] K. Takahashi, Y. Kobayashi, and T. Fujii, "Displaying real world light fields using stacked LCDs," in *Proc. Int. Display Workshops*, 2016, pp. 1323–1326.

[23] 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 Trans. Media Technol. Appl.*, vol. 5, no. 3, pp. 88–95, 2017.

[24] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2000, pp. 297–306.

[25] R. Ng, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 735–744, 2005.

[26] K. N. Kutulakos and S. W. Hasinoff, "Focal stack photography: High-performance photography with a conventional camera," in *Proc. IAPR Conf. Mach. Vis. Appl.*, 2009, pp. 332–337.

[27] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1831–1838.

[28] K. Kodama and A. Kubota, "Efficient reconstruction of all-in-focus images through shifted pinholes from multi-focus images for dense light field synthesis and rendering," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4407–4421, Nov. 2013.

[29] X. Lin, J. Suo, G. Wetzstein, Q. Dai, and R. Raskar, "Coded focal stack photography," in *Proc. IEEE Int. Conf. Comput. Photograph. (ICCP)*, Apr. 2013, pp. 1–9.

[30] Y. Kobayashi, K. Takahashi, and T. Fujii, "From focal stacks to tensor display: A method for light field visualization without multi-view images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2007–2011.

[31] K. Takahashi. (2018). *Light Field Display Project*. [Online]. Available: http://www.fujii.nuee.nagoya-u.ac.jp/~takahasi/Research/LFDisplay/

[32] S. Kondo, Y. Kobayashi, K. Takahashi, and T. Fujii, "Physically-correct light-field factorization for perspective images," *IEICE Trans. Inf. Syst.*, vol. E100-D, no. 9, pp. 2052–2055, 2017.

[33] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proc. ACM SIGGRAPH*, 2000, pp. 307–318.

[34] V. Vaish. (2018). *The (New) Stanford Light Field Archive*. [Online]. Available: http://lightfield.stanford.edu/lfs.html

[35] G. Wetzstein. (2018). *Synthetic Light Field Archive*. [Online]. Available: https://www.media.mit.edu/~gordonw/SyntheticLightFields/

[36] I. Viola, M. Řeřábek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *Proc. Picture Coding Symp. (PCS)*, Dec. 2016, pp. 1–5.

[37] P. Paudyal, J. Gutiérrez, P. Le Callet, M. Carli, and F. Battisti, "Characterization and selection of light field content for perceptual assessment," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May/Jun. 2017, pp. 1–6.

[38] P. Paudyal, F. Battisti, M. Sjöström, R. Olsson, and M. Carli, "Towards the perceptual quality evaluation of compressed light field images," *IEEE Trans. Broadcast.*, vol. 63, no. 3, pp. 507–522, Sep. 2017.

[39] Y. Kobayashi, K. Takahashi, and T. Fujii, "Using higher resolution and lower bit-depth panels for stacked-layer light-field display," in *Proc. 24th Int. Display Workshops*, Dec. 2017, pp. 901–904.

[40] S. Lee, C. Jang, S. Moon, J. Cho, and B. Lee, "Additive light field displays: Realization of augmented reality with holographic optical elements," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 60:1–60:13, Jul. 2016.

Keita Takahashi received the B.E., M.S., and Ph.D. degrees in information and communication engineering from The University of Tokyo in 2001, 2003, and 2006, respectively. He was a Project Assistant Professor with The University of Tokyo from 2006 to 2011 and an Assistant Professor with the University of Electro-Communications from 2011 to 2013. He is currently an Associate Professor with the Graduate School of Engineering, Nagoya University, Japan. His research interests include computational photography, image-based rendering, and 3D displays.

Yuto Kobayashi received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Japan, in 2016 and 2018, respectively. His research topics were light-field acquisition and rendering for 3D displays when he was a Student.

Toshiaki Fujii received the B.E., M.E., and Dr. E. degrees in electrical engineering from The University of Tokyo in 1990, 1992, and 1995, respectively. Since 1995, he has been with the Graduate School of Engineering, Nagoya University. From 2008 to 2010, he was with the Graduate School of Science and Engineering, Tokyo Institute of Technology. He is currently a Professor with the Graduate School of Engineering, Nagoya University. His current research interests include multi-dimensional signal processing, multi-camera systems, multi-view video coding and transmission, free-viewpoint television, and their applications.