A 3-D Display Pipeline: Capture, Factorize, and Display the Light Field of a Real 3-D Scene

Yuto Kobayashi†, Shu Kondo†, Keita Takahashi (member)†, Toshiaki Fujii (member)†

Abstract  Inspired by pioneering work on modern light field displays, we developed a prototype display in which three liquid crystal display (LCD) panels are stacked in front of a backlight. The stacked LCD panels constitute a set of semi-transparent layers, which modulate the outgoing light rays. Even though only three layers are used, this display can emit a light field, i.e., a dense set of many multi-view images, in different viewing directions simultaneously. We established a process pipeline from capture to display of the light field of a real 3-D scene. We analyzed the amount of pop-out and motion parallax that can be presented by the display using a given light field data. We used a light field camera (Lytro Illum) and a multi-view camera (ViewPLUS ProFUSION 25) to capture the light field, which was then factorized into layer representations to be displayed. We show several successful results using our prototype display.

Key words: light field, 3-D display, image based rendering, non-negative tensor factorization, light field camera

1. Introduction

Light field displays are display devices that aim to reproduce a light field, i.e., a set of light rays that is emitted from arbitrary points on the display surface in arbitrary directions. These displays have attracted much attention from researchers because of their ability to provide not only a binocular depth sensation but also natural motion parallax in relation with head motion. Technically, a light field is represented as a dense set of many multi-view images, and thus, a light field display should be capable of emitting many views simultaneously. To develop such a display, several approaches using parallax barriers\(^1\), lenticular lenses\(^2\), and rear projections\(^3\)\(^4\) have been proposed. Among them, we focus on a newly emerging approach using a stack of semi-transparent layers\(^5\)\(^6\)\(^7\). This approach enables us to display many views simultaneously without sacrificing the resolution for each view. As proposed by Wetzstein et al.\(^7\), an efficient layer representation is made possible by light field factorization, where the layers’ transmittances are iteratively optimized so as to reproduce the given light field data as accurately as possible. This design can also be applied to light field projection\(^8\)\(^9\) and head mounted displays\(^10\)\(^11\). However, most of the contents tested with those layered representations have been CG contents; this limitation is seemingly due to the difficulty of capturing input light field data from a real 3-D scene.

In this paper, we followed the design of a “tensor display”\(^2\) and developed prototype hardware using three liquid crystal display (LCD) panels and a backlight. Moreover, we established a process pipeline from capture to display of a real 3-D scene. This is the main contribution of our study, and it is a non-trivial task, because the configuration of the input light field data has significant impact on the quality and 3-D sensation that are perceived by the observer. An important fact is that the required density (the viewpoint interval) for the input light field data is a rather strict condition. For example, a ViewPLUS ProFUSION 25 camera has 25 \((5 \times 5)\) viewpoints with 12 mm viewpoint intervals. However, with a practical setup of the target scene, the density of captured data is too low (the viewpoint interval is too large) for a high-quality display. To resolve this problem, we used image-based rendering to produce sufficiently dense multi-view images (virtual images) from real photographs\(^12\)\(^13\). Meanwhile, if we use a Lytro Illum camera\(^14\) to capture the target scene, the requirements on the data density can easily be satisfied. However, a small disparity range between the outermost viewpoints, which is due to the limited aperture of the Lytro Illum camera, limits the 3-D sensation that can be obtained from the display. To provide a sufficient
depth sensation, the target scene should be configured to have a larger depth range compared with the case of using a ProFUSION 25 camera. In both cases, we got successful results with our prototype display.

This paper is an extension of our previous papers. In those articles\(^{12,13}\), we analyzed the requirements placed on light field data for creating high-quality 3-D displays and proposed using image-based rendering to synthesize sufficiently dense multi-view images from captured real images. However, only computer simulation results were presented, because our prototype display was still under development. Later, in a workshop\(^5\), we reported on our prototype display, which is capable of displaying real 3-D objects using data captured by a multi-view camera. In the present paper, we use both a multi-view camera and a light field camera as input devices and provide a comprehensive discussion on the process pipeline from capture to display of a real 3-D scene.

2. Principle and Parameters

2.1 Displaying A Light Field through Factorization

Here, we summarize the principle of the layered light field display, dubbed a tensor display\(^7\).

A few light attenuating layers, such as liquid crystal display (LCD) panels, are stacked in front of a backlight as shown in Fig. 3. With this setup, different views are emitted in different directions, as a result of the difference in the relative shifts among the layers depending on the direction. With three layers, an outgoing light ray (which constitutes the light field produced by the display) illustrated in Fig. 1 can be described as

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

Here, \(T_n(x,y)\) 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, which can be omitted for simplicity. As can be seen from Fig. 1, a light ray with a tuple of \((s,t,x,y)\) intersects the middle layer at \((x,y)\) and goes in the direction denoted by \((s,t)\).

To display 3-D content on this display, we first need to prepare light field data consisting of a set of multi-view images. These images correspond to the views that are to be observed from different directions. More specifically, each image is associated with the target light field as follows:

\[
\hat{L}(ki, kj, x, y) = I_{i,j}(x,y), \quad (2)
\]

where \(I_{i,j}(x,y)\) is the image observed from viewpoint \((i,j)\). The symbol \(k\) denotes a positive integer used to control the amount of pop-out; a larger \(k\) yields less pop-out for the same set of multi-view images, because it leads to a larger direction difference for the same viewpoint change. At the same time, a larger \(k\) means a lower sample density, which causes aliasing, as will be described later.

Next, the transmittance patterns for the layers are optimized so as to reproduce the given data (target light field) as accurately as possible: \(L(s,t,x,y) \simeq \hat{L}(ki,kj,x,y)\) with \((s,t) = (ki,kj)\). Equivalently, the optimization is

\[
\arg \min_{T_n} \sum_{i,j,x,y} \left| I_{i,j}(x,y) - \prod_{n=-1}^{1} T_n(x + nki, y + nkj) \right|^2 \quad (3)
\]

where \(0 \leq T_n(x,y) \leq 1\). This optimization is formulated as non-negative tensor factorization (NTF). Although the number of layers is limited to a small value (3 in our case), they can reproduce the target light field consisting of many views (typically dozens to hundreds of views) with high quality. We refer the reader to the original paper\(^7\) for descriptions of the optimization method and the extension to time multiplexing.

2.2 Parameters of A Light Field

In this section, we analyze two factors, i.e., the pop-out range and the total amount of motion parallax of the displayed content, which should be considered when we arrange the target 3-D scene.

Let us consider the configuration illustrated in Fig. 2,
which describes the layout of the input cameras with which the input light field is captured. If an object is located at \( Z \), the disparity among the neighboring viewpoints (in pixels) is given by

\[
d = f \left( \frac{1}{Z} - \frac{1}{Z_0} \right)
\] (4)

where \( f \) is the focal length of each camera, \( b \) is the baseline connecting the nearest cameras, and \( Z_0 \) is the depth of the converging plane. These images are associated with the target light field in accordance with Eq. (2).

An important factor for every 3-D display is the pop-out range of the displayed content, which is directly related to the 3-D sensation perceived by the observers. A larger amount of pop-out induces a stronger 3-D sensation. However, an object with a larger amount of pop-out is reproduced with stronger blurring, because the resolution (the upper bound of the spatial frequency) of the display is depth dependent\(^7\)\(^13\). According to the previous literature\(^7\)\(^13\), it is safe to keep the pop-out range within the display volume: the volume between the front and rear layers. As shown in the appendix, an object point that has a disparity value \( d \) among the nearest input images is displayed at \( z = d \Delta L \) in the display space, where \( \Delta L \) is the interval between layers. Therefore, keeping the content within the volume is equivalent to \( |d| \leq 1 \), or equivalently,

\[
f b \left| \frac{1}{Z} - \frac{1}{Z_0} \right| \leq 1.
\] (5)

As long as the above condition is satisfied, a larger amount of pop-out is desirable for obtaining a stronger 3-D sensation. Here, the range of pop-out \( \Delta z \) is defined using the minimum and maximum disparities in the target scene, as

\[
\Delta z := \Delta L \cdot (d_{\text{max}} - d_{\text{min}})
\]

\[
= \Delta L \cdot f b \left( \frac{1}{Z_{\text{min}}} - \frac{1}{Z_{\text{max}}} \right)
\] (6)

where \( Z_{\text{min}} \) and \( Z_{\text{max}} \) are the minimum and maximum depths of the target scene. The right-hand side of Eq. (6) is rewritten as

\[
\Delta L \cdot f b \left( \left( \frac{1}{Z_{\text{min}}} - \frac{1}{Z_{\text{max}}} \right) + \left( \frac{1}{Z_0} - \frac{1}{Z_{\text{max}}} \right) \right).
\] (7)

Obviously, from Eq. (5), \( \Delta z \leq 2\Delta L \).

Another important factor for 3-D displays is the total amount of motion parallax presented by the display. This can be measured by the amount of disparity between the outermost viewpoints; that is, how much difference is observed between the nearest and farthest objects if the observer moves from the leftmost to rightmost viewpoints. It can quantitatively be described as \( \Delta D = (N+1) \cdot (d_{\text{max}} - d_{\text{min}}) \), where \( N+1 \) is the number of viewpoints along the horizontal (or vertical) direction. Using Eq. (4), \( \Delta D \) is written as

\[
\Delta D = f B \left( \frac{1}{Z_{\text{min}}} - \frac{1}{Z_{\text{max}}} \right)
\] (8)

where \( B = Nb \) is the total baseline.

3. Hardware and Process Pipeline

We have developed an end-to-end system wherein a real 3-D scene captured by a multi-view camera or light field camera is reproduced in 3-D on a layered light field display. We first describe two input devices used to capture light fields, and then present two process pipelines for these input devices.

3.1 Input Devices

In this study, we used two different cameras as input devices: a multi-view camera and a light field camera. As a multi-view camera device, we chose a ViewPLUS ProFUSION 25, which has 25 cameras that are compactly arranged in a 5 × 5 array at 12 mm intervals and are gen-locked with each other. As a light field camera, we used a Lytro Illum camera, which can take a set of light field data with a single shot using a lenslet array. As summarized in Table 1, these two cameras are quite different in their configurations. Therefore,
the process pipeline and configuration of the target 3-D scene should be different depending on which camera is used, as will be detailed in sections 3.2 and 3.3.

As for the ProFUSION 25, the total baseline $B$ is about 48 mm, which is long enough to obtain sufficient motion parallax $\Delta D$ according to Eq. (8). However, the baseline connecting the neighboring viewpoints, $b$, which is 12 mm, is too long to satisfy Eq. (5) with a practical setup. Therefore, we use image-based rendering to synthetically generate a sufficiently dense set of multi-view images, in a similar manner to our previous study\cite{12,13}. In doing so, we can satisfy Eq. (5) while keeping the motion parallax $\Delta D$ unchanged.

Meanwhile, as for the Lytro Illum camera, the situation is opposite. This camera can capture very dense light fields; as shown in Table 1, $fb$ is so small that Eq. (5) can be easily satisfied with a practical setup. However, the total baseline $B$ is short, due to the limited aperture of the camera hardware. The baseline connecting neighboring viewpoints, $b$, which is determined by the hardware configuration of the Lytro Illum camera, is also very short. To gain sufficiently large pop-out $\Delta z$ and motion parallax $\Delta D$, we should make $(1/Z_{\text{min}} - 1/Z_{\text{max}})$ as large as possible; that is, the target scene should have a wider depth range compared to the case of using the ProFUSION 25 camera.

### Table 1

| Design parameters for two input devices. | ProFUSION 25 | Lytro Illum |
|-----------------------------------------|-------------|-------------|
| $fb$ [pix mm]                           | $1.1 \times 10^4$ | $7.9 \times 10^2$ |
| $B$ [pix mm]                            | $4.2 \times 10^4$ | $6.3 \times 10^3$ |

### Table 2

| Specifications of our system |
|------------------------------|
| Camera | ViewPLUS ProFUSION 25 (5x5 viewpoints, 640x480 pixels, 25fps) |
| Camera PC | Intel Core i-7 3770 CPU NVIDIA Geforce GTX 650 GPU 8GB main memory |
| Display PC | Intel Core i-7 860 CPU NVIDIA Geforce GT 730 GPU GF-QUAD-DISP/4DVI/LP video card 8GB main memory |
| LCD controller | METASIGN MF-403AD |
| LCD panels | METASIGN 9.7 inch (1024x768 pixels) |

### 3.2 Pipeline with a Multi-View Camera

The overview and specifications of our system are presented in Fig. 3 and Table 2, respectively. Hereafter, we describe our system and the process pipeline from input to output.

The input device, ProFUSION 25, is connected to a PC, the first one, in which a set of multi-view images (Fig. 3(a)) is captured and converted into a denser set of multi-view images (Fig. 3(b)). For example, we interpolate (or densified) the original $5 \times 5$ viewpoints into $17 \times 17$ viewpoints. This conversion is conducted with our in-house image-based rendering software, which implements a variant of the algorithm presented in our previous paper\cite{17}. This first PC is connected via Ethernet to another PC, the second one, which is responsible for displaying the images.

The second PC receives the target light field from the first PC and then calculates the corresponding layer representation (Fig. 3(c)) using non-negative tensor factorization (NTF). The second PC is equipped with a video card that has four synchronized DVI outputs,
from which the video signals are fed to the light field display via dedicated controller circuits.

Our light field display consists of three semi-transparent LCD panels, each of which is recognized as an external display by the second PC, and a hand crafted backlight, which is brighter than off-the-shelf display backlights. When stacking the LCD panels, we should be aware of the directions of the polarizers, which are attached to both sides of the LCD panel in perpendicular directions. Furthermore, diffuser panels are placed between the LCD panels to eliminate annoying moiré effects. Figure 5 shows the hardware components and appearance of the display.

### 3.3 Pipeline with a Light Field Camera

As illustrated in Fig. 4, a light field camera\textsuperscript{15} can be used instead of a multi-view camera. Here, we used a Lytro Illum camera. Note that different from the case with the ProFUSION 25 camera, we should capture the input light field offline, because the Lytro Illum camera does not support live video readout from a computer. A RAW image taken by the Lytro Illum is shown in 4(a), whose size is 7728 × 5368 pixels. The RAW image is converted into a set of multi-view images, 15 × 15 views, each having 625 × 433 pixels. For this conversion, we used Light Field Toolbox v4.0\textsuperscript{18}. Only the 9 × 9 views in the central part, shown in Fig. 4(b), are finally used because the other views are affected by severe lens distortion, aberration, and vignetting. As mentioned in section 3.1, the obtained light field data is already sufficiently dense, so that interpolation (densification) of the viewpoints is unnecessary. From these multi-view images, the corresponding layer representation (Fig. 4(c)) is calculated using non-negative tensor factorization (NTF).
4. Experimental Results

We displayed several real 3-D scenes on our light field display to examine the effectiveness of our pipeline.

First, we present results obtained with the ProFUSION 25 camera. The shooting conditions are shown in Fig. 6. Several images displayed on our prototype display are presented in Fig. 7. For the top image, we used the 25 images captured by the multi-view cameras directly to obtain a layer representation. We set $k = 4$ to limit the range of the displayed content inside the display volume, because the upper bound of the spatial frequency falls off outside the display volume, as has been analyzed in previous literature\(^7\)\(^{12}\). In this case, the displayed images are very noisy; the insufficient sample density leads to aliasing artifacts. For the central image in Fig. 7, we set $k = 1$ with the same input images. In this case, no aliasing occurs, but the displayed images are blurry except the dog image because of the significant amount of pop-out outside the display volume. Meanwhile, for the bottom image in Fig. 7, we first generated a denser set of multi-view images (289 images) and used them to obtain the layer representation. We can set $k = 1$ while keeping the pop-out range of the displayed content inside the display volume. In this case, we can see fine images that are clear throughout the depth. Therefore, we conclude that the sample density of the target light field is an essential consideration for this light field display, and a sufficient density can be achieved by image-based rendering. See also our supplementary video\(^{19}\).

Second, we show results obtained with the Lytro Illum camera. Two different shooting conditions are shown in Fig. 8, where the target scenes were configured in different depth ranges. Two displayed images, obtained from the first shooting condition with the wider depth range, are shown in Fig. 9. For both conditions, we extracted $9 \times 9$ images from RAW images and associated them directly with the target light field with $k = 1$. Since the obtained light fields were already dense enough to satisfy Eq. (5), we did not need to conduct image-based rendering to interpolate (or densify) the viewpoints. However, it remains challenging to obtain reasonably large amounts of pop-out and motion parallax. In the second shooting condition, we cannot observe sufficient disparity (see also the supplementary video\(^{19}\)). These results clearly show that to obtain a reasonable amount of 3-D sensation with a Lytro Illum camera, we should configure the target scene to have a very wide depth range.
Finally, we briefly mention the processing time of our current implementation. Our system with the ProFUSION 25 camera can process the data online without any human intervention. We implemented two modes: a wide viewing-range mode and a nearly live mode. The former mode increases the viewpoints such that the captured $5 \times 5$ images are converted into $17 \times 17$ images while keeping the outermost viewpoints fixed to preserve the total amount of motion parallax. This mode provides a wide viewing range, but requires significant computational costs for both image-based rendering and non-negative tensor factorization. Meanwhile, the latter nearly live mode generates denser but only $5 \times 5$ multi-view images from the captured $5 \times 5$ images. This mode provides a narrower viewing range, but its processing speed is closer to the interactive video rate. The processing time for each step and the total computational time with additional latency (e.g., data transfer) are summarized in Table 3. Due to the many images ($17 \times 17$ images) to be handled, the wide viewing-range mode is much slower than the nearly live mode. The step used to calculate a layer representation (non-negative tensor factorization) is implemented on a GPU. However, in the case of the wide viewing-range mode, this step cannot be executed on a GPU because of the limited video memory.

5. Conclusion

We presented a 3-D display pipeline to capture, factorize, and display the light field of a real 3-D scene. We developed a prototype display hardware consisting of three semi-transparent LCD panels and a backlight and visualized real 3-D scenes on it. To capture real 3-D scenes, we used a multi-view camera and a light field camera. When a multi-view camera, which has relatively long baselines, is used, the key to achieve high-quality 3-D visualizations is to generate sufficiently dense light field data from sparser samples obtained from the camera by using image-based rendering. Meanwhile, when a light field camera with a small aperture is used, the target scene should be configured to have a sufficient range of depth for the display to provide a reasonable 3-D sensation. In the current situation, the multi-view camera (ProFUSION 25) is
the only practical choice for us to capture moving 3-D objects because a Lytro Illum camera cannot capture videos. However, this situation will soon be changed by the rapid progress of lenslet-based light field cameras. Our future work will include speeding up of the process pipeline to achieve video-rate online visualization of real 3-D scenes.

Acknowledgement: This work was supported by JSPS Kakenhi under Grant 15H05314.

Appendix

Consider an object point that is observed at \((x_0, y_0)\) in the central input image \(I_{0,0}\). If the object point has a disparity \(d\) among the neighboring viewpoints in the input images, it will appear at \((x_0 + id, y_0 + jd)\) in image \(I_{ij}\). When these images are associated with the target light field in accordance with Eq. (2) with \(k = 1\), the object point is associated with the bundle of light rays \(L(i, j, x_0 + id, y_0 + jd)\). It can be easily proven that all of these light rays pass through the same 3-D point \((x, y, z) = (x_0, y_0, d\Delta L)\) in the display space, where the \(xy\) plane is in the middle layer, and \(\Delta L\) is the distance between the layers. Therefore, the image of the object point is formed at \((x_0, y_0, d\Delta L)\) in the display space.

References

1) F. E. Ives: “Parallax stereogram and process of making same,” U.S. Patent US72556A, 1903.
2) G. Lippmann: “Epreuves reversibles donnant la sensation du relief,” Journal of Physics, vol. 7, no. 4, pp. 821–825, 1908.
3) W. Matusik and H. Pfister, “3D TV: A Scalable System for Real-Time Acquisition, Transmission and Autostereoscopic Display of Dynamic Scenes,” ACM Transactions on Graphics, vol. 23, issue 3, pp. 814–824, 2004.
4) Y. Takaki and N. Nago: “Multi-projection of lenticular displays to construct a 256-view super multi-view display,” Optics express, vol. 18, no. 9, pp. 8824–8835, 2010.
5) 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, no. 95, 2011.
6) D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, R. Raskar: “Polarization Fields: Dynamic Light Field Display using Multi-Layer LCDs,” SIGGRAPH Asia 2011 papers, no. 186, 2011.
7) G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar: “Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting,” SIGGRAPH 2012 papers, no. 80, 2012.
8) M. Hirsch, G. Wetzstein, R. Raskar: “A Compressive Light Field Projection System,” ACM Proc. of SIGGRAPH (Transactions on Graphics 33, 4), 2014.
9) L. Seungjae et al. “Additive light field displays: realization of augmented reality with holographic optical elements,” SIGGRAPH 2016, vol. 35, issue 4, pp. 1–10.
10) J. Heide et al., “Cascaded Displays: Spatiotemporal Superresolution using Offset Pixel Layers,” ACM Transactions on Graphics (TOG) 33.4 (2014): 60.
11) F. Huang, K. Chen, G. Wetzstein, “The Light Field Stereoscope: Immersive Computer Graphics via Factored Near-Eye Light Field Displays with Focus Cues”, ACM SIGGRAPH (Transactions on Graphics 33, 5), 2015.
12) T. Saito, K. Takahashi, M.P. Tehrani, T. Fujii: “Data Conversion from Multi-View Cameras to Layered Light Field Display for Aliasing-Free 3D Visualization,” IST&/SPIE Electronic Imaging, Stereooscopic Displays and Applications XXVI, Paper 9391–42, 2015.
13) 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,” OSA/IEEE Journal of Display Technology, vol. 12, no. 11, pp. 1290–1300, 2016.
14) “Lytro,” http://lytro.com/.
15) K. Takahashi, Y. Kobayashi, T. Fujii: “Displaying Real World Light Fields Using Stacked LCDs,” International Display Workshops in conjunction with Asia Display 2016, DEStIS-D8-1, Invited talk, 2016.
16) R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” Computer Science Technical Report CSTR, vol. 2, no. 11, pp. 111, 2005.
17) K. Takahashi, T. Naemura: “Layered Light-Field Rendering with Focus Measurement,” EURASIP Signal Processing: Image Communication, vol. 21, no. 6, pp. 519–530 (2006.6).
18) G. Dansereau et al.: “Decoding, calibration and rectification for lenslet-based plenoptic cameras,” Computer Vision and Pattern Recognition, IEEE Conference, pp. 1027–1034, 2013.
19) https://youtu.be/6CM9sqTLRk0

Yuto Kobayashi received a B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 2016, where he is currently working toward a graduate degree in electrical engineering and computer science. His research project includes software/hardware development of a light field display.

Shu Kondo received a B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 2015, where he is currently working toward a graduate degree in electrical engineering and computer science. He is working on light field acquisition and rendering for light field displays.

Keita Takahashi received B.E., M.S., and Ph.D. degrees in information and communication engineering from the University of Tokyo, Tokyo, Japan, in 2001, 2003, and 2006. He was a Project Assistant Professor at the University of Tokyo from 2006 to 2011 and was an Assistant Professor at the University of Electro-Communications from 2011 to 2014. He is currently an Associate Professor at the Graduate School of Engineering, Nagoya University, Nagoya, Japan. His research interests include computational photography, image-based rendering, and 3-D displays.

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