https://doi.org/10.2352/ISSN.2470-1173.2019.3.SDA-625

3D Image Processing – from Capture to Display –

Toshiaki Fujii: Nagoya University; Nagoya, Japan

Abstract

Over the last 25 years, we have been involved in 3D image processing research field. We started our researches related to 3D image processing with "Data Compression of an Autostereoscopic 3-D Image" and presented our work in SPIE SD&A session in 1994. We first proposed the ray space representation of 3D images which is a common data format for various 3D capturing and displaying devices. Based on the ray space representation, we have conducted various researches on 3D image processing, which include: ray space coding and data compression, view interpolation, ray space acquisition, display, and a full system from capture to display of ray space. In this paper, we introduce some of our 25-year researches in terms of 3D image processing – from capture to display –.

Introduction

3D TV has become a popular media, which can provide viewers with stereoscopic visual effects and hence immersive viewing experiences. In accordance with its increasing growth, many works regarding end-to-end 3D TV systems have been proposed, which are composed of stereo/multiview cameras, 3D display with/without glasses. In principle, 3D TV requires multiple of images with the view number ranging from two to tens or hundreds, but the parameters such as camera distances for capturing multiview images vary from systems to systems and it is difficult to develop common capturing/display systems that meet these various requirements. We proposed, in 1994, the ray space concept [2] that provides the solutions for this problem. It was originally proposed as a common data format for 3D visual communications. It is a novel method to genarate photo-realistic virtual scenes, free-viewpoint images, and variable-focused images without complicated analysis and rendering processes.

In this paper, we review our researches related to 3D image processing in terms of ray space data processing. We first introduce the ray space representation of 3D scene, and then we describe ray space acquisition systems and introduce some example systems developed so far, including 100-camera system, time-division acquisition system, and coded aperture camera system. We finally introduce 3D display researches, which is incorporated to our end-to-end 3D communication system.

Definition of ray space

We see a 3D scene by our eyes that have the same mechanism as optical cameras. An optical camera is a device which record light rays from a 3D scene. This means that we obtain our visual information from a collection of light rays from the scene. Therefore, if we can represent a collection of light rays, we can describe 3D visual information of the scene. Based on this observation, we proposed ray space concept for a common data format for 3D image communications [2, 3]. In the Computer Graphics field, a similar idea was proposed in 1996 as a method of generating photo-realistic images. It was named as light field [5] or Lumigraph [6]. The ray space and light field are mathematically the same and mutually convertible. They can be seen as a reduced-dimensional (projected) version of Plenoptic function which was first introduced as a model of human early vision [1] .

Figure 1 shows one of the parameterization methods of the ray space [2]. A light ray in a 3D scene is parameterized using 4 variables; in Fig. 1, the ray parameters are defined using the position (x,y) where the ray intersects the reference plane (i.e. $Z=0$), and its outgoing direction (θ,ϕ). A view image corresponds to a collection of light rays which pass through the optical aperture of a camera. If we assume a pinhole camera model, we can show that the trajectory of the rays in the ray space form a 2D hyperplane in the 4D ray space (x,y,θ,ϕ) from the geometrical relationship. When we set a pinhole camera at a certain position, the camera "samples" the data on the 2D hyperplane of the 4D ray space. If we put many cameras in the scene, they sample 2D hyperplanes of the 4D ray space with different parameters. Thus, we can construct the whole 4D ray space by putting many cameras in the scene. Once we obtain the whole 4D ray space, free-viewpoint images can be easily generated by simply "resampling" a new hyperplane whose parameters corresponding to a

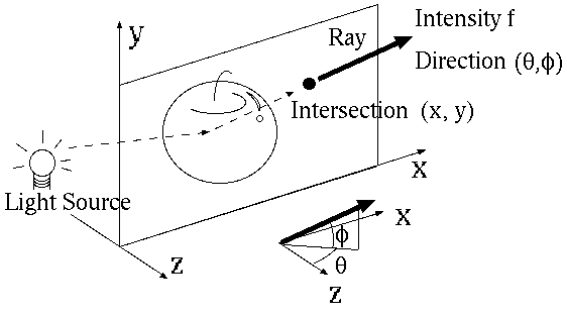


Figure 1. *Definition of ray space.*

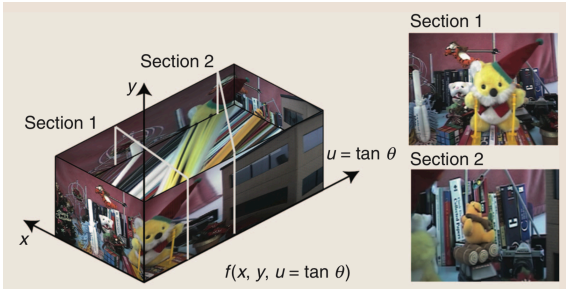


Figure 2. *Ray-space representation and free-viewpoint image generation [4].*

new camera position shown in Fig. 2. To sum up, capturing and displaying multiview images mean sampling and resampling of 4D ray space, and therefore, it can be used as a common 3D TV data format. In the following, we describe ray space acquisition and ray space display from this perspective.

Acquisition of Ray Space

Ray space data acquisition is a very challenging task and a lot of works have been done on this topic. An acquisition system of ray space requires an optical sensor that is located at the position (x,y) and senses the intensity of a ray coming from the direction (θ,ϕ). In practical use, a 2-D sensor is usually used but note that it captures a collection of rays that pass through the point (x,y) at the same time.

Ray space capturing systems are categorized into 4 types in terms of how to separately obtain each light ray: (1) space division acquisition, (2) multiple cameras, (3) time division acquisition, and (4) computational photography. Among them, Integral Photography (IP) and commercialized light field cameras such as Lytro and Raytrix fall into the category (1). An IP and light field camera consist of an imaging device and a lens plate which is located in front of the imaging device with a large number of small lens on the plane. Through the IP lens plate, 4D ray space is "encoded" to an array of thousands of small "pictures", thus forming one 2D "Integrated Image". Here, we focus on the categories (2)-(4) and introdue our researches conducted so far.

100-camera system

We reported a multi-dimensional multi-point measuring system, which is also called by its function a 100-camera and 200-microphone system [13]. Although similar camera array systems have already been reported (*e.g.* by Wilburn *et al.* [12]), our system has the following special features:

- Scalable multi-channel recording system (no limitation of channels)
- Simultaneous recording of video and analog signals
- High accuracy of synchronization ($< 1\ \mu$s)
- High image resolution (1392 H $\times$ 1040 V)
- "Uncompressed" raw data capturing
- C-mount lenses for high resolution cameras are available
- Synchronization among remote sites (using GPS, $<$ 1 ms)
- Long recording time ($>$ 1 hour)

The system consists of a system control unit, a system server unit, a synchronous control unit, and one hundred of recording units (nodes). The server unit consists of a system control unit and a system server unit. Although the system control unit and the system server unit are separate, one PC can serve as the both units. The performance of the PC does not need to be very high, and therefore, a commercially available PC can be used. The system server unit also serves as a user interface. The system control unit is connected to the synchronous control unit with RS-232C. It controls the generation of synchronization signal and therefore recording timing. The recording unit (called node) is a PC-based system which is equipped with specially developed custom boards. These boards are: (1) a module which controls the record of video data, (2) a base module which controls the record of analog signal data, and (3) an analog signal processing module. A node inputs one video data via CameraLink interface and 2ch (4ch maximum) analog signal. As for video capturing, since the dot clock is very high (50 MHz), transfer speed exceeds to 32 bit PCI bus and single HDD interface. We overcame this problem by adopting RAID technique to record high-bandwidth video data. The high-bandwidth data is divided into two and recorded on the two HDDs simultaneously. The nodes receive synchronization signal and sample video and analog signal in time with the sync signal. Since a node is driven by Linux operating system, it can flexibly execute remote commands via network. This feature enables us to construct flexible software environment.

The image resolution is 1392 (H) $\times$ 1040 (V), 8 bits/pixel. The camera has a CCD imager with a Bayer color filter. The interface between camera and PC is CameraLink (TM). The camera accepts external exposure signal, so generated synchronization signal is used as the external exposure signal. Considering accurate synchronization of video and analog signal, we set the frame rate 29.4118 frames per second for the system. As for analog signal input, various signals can be input. If we use many microphones, we can construct a high-channel microphone array. One of interesting applications is for ITS applications; various types of sensor can be used to sense such data as car speed, rotating speed of the engine, air temperature, heart rate of a driver, etc.

A synchronous control unit is composed of three components video synchronization signal generator, analog synchronization signal generator, and GPS (Global Positioning System) module. The sampling interval of video is set to be multiple of analog signal interval. This enables us to avoid frame drop and high accuracy of synchronization between video and analog signal is realized. The sampling interval of video is 29.4118 frames per second, and that of analog signal is up to 96kHz. The synchronization signal is transmitted via the cable with the delay of 5 ns/m. One buffering of the synchronization signal can cause 40 ns delay.

MPEG Test Sequences

We provided MPEG (Moving Picture Experts Group) with test sequences captured by the 100-camera system for Multiview Video Coding (MVC) and 3D Video (3DV) activities. Since MPEG is targeting to decide international standards for video coding, test sequences for the multiview and 3D video coding experiment must be uncompressed. In this sense, our 100-camera sys-

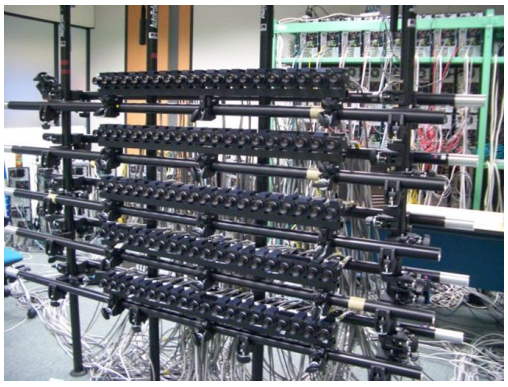


Figure 3. *100 camera system [13].*

tem which can capture raw video data is suitable for the purpose.

For the MVC experiment, we captured two test sequences with different camera arrangement: 1D line, and 2D array. The first sequence is "Rena" captured with 1D line arrangement, in which 100 cameras are aligned in a line with the camera interval 5 cm, and hence, the viewing zone is 5 meters in length. The second sequence is "Akko&Kayo" captured with 2D array camera arrangement, in which 100 cameras are aligned in 20 (H) x 5 (V) in camera interval 5 cm and 20 cm, respectively.

For the 3DV experiment, we provided "Champagne_tower" and "Pantomime" that were captured with 1D parallel camera arrangement. As MPEG called for multiview sequences captured with moving camera set, we developed a movable camera mount and captured multiview images by the moving camera array. Two sequences were adopted for 3DV experiment: Balloons and Kendo shown in Fig. 4

Time-division Acquisition System

One of the alternative acquisition systems is a time-division ray space acquisition system. We developed such a system, which consists of an optical imaging system, an optical scanner, and a high-speed camera. The optical system consists of a set of two parabolic mirrors which produces a real image of an object at its focal point, like a "floating" image. We set a galvanometer mirror at the position of the real image, which reflects the real image to various directions depending on the angle of the mirror. A high speed camera captures the reflected image at very high speed in a synchronized manner with the angle of the mirror. This system can capture equivalently about 300 multiview images at 30 fps with the viewing angle 55 degree. Figure 5 shows the overview of the system.

Coded Aperture Camera

The acquisition techniques mentioned above directly capture a ray space on a one-light-ray-by-one-pixel basis. On the other hand, there are other approaches called coded aperture/mask cameras, which can resolve the trade-off between spatial resolution and angular resolution in acquiring the ray space. Coded aperture camera is a ray space capturing system that records a linear combination of sub-aperture images and recovers the original full light field through computation shown in Fig. 6. Using the coded aperture cameara, we can reconstruct the entire ray space, which is equivalent to many view images, from only a few images that are captured through different aperture patterns. The problem here is how to reconstruct the target ray space from the set of captured images in Fig. 6. In previous works, this problem has often been discussed from the context of compressed sensing (CS), which

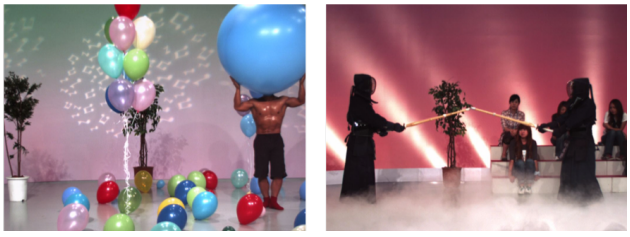


Figure 4. *MPEG test sequence: Balloons and Kendo.*

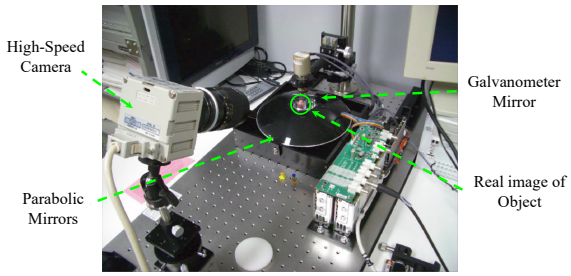


Figure 5. *Time-division ray space acquisition system.*

provides a sophisticated framework for signal reconstruction from a limited number of samples, where sparse representations on a pre-trained dictionary or basis are explored to reconstruct the target signal [10].

In contrast, we formulated this problem from the perspective of principal component analysis (PCA) and non-negative matrix factorization (NMF) [11]. In this method, only a small number of basis vectors are selected in advance based on the analysis of the training dataset. From this formulation, we derived optimal non-negative aperture patterns and a straight-forward reconstruction algorithm. Experimental results validated the effectiveness of our proposal; our method is superior to the state-of-the-art CS-based method in speed and accuracy of ray space reconstruction.

There is a drawback in the above mentioned PCA/NMF-based method , however, that it requires complex computation to reconstruct the original ray space, although it is far less than the original CS-based method. We can successfully replace the complex computation with a deep neural network (DNN) based method. In the following, we will explain our approach according to [18]. In our method, we formulated the entire pipeline of ray space acquisition as an auto-encoder. We implemented this auto-encoder as a fully convolutional neural network (CNN) and trained it end-to-end by using a collection of training samples.

Figure 7 shows how we modeled the ray space acquisition using CNN. The basic structure of the network is an auto-encoder. The auto-encoder employs a simple structure, in which an encoder network is connected to a decoder network, and it is trained to best approximate the identity mapping between the input and output. In our method, an encoder and decoder correspond to the image acquisition and reconstruction processes, respectively, because the original ray space is once reduced (encoded) to only

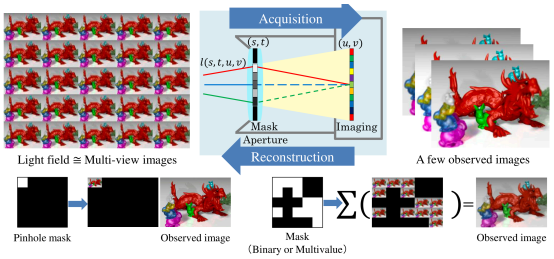


Figure 6. *Coded aperture camera model. Each light ray passing through (s,t) on the aperture plane is attenuated by the transmittance at (s,t) and reaches a pixel (u,v) on the imaging plane [11].*

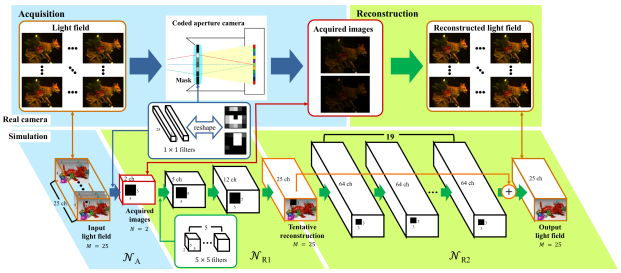


Figure 7. *Modeling light-field acquisition using CNN.*

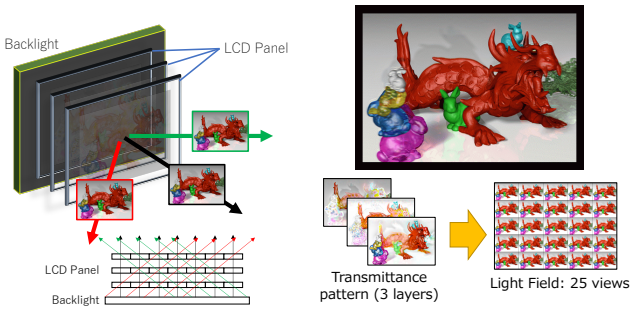


Figure 8. *3D layer (Tensor) display.*

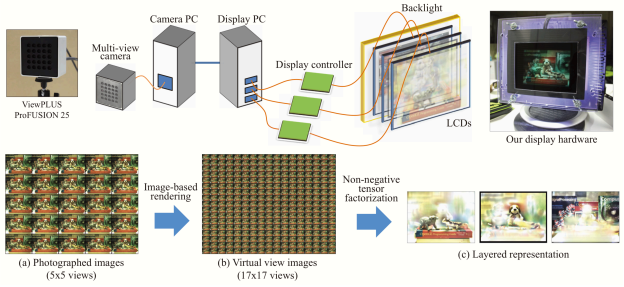


Figure 9. *End-to-end system and process pipeline of our light field display.*

a few images through the physical imaging process of the coded aperture camera, and these images are then combined to reconstruct (decode) the ray space with the same size as the original one. The parameters of the trained network correspond to the aperture patterns and reconstruction algorithm that are jointly optimized over the training dataset. In short, our method can learn to capture and reconstruct a ray space through a coded aperture camera by utilizing the powerful framework of a DNN.

Display of Ray Space

Light field display is a kind of 3D display which is capable of reconstructing a light field. There are many types of light field displays, such as a lenslet-based display (e.g. Integral Photography) and a barrier-based display, etc. Here we focus on a stacked layer type display [19]. In the following, we describe how the stacked layer light field display works and show that it requires huge computation to calculate the layer pattern.

The principle of the stacked layer display is shown in Fig. 8. It consists of a few light-attenuating layers located in front of a backlight. Each pixel of the light attenuating layers has an individual transmittance. When a viewer sees the display, the viewer observes the light rays that are emitted from the backlight and attenuated by the multiple layers. These layers overlap with different amount depending on the viewing direction. By designing the layer transmittance pattern appropriately, we can make the observed light rays to correspond to the target light field. The problem here is how to design the layer pattern under the condition that the target light field is given. 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, 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. The optimization is formulated as non-negative tensor factorization (NTF). Please reader to the original paper [19] for descriptions of the optimization method and the extension to time multiplexing. Since the transmittance values are alternately updated layer by layer, it requires heavy computations.

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 mitigate the heavy computation problem, we conducted the experiment where we adopted DNN for this calculation instead of iterative updates. The input to the network is the patches of the target light field data $I_{(i,j)}(x,y)$ and output is the patches of transmittance pattern $T_n(x,y)$. This one directional computation greatly reduces computations and increases the calculation speed, while avoiding computationally heavy iterations.

Full-Chain from Capture to Display

We developed a process pipeline from capture to display of a real 3-D scene [20]. To capture real 3-D scenes, we used a light field camera (Lytro Illum) and a multi-view camera (ViewPLUS ProFUSIOIN 25) to capture the ray space, which was then factorized into layer representations to be displayed and fed to the layer type display. In [20], we analyzed the amount of pop-out and motion parallax that can be presented by the display using a given light field data.

An important fact is that the required density (the viewpoint interval) for the input ray space data is a rather strict condition. For example, a ViewPLUS ProFUSION 25 camera has 25 (5 × 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. 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. To resolve this problem, we used image-based rendering to produce sufficiently dense multi-view images (virtual images) from real photographs. Figure 9 shows the developed pipeline from capture to display based on this concept.

Conclusion

In this paper, we introduced our research works on 3D image processing. First, we started with the ray space definition. The ray space method requires a very high number of "rays", so its acquisition system needs a large amount of equivalent "pixels"

accordingly. As such acquisition systems, we first introduced a multi-dimensional multi-point measuring system, so-called 100-camera system. The system has the special features: capable of acquiring high-resolution uncompressed raw video, high accuracy of synchronization between video and analog signal, and synchronization in remote sites using GPS sensor. Secondly, we introduced "ray space camera", which falls into two categories: a space-division system and a time-division system. The bottleneck of the ray space camera is the needs for high-resolution devices for acquisition. The Integral Photography based system is such an example which requires high-resolution devices. On the other hand, the space-division system can mitigate this high requirement by encoding 4D ray space into 2D integral images. The time-division system introduced here is composed of a special optical imaging system and a high-speed video camera. We showed that the time-division system can capture a dynamic ray space with equivalently 300 views in 30 fps. We then introduced a layer 3D display which was originally proposed by Wetzstein *et al*. in [19]. Finally, we introduced a full-chain system from capture to display. The 3D scene is captured by multiple cameras and, through Image Based Rendering, layer patterns for the layer display are calcurated and fed to the display. Our future work is to develop a real-time ray space communication system which is composed of coded aperture camera, coding and transmission part, and 3D display. We have already succeeded the fast reconstruction of the views from a few shots obtained from coded aperture camera, and fast calculation of the layer pattern of 3D display. By cascading these components and optimizing the network configuration for CNN calculation, we would be able to develop the real-time full-chain system.

References

[1] E. H. Adelson, and J. Bergen, "The Plenoptic Function and the Elements of Early Vision," In Computational Models of Visual Processing, Cambridge, MA: MIT Press, pp. 3–20 (1991).

[2] T. Fujii, "A Basic Study on the Integrated 3-D Visual Communication," Ph.D. thesis of engineering, The University of Tokyo, 1994 (in Japanese).

[3] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray Space Coding for 3D Visual Communication," Picture Coding Symposium '96, pp. 447–451, (Mar. 1996).

[4] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-Viewpoint TV," IEEE Signal Processing Magazine, Vol. 28, Issue 1, pp. 67–76 (2011).

[5] M. Levoy and P. Hanrahan, "Light Field Rendering," In ACM SIGGRAPH '96, pp. 31–42 (Aug. 1996).

[6] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen, "The Lumigraph," In ACM SIGGRAPH '96, pp. 43–54 (Aug. 1996).

[7] T. Fujii and M. Tanimoto, "Freeviewpoint TV system based on Ray-space Representation," SPIE ITCom 2002, vol. 4864–22, pp. 175–189 (2002).

[8] T. Fujii, T. Kimoto, and M. Tanimoto, "A New Flexible Acquisition System of Ray-Space Data for Arbitrary Objects," In IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 2, pp. 218–224 (2000).

[9] T. Fujii, and M. Tanimoto, "Real-Time Ray-Space Acquisition System," In Proc. of SPIE Electronic Imaging, Vol. 5291, pp. 179–187 (2004).

[10] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive Light Field Photography Using Overcomplete Dictionaries and Optimized Projections," ACM Trans. Graphics (TOG), vol.32, no.4, Article No. 46 (2013).

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

[12] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High Performance Imaging Using Large Camera Arrays," ACM Transactions on Graphics (TOG) Vol. 24, Issue 3, pp. 765–776 (2005).

[13] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto, and Y. Suenaga, "Multipoint Measuring System for Video and Sound – 100–camera and microphone system –," In IEEE 2006 International Conference on Multimedia & Expo (ICME2006), pp. 437–440 (2006).

[14] R. Ng, M. Levoy, M. Bredif, G. Duval M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-held Plenoptic Camera," Computer Science Technical Report CSTR 2(11) pp. 1–11 (2005).

[15] E. H. Adelson, and J. Y. Wang, "Single Lens Stereo with a Plenoptic Camera," IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 14, No. 2, pp. 99–106 (1992).

[16] J. Arai, F. Okano, H. Hoshino, and I. Yuyama, "Gradient-index Lens-array Method Based on Real-time Integral Photography for Three-dimensional Images," Applied Optics Vol. 37, No. 11, pp. 2034–2045 (1998).

[17] R. Ng, "Digital Light Field Photography," Ph.D. Thesis, Stanford University (2006).

[18] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, "Learning to Capture Light Fields through a Coded Aperture Camera," European Conference on Computer Vision (2018).

[19] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor Displays: Compressive Light Field Synthesis Using Multilayer Displays with Directional Backlighting," ACM Transactions on Graphics (TOG) , Vol. 31, Issue 4, pp. 1–11 (2012).

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

Author Biography

Toshiaki Fujii received his B.E., M.E., and Dr.E. degrees in electrical engineering from the University of Tokyo in 1990, 1992, and 1995, respectively. He is currently a Professor at the Graduate School of Engineering, Nagoya University, Japan. His current research interests include multi-dimensional signal processing, multi-view video coding and transmission, and 3D imaging system based on light field acquisition and display. Prof. Fujii is a member of IEEE Signal Processing Society.

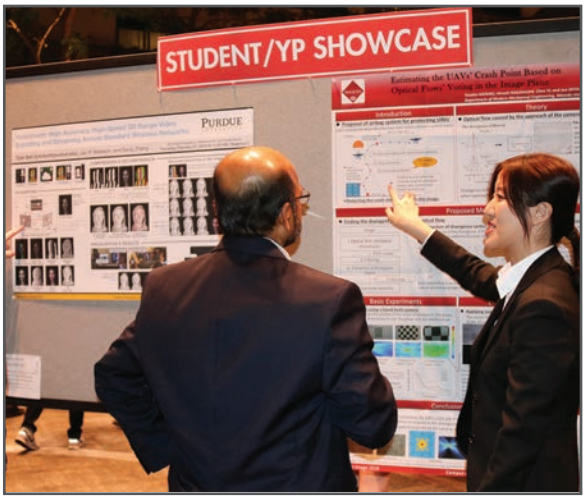
JOIN US AT THE NEXT EI!

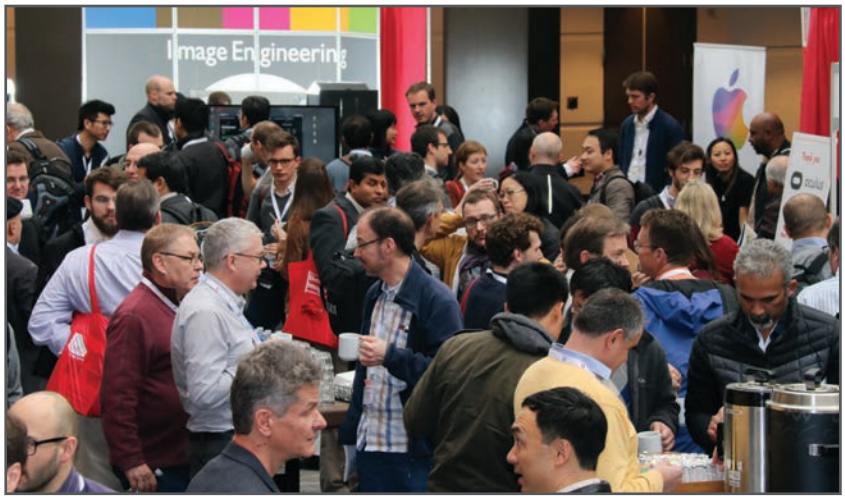
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org