Video Processing/Display Technology for Reconstructing the Playing Field in Sports Viewing Service Using VR/AR

Megumi Isogai, Kazuki Okami, Masaaki Matsumura, Munekazu Date, Akio Kameda, Hajime Noto, and Hideaki Kimata

Abstract

Sports viewing services using virtual reality (VR) and augmented reality (AR) technology have been introduced recently due to the development of sensor devices and video production/delivery technology. NTT Media Intelligence Laboratories aims to provide a sports viewing service using VR/AR to display reconstructed playing field videos without user operation. In this article, we present a video processing technology to reconstruct the playing field for VR/AR displays and a three-dimensional display technology for representing a playing field three-dimensionally on a table.

Keywords: virtual reality/augmented reality, video processing technology, 3D display technology

1. Introduction

Advances in sensor devices and video production and delivery technology have led to the introduction of sports viewing services using virtual reality (VR) and augmented reality (AR) technologies. NTT DOCOMO demonstrated an AR sports viewing service for rugby games that displays player information and video in a high visibility area through a user's smart glass [1]. KDDI has already launched a VR sports watching service that provides VR views while switching among viewpoints from five cameras in a baseball stadium [2].

However, neither AR nor VR can cover all of the game scenes in a stadium because suitable scenes for VR/AR change in response to the constantly varying game state. To enable viewers to have a good game watching experience, it should be possible to change VR/AR viewing modes without any need for user operation. NTT Media Intelligence Laboratories thus aims to give the crowd in the sports stadium an additional element of excitement by using VR/AR to display reconstructed playing field videos without user operation. To achieve this, we have been studying video processing technology to generate common content for VR and AR, as well as video display technology to create a feeling of physical presence, as if real objects were actually there.

In this article, we first give an overview of a sports viewing system we developed that uses VR/AR. We then present a three-dimensional (3D) reconstruction technology we developed as a means for processing videos to reconstruct the playing field for VR/AR displays. It generates arbitrary viewpoint images from places where cameras cannot be placed on the playing field. We also describe our diminished reality technology that focuses on particular players by removing everything except the players from the video. Finally, we introduce a novel visually equivalent light field 3D display technology that we propose

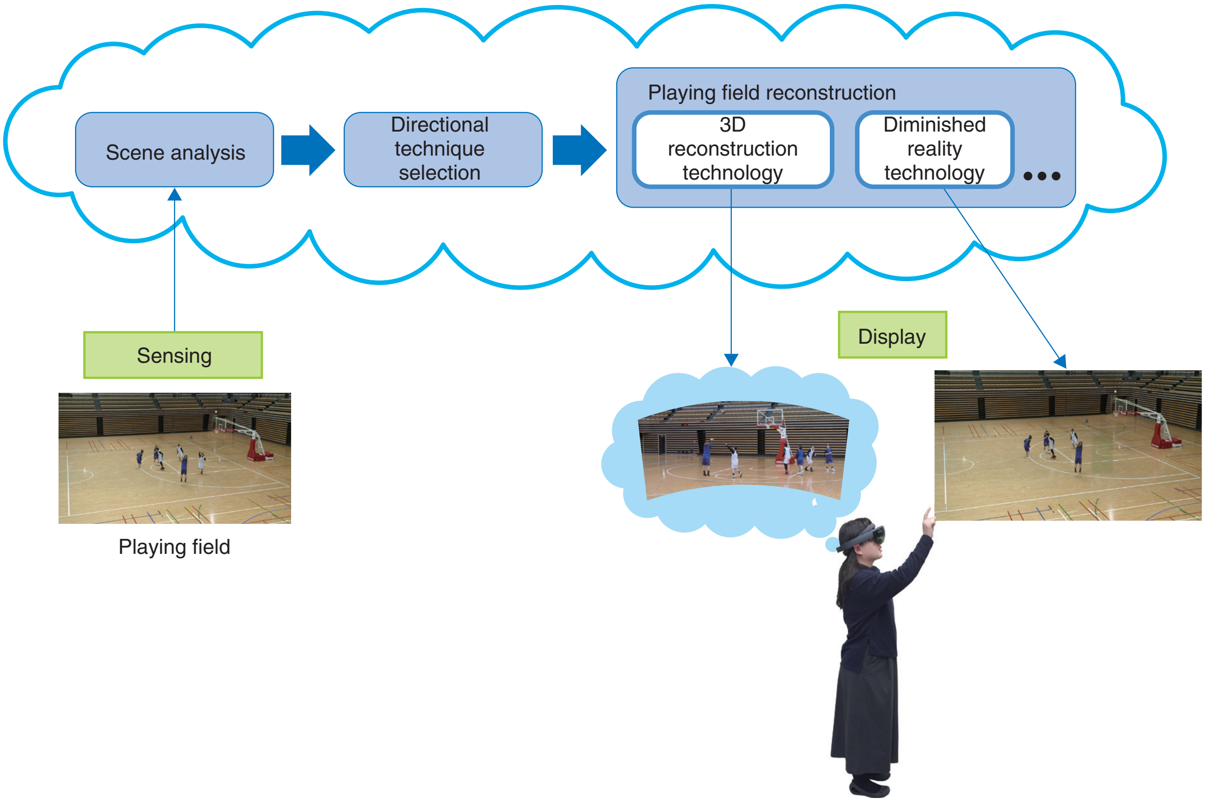

Playing field reconstruction
Scene analysis
Directional technique selection
3D reconstruction technology
Diminished reality technology
•••
Sensing
Display
Playing field

Fig. 1. Overview of sports viewing system using VR/AR technologies.

as a future AR method for representing a playing field three-dimensionally on a table.

2. Overview of sports viewing system

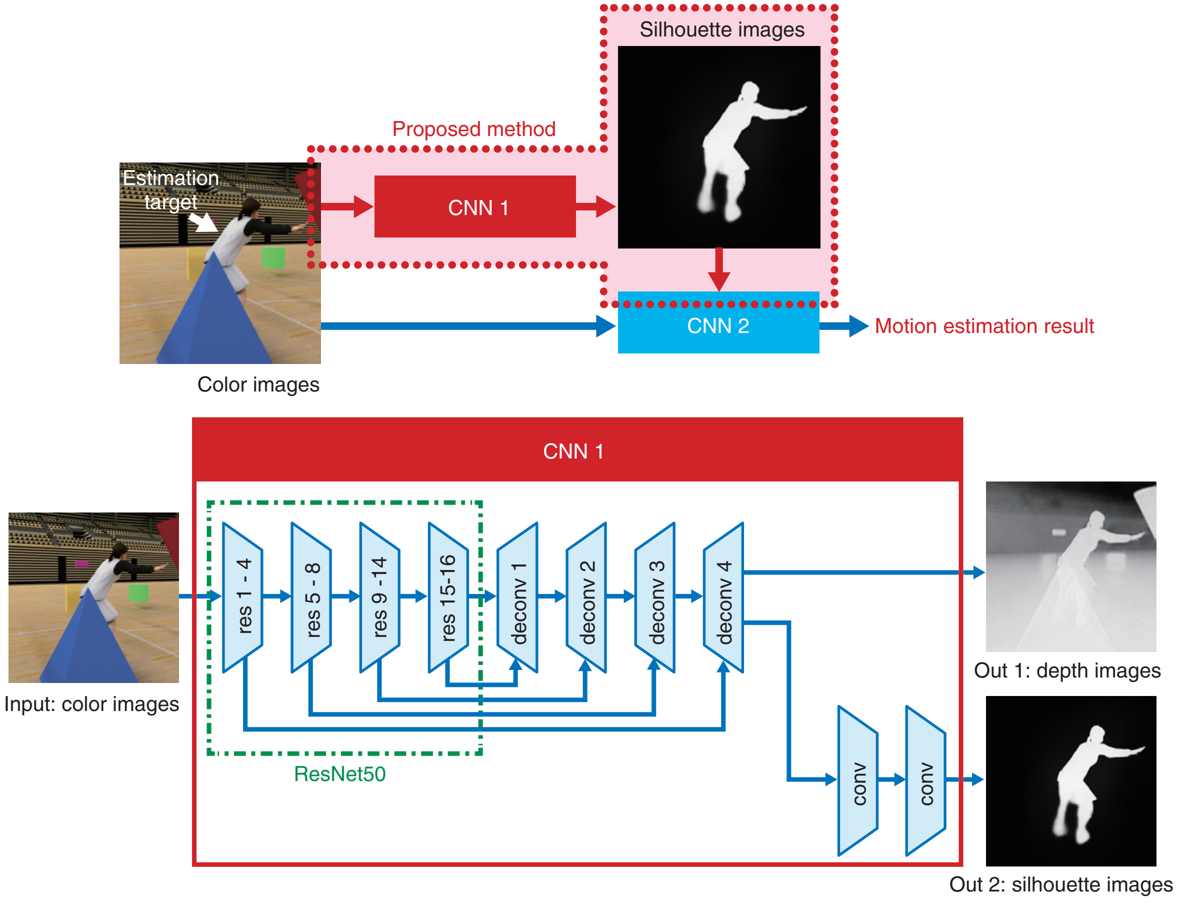
Our system for watching sports in a stadium using VR/AR technologies is shown in **Fig. 1**. It features multiple cameras that surround the stadium and capture videos of the playing field. The videos are delivered to a server on a cloud. The server uses the videos to analyze the event that occurred on the playing field and selects the videos to provide to the users from the analysis results. The playing field is then reconstructed from the videos based on preprocessing results and distributed to the user viewing devices.

3. 3D reconstruction technology for the playing field

One of the video processing technologies used to reconstruct the playing field is 3D reconstruction technology. It enables viewpoint images to be generated from places inside the playing field where cameras cannot be installed. Since the spectators' seats are fixed in stadiums and arenas, the spectators can only watch the event going on from a limited direction. They cannot see players if their seats are far away from the playing field, and often cannot watch the event from the viewpoint they would like in order to get the best view of how the event is developing.

A method has been developed to generate arbitrary viewpoint images from multiple cameras that surround the playing field. However, it requires the installation of a very large number of cameras, which is difficult to do in a stadium. Also, the method cannot generate high quality viewpoint images of what is happening in front of the camera because of insufficient video resolution and because 3D information cannot do estimations in occluded areas.

To address these problems, we propose a new form of 3D reconstruction technology based on computer graphic (CG) characters. It estimates a player's motions from videos and applies the estimation results to a preprepared high quality CG model of the player. With the notably improved CG quality for movies and games achieved in recent years, our



CNN: convolution neural network
conv: convolution
deconv: deconvolution
ResNet: Residual Network

Fig. 2. Overview of proposed method.

method generates high quality viewpoint images that make users feel as if they were watching scenes in the real world.

To achieve this, it is necessary to estimate player motions from videos. While conventional methods have been proposed to estimate human motions from videos using deep learning [3], the pose estimation accuracy they provide is substantially reduced when the people being videoed are partly occluded by objects. An important task is to improve robustness in occluded areas, because in sports scenes many athletes are frequently intertwined in one scene and can easily shield each other.

Thus, we propose the idea of estimating athletes' motions not from color images directly but from color images and silhouette images of people, including those in occluded areas (**Fig. 2**). By using silhouette images to limit the area for searching human motions, our method achieves higher human motion estimation accuracy than conventional ones can provide.

4. Diminished reality technology for the playing field

Here, we describe the diminished reality technology we developed that removes unnecessary players and objects from videos in order to focus on target players.

In movie and cartoon scenes, only the persons to be focused on remain in the scene, and effects such as a spotlight are added. We anticipate that applying this to sports events held in stadiums and arenas will ensure that increased attention is directed towards players involved in decisive moments of the game, which will make watching the game more exciting.

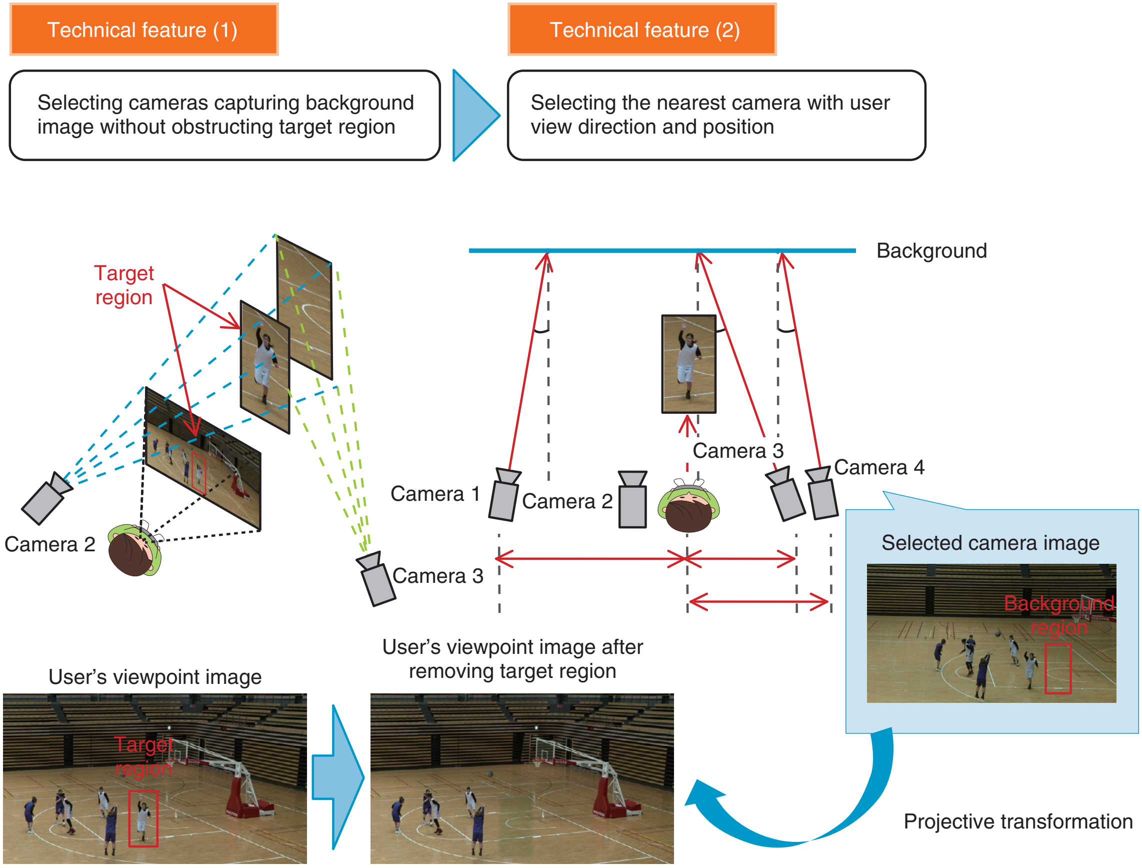

Technical feature (1)
Selecting cameras capturing background image without obstructing target region
Technical feature (2)
Selecting the nearest camera with user view direction and position
Target region
Camera 2
Camera 3
Background
Camera 1
Camera 2
Camera 3
Camera 4
Selected camera image
Background region
User's viewpoint image
Target region
User's viewpoint image after removing target region
Projective transformation

Fig. 3. Technical features of proposed method.

Diminished reality (DR) technology removes unnecessary target regions from video images and reconstructs the missing regions with background images. To apply this to sports events, DR technology must reconstruct only target regions while considering the relationship between players in 3D space. Furthermore, since spectators watch sports events in various positions and postures, it needs to present images in which the target region does not stand out above all others for each spectator's viewpoint.

To address these points, we propose a DR method that selects the optimal camera capturing the background of the target region from multiple cameras installed in a stadium and overlays the transformed image so that it matches that seen from the user viewpoint. This prevents the target region from becoming overly conspicuous [4].

The technical features of our proposed method are shown in **Fig. 3**. The first feature is that the playing field is treated as a multilayered plane, and 3D information is estimated on that basis, since a wide range of depth is unnecessary to reconstruct the playing field. This enables it to select the group of cameras capturing the background of the target region with less calculation than that needed in methods estimating 3D information of the entire playing field.

The second feature is that the method automatically determines the optimum camera—one that is near the user's location and thus catches the action from a direction and position similar to that of the user. This makes it possible to provide images that are natural from a user's viewpoint by simply applying projective transformation to selected camera images.

5. Visually equivalent light field display technology

Finally, we introduce a novel visually equivalent light field 3D display technology, which we propose as a future AR method that three-dimensionally represents a playing field on a table. We believe that showing game highlights at the box office, lobby, or

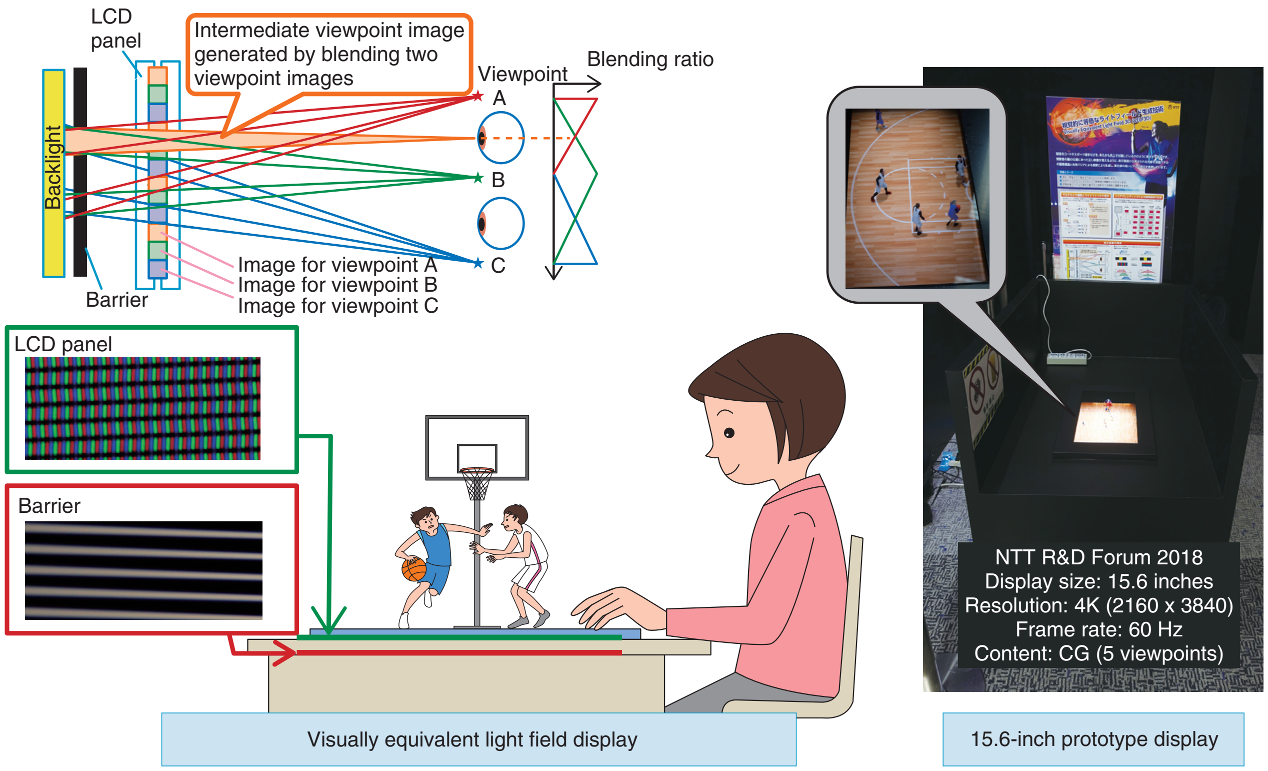

LCD panel
Intermediate viewpoint image generated by blending two viewpoint images
Viewpoint
Blending ratio
A
B
C
Backlight
Barrier
Image for viewpoint A
Image for viewpoint B
Image for viewpoint C
LCD panel
Barrier
NTT R&D Forum 2018
Display size: 15.6 inches
Resolution: 4K (2160 x 3840)
Frame rate: 60 Hz
Content: CG (5 viewpoints)
Visually equivalent light field display
15.6-inch prototype display

Fig. 4. Overview of visually equivalent light field display.

some other place outside the playing venue will increase people's interest in certain players and teams so they will come to see them again and again. To provide spectators with a more attractive viewing experience, it is important to show them the players in a way that makes them feel as if the players were actually there in front of them.

Objects such as players, fields, and balls emit rays with different colors and brightness in each direction by reflecting light or by emitting light themselves. Fields generated with these rays, called light fields, can be made to seem highly realistic by accurate ray reproduction. However, to display different rays depending on the direction, we allocate pixels in accordance with the number of directions. For example, generating rays in 100 directions requires 100 times more pixels than needed for a 2D display with the same resolution.

Human vision perceives objects from incident rays entering the pupils of the eyes. This perception involves not only acquiring simple images like pictures taken by a camera, but also using differences in rays between the left and right eyes and small temporal variations of rays due to viewing position changes induced by fluctuations of posture or eye movement. This perception seems to be high at first glance but is insensitive to elements that are not necessary. Therefore, a visually equivalent light field should exist. Though it is different from the light field of an actual object, human vision perceives it as the same as that of an actual object.

This is the concept of a visually equivalent light field display. With this technology, light rays to an intermediate viewpoint are interpolated with visual equivalence. The interpolation is a weighted average of rays to discrete viewpoints that is optically created in the display [5].

An overview of a visually equivalent light field display is given in **Fig. 4**. As shown in the top left corner, the display consists of an LCD (liquid crystal display) panel, a stripe-pattern barrier, and a backlight, in order from the rear. When an observer's pupils are at viewpoint A, rays from the barrier spacing illuminate the pixels of the viewpoint A image only, and that image can be seen. When an observer's pupils are at viewpoint B, the image for viewpoint B can be seen. When an observer's pupils are midway between viewpoints A and B, the pixels for both images A and B are partially illuminated, and a weighted average is achieved. Since the weights depend on the distance between pupil position and viewpoints A and B, an intermediate viewpoint image is perceived as expected. Because the image quality is high when a displayed object is close to the display panel, we placed the panel horizontally as shown at the bottom left in the figure. This made it possible to display and maintain high resolution, high quality images in every corner of the court.

Reproducing the light field in this way enables the correct depth to be perceived even if the distance between the left and right eyes is different (e.g., for an adult and a child). We believe the efficacy of interpolation can not only improve pixel usage efficiency but also improve feelings of object existence or reality.

6. Future work

We have improved VR/AR technologies and expect that sports viewing services using them will increase as interest in sports increases with the approach of the major international sports event in 2020. The research on scene analysis technology NTT Media Intelligence Laboratories is conducting will facilitate the selection of video effects and VR/AR displays that will accord with game situations. The development of video processing technology will facilitate the reconstruction of playing fields and visually equivalent light field 3D displays. This will enable us to contribute to the development of innovative ways to provide people with a new and enhanced sports watching experience.

References

[1] Press release issued by NTT DOCOMO on December 21, 2017 (in Japanese).
https://www.nttdocomo.co.jp/binary/pdf/info/news_release/topics/topics_171221_01.pdf

[2] Press release issued by KDDI on July 25, 2018 (in Japanese).
http://news.kddi.com/kddi/corporate/newsrelease/2018/07/25/3281.html

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields,” Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 1302–1310, Honolulu, HI, USA, July 2017.

[4] M. Isogai, D. Ochi, and H. Kimata, “Diminished Reality Using Plane-based Reconstruction Method,” The 17th International Meeting on Information Display (IMID 2017), F38-4, 3DSA, Busan, Korea, Aug. 2017.

[5] M. Date, D. Ochi, and H. Kimata, “Visually Equivalent Light Field Flat Panel 3D Display,” Proc. of the 22nd Annual Conference of the Virtual Reality Society of Japan, Tokushima, Japan, Sept. 2017 (in Japanese).

Megumi Isogai

Senior Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.

She received a B.E., M.E., and Ph.D. in communication network engineering from Okayama University in 2004, 2005, and 2010. She joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2006 and has been studying 3D video processing and high-reality communication technology.

Kazuki Okami

Researcher, Visual Media Project, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Waseda University, Tokyo, in 2012 and 2014. He joined NTT Media Intelligence Laboratories in 2014, where he has been engaged in research and development (R&D) of free-viewpoint video synthesis.

Masaaki Matsumura

Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.

He received a B.S. in design from Kyushu Institute of Design, Fukuoka, in 2005, and an M.E. in design from Kyushu University, Fukuoka, in 2007. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2007 and has been researching a video coding algorithm and GPU (graphics processing unit)-accelerated high-performance computing.

Munekazu Date

Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.

He received a B.S. in physics from Gakushuin University, Tokyo, in 1990, an M.E. in applied electronics from Tokyo Institute of Technology in 1992, and a Ph.D. in chemistry from Tokyo University of Science in 2003. He has been with NTT since 1992, where he has been researching holographic optical devices using polymer/LC composites and 3D displays. He joined NTT COMWARE in 2010 and rejoined NTT in 2012. Dr. Date is a member of the Institute of Electrical and Electronics Engineers, the Society for Information Display, the Japanese Liquid Crystal Society, the Institute of Electrical Engineers of Japan, and the Holographic Display Artists and Engineers Club.

Akio Kameda

Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.

He received a B.E and M.E in electrical engineering from Tokyo University of Science, Chiba, in 1993 and 1995. In 1995, he joined NTT Human Interface Laboratories and has been engaged in R&D of video coding, communication, and distribution systems.

Hajime Noto

Senior Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. in industrial engineering from Kansai University, Osaka, in 1997 and 1999. In 1999 he joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories), where he researched 3D input systems. He is a member of the Institute of Image Information and Television Engineers and is currently involved in R&D of high-reality systems.

Hideaki Kimata

Senior Research Engineer, Supervisor, High-Reality Visual Communication Group Leader, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. in applied physics in 1993 and 1995 and a Ph.D. in electrical engineering in 2006 from Nagoya University, Aichi. He joined NTT in 1995 and has been researching and developing a video coding algorithm, visual communication systems, and machine learning. He is a Chief Examiner of the Information Processing Society of Japan (IPSJ) Special Interest Group on Audio Visual and Multimedia Information Processing.