Surround Video Stitching and Synchronous Transmission Technology for Immersive Live Broadcasting of Entire Sports Venues

Takako Sato, Koji Namba, Masato Ono, Yumi Kikuchi, Tetsuya Yamaguchi, and Akira Ono

Abstract

NTT Service Evolution Laboratories is promoting the research and development of the live immersive reproduction of entire spaces and environments by simultaneously transmitting video, audio, and related information from a remote location. In this article, we propose a technique to combine multiple 4K video pictures horizontally and vertically in real time to create a panorama video with high resolution and a wide viewing angle (at least 180°)—much higher than the resolution that can be achieved with a single camera. We also introduce technology to synchronously transmit other video and audio data to remote venues with low latency in addition to the stitched video.

Keywords: video stitching, media synchronous transmission, low latency

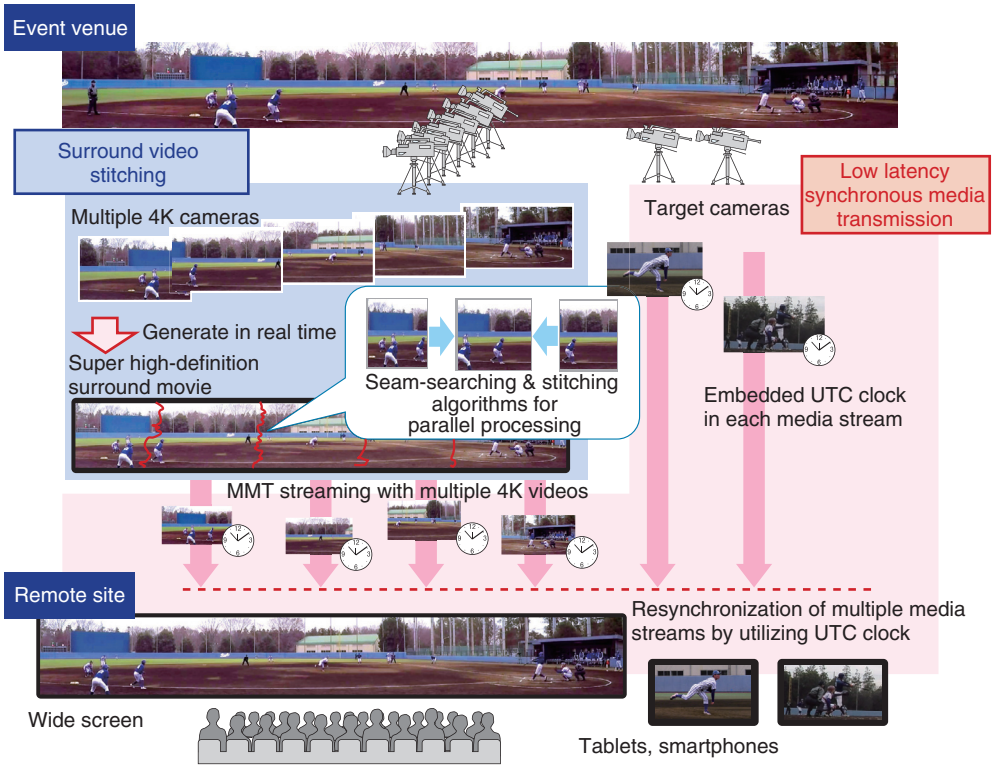
1. Introduction

NTT Service Evolution Laboratories is actively driving research into the live immersive reproduction of entire spaces and environments by simultaneously transmitting live video, audio, and related information from the source location. By establishing technologies for transmitting immersive live video to remote locations, we aim to provide immersive public viewing services that can provide captivating experiences of events such as sports matches and live shows; the goal is to enable viewers to experience such events as though they were actually at the venue.

In this article, we introduce surround video stitching/synchronous transmission technologies that can transmit live video of an entire sports event taking place at, for example, a large sports field, and reproduce it at a remote location so that it appears as if the action was taking place right in front of the audience. This technology can also be applied to public viewing for productions such as theater shows and musical concerts that take place on a large stage.

2. Surround video stitching/synchronous transmission technologies

In recent years, Hi-Vision and 4K video have made inroads in a wide variety of fields, and they are being increasingly incorporated into the development of higher-resolution 8K services, mainly in broadcasting applications. However, these services are not yet able to provide complete, unbounded views of sporting events taking place on large sports fields (e.g., baseball, soccer, and rugby). We are working to establish surround video stitching and synchronous transmission technologies that offer the real-time stitching and transmission of high-resolution video (beyond 4K/8K) covering a wide viewing angle so that people can view events as if they were actually there. These technologies consist of two core



MMT: MPEG Media Transport
UTC: Coordinated Universal Time

Fig. 1. Surround video stitching/synchronous transmission technologies.

components (**Fig. 1**):

- Surround video stitching: This combines multiple 4K video images both horizontally and vertically in real time to produce high-definition surround video with a wide viewing angle (180° or more) that greatly exceeds the resolution of images that can be taken with a single camera.
- Low latency synchronous media transmission: Surround video covering the entire sporting venue is partitioned into multiple images and combined with sounds from the event and close-up videos of the players, and then synchronously transmitted and played back with low latency on large-scale Hi-Vision systems at remote venues or on spectator tablet devices. MMT (MPEG* Media Transport) is used to transmit the content, and performing delay compensation according to the playback environment makes it possible to accurately synchronize multiple video and audio sources even on disparate playback devices.

3. Surround video stitching technology

In this section, we explain the various elements of surround video stitching technology.

3.1 Real-time video stitching issues and solutions

Various conventional algorithms can be used to create 4K video panoramas and 360° views, but these are all centered on techniques that demand off-line processing. For real-time processing of multiple 4K video images, amounting to approximately 15 gigabits of data per second, it is necessary to overcome many technical issues in terms of improving the processing speed. When creating panoramas from multiple video images, correction processing (homography transformation) is performed on each video image to reduce the effects of parallax arising from the use of multiple cameras. The stitching lines

* MPEG: Moving Picture Experts Group. A working group established by ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) to develop standards for digital audio and video compression.

(seams) between adjacent video images are then located by considering the movements and other attributes of people and other moving objects in the video, and finally, the multiple video images are merged to yield panorama videos. The seam positions are determined so as to avoid moving objects and thus prevent artifacts such as the stretch and truncation of moving objects that occur when moving objects cross a seam.

Our surround video stitching technology makes it possible to stitch, in real time, multiple 4K/60 fps video images that adjoin one another in the horizontal and vertical directions. The computationally intensive process of seam search, which involves video analysis, is conducted on multiple servers in parallel. By combining eleven 4K images on the horizontal plane, we can generate surround video with a 180° viewing angle. We describe below the real-time processing operations in more detail and introduce a camera rig that can shoot video appropriate for stitching.

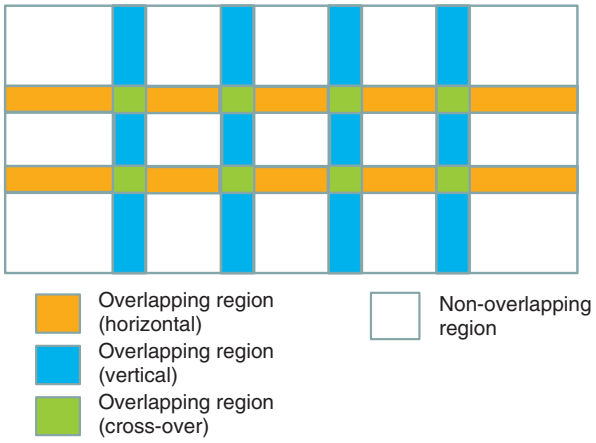


Fig. 2. Example of the overlapping regions used in 5(H) × 3(V) 4K video stitching.

3.2 Faster seam search

In searching for seams, information such as luminance differences (motion information) and edge intensities (shape information) obtained by analyzing the video frames is used, but in our surround video stitching technology, this analysis has to be performed in parallel in order to achieve sufficient speed. It is possible to use information from previous video frames to stabilize seam positioning so that the seams do not jump about unnecessarily, even for fast-moving objects. However, when video analysis and seam search are performed in parallel, it may not be possible to refer to the seam information of previous video frames. To address this problem, we perform seam detection processing on scaled-down video images to obtain approximate seam information for previous video frames, thereby achieving both high processing speeds and accurate seam detection.

To combine multiple 4K images that are tiled both horizontally and vertically, we have to perform stitching processing in many overlapping regions (**Fig. 2**). Since the sequential execution of stitching processing in the horizontal and vertical directions would increase the overall processing time, the seam search processing in each overlapping region is performed in parallel. If we search for seams separately in each area, it is possible that these seams will not line up properly from one region to the next. This makes it necessary to share information about the seam endpoints between adjacent areas so that continuous seams can be formed between regions in the horizontal and vertical directions (**Fig. 3**).

3.3 Synchronous distributed processing architecture

With our surround video stitching technology, the abovementioned correction processing and seam search/stitching processing are done across multiple servers in parallel, enabling even huge video images to be processed in real time. To finally merge the video images processed by each server into a single 4K × *N* surround video, we need to synchronize the various processes performed on each video frame and output the processing results synchronously. In this technology, synchronous control is facilitated by applying the same time stamp to video frames captured at the same time before they are distributed among the servers.

Instead of generating new time stamps every time a video frame is input, it is possible to generate time stamps from the time codes that are added by cameras or superimposed on the input video. Conversely, when the stitched video is output, the original time codes can be recovered and superimposed on the picture. In this way, we can propagate time codes assigned by cameras or other equipment directly to the later stages of our system (e.g., encoding or transmission equipment), enabling stitched video to be played back in perfect synchronization with other video materials that are transmitted directly without surround video stitch processing.

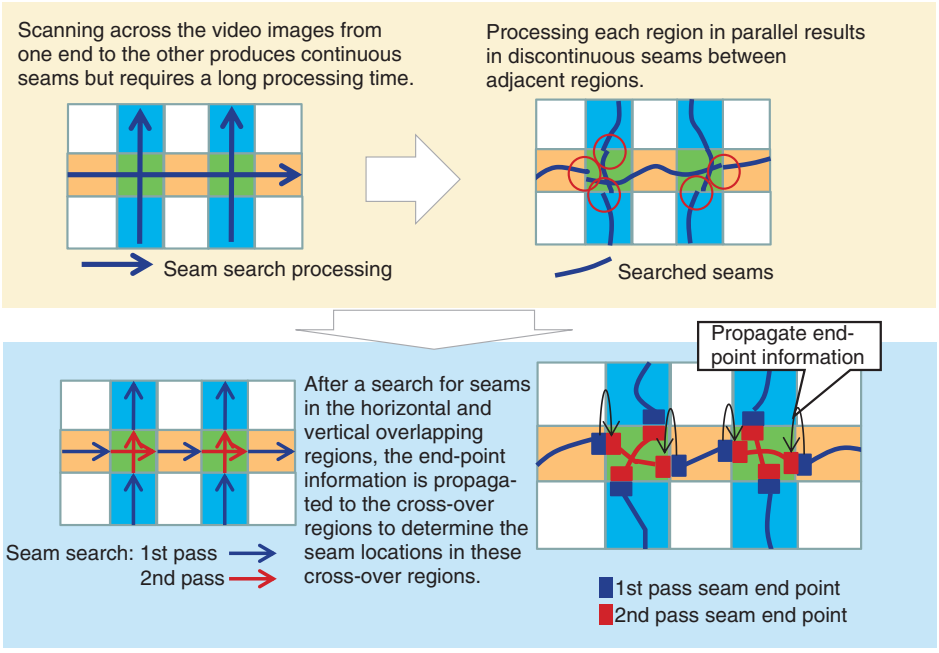


Fig. 3. Parallel processing to ensure seam continuity.

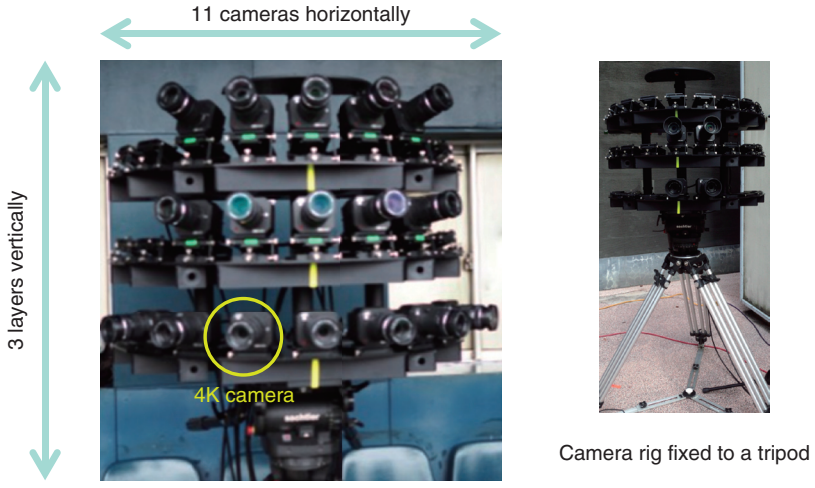


Fig. 4. Video camera rig with multiple 4K cameras mounted horizontally and vertically.

3.4 Precise video capture from multiple 4K devices

Before the video stitching is done, the effects of parallax arising from the use of multiple cameras are reduced by performing correction processes such as homography transformation. However, it is not possible to completely eliminate the effects of parallax in all regions of the videos because they include depth and movement. To create stitched images of the highest quality, we must accurately adjust the camera positions so as to minimize parallax in the source material. We have developed a video capture rig that can quickly and accurately adjust camera positions in the horizontal and vertical directions, and we use this rig to reduce the video capture workload and improve the quality of the stitched video (**Fig. 4**).

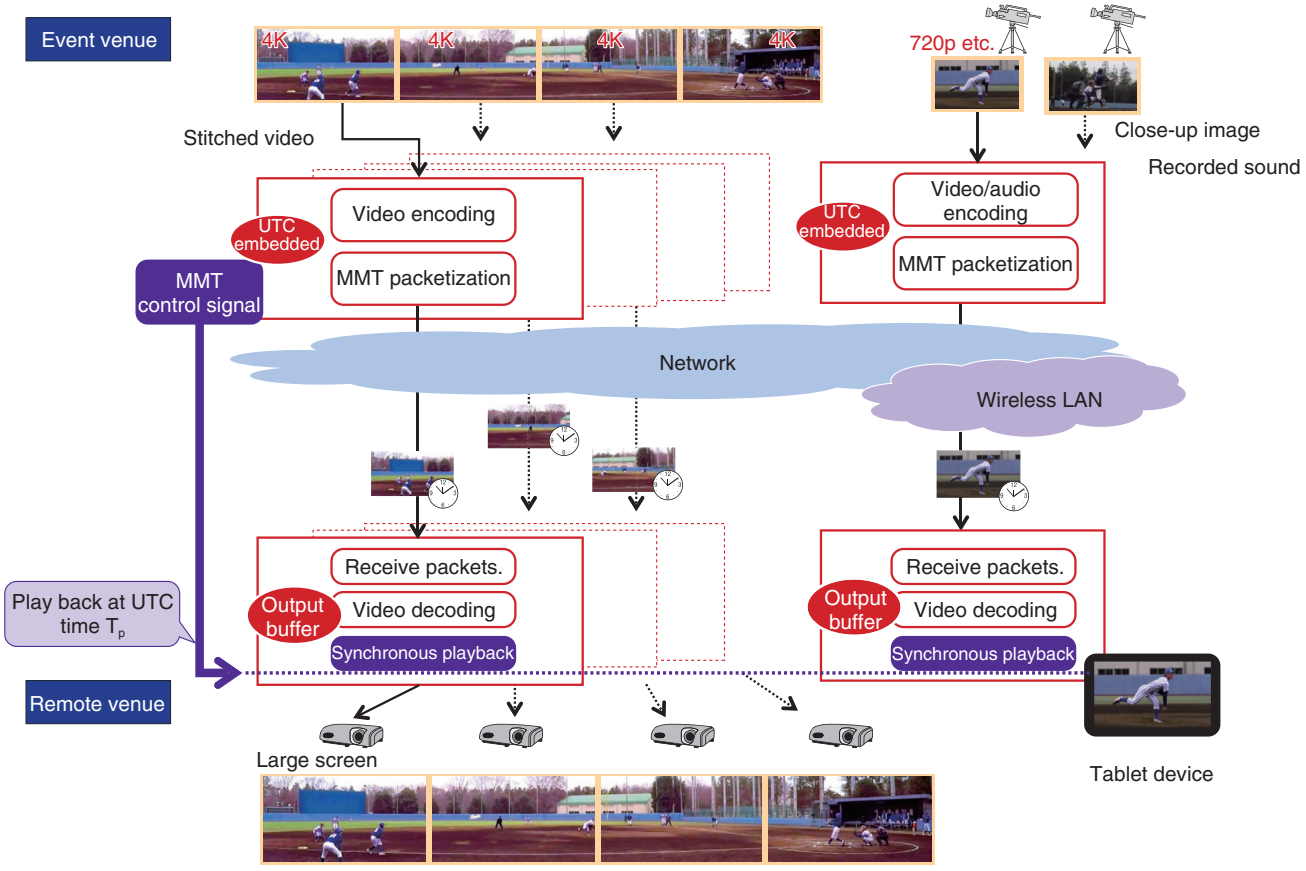


Fig. 5. Low-latency media synchronization and transmission.

4. Low-latency media synchronization and transmission technology

A variety of different display devices may be used when viewing content at a remote location, including large-scale displays and personal tablet devices. Content may also be delivered via a transmission network that includes various paths such as leased lines, the Internet, and wireless local area networks (LANs). If multiple video and audio signals are transmitted across these mixed environments, the playback timing of individual video and audio streams will vary due to differences in the propagation delay and processing latency on each route.

Our low-latency media synchronization and transmission technology achieves low-latency transmission while controlling the playback terminal buffer and using MMT-conformant synchronization control signals based on UTC (Coordinated Universal Time) to suppress the variations in the multiple video and audio sources. MMT also specifies the use of error correction coding to accommodate the loss of packets in transit. We implement error correction using a lightweight low-latency coding technique called LDGM (low density generator matrix) [1] developed by NTT Network Innovation Laboratories to enable stable synchronous playback even when close-up videos are transmitted to tablet devices via a wireless LAN.

The low-latency media synchronization processing flow is described below, taking the transmission of surround video as an example (**Fig. 5**). The surround video is transmitted after it has been partitioned into 4K units and encoded using a technique such as HEVC (High Efficiency Video Coding). During this process, the playback time, T_p, is transmitted to the display terminal as a UTC time stamp together with the video signal. Once the display terminal has received and decoded the video, it is stored in a buffer region until time T_p and then synchronously played back. Our low-latency media synchronization and transmission technology allows the buffer content to

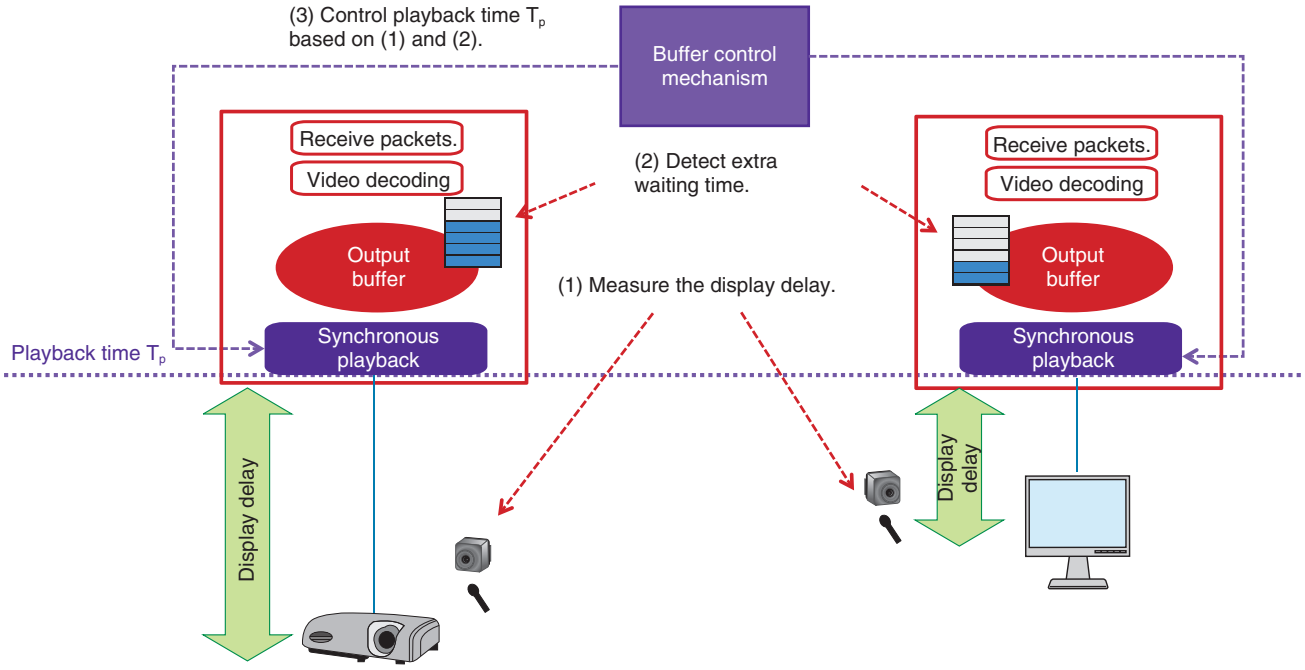


Fig. 6. Controlling the playback buffer region.

be controlled according to the playback environment so as to simultaneously achieve 1) guaranteed synchronization in different playback environments and 2) minimal video transmission delays (**Fig. 6**).

4.1 Guaranteed synchronization in different playback environments

In media transmission using MMT, synchronization of the playback timing can be guaranteed at the stage where video is output from the decoder, but the synchronization can still deviate due to differences in playback environments. This is because the time at which the video is actually presented is later than T_p due to differences in the display delays of different playback devices (projectors, monitors, tablets, etc.) and the delays of processes such as geometry correction in projectors, which are liable to vary. This can be fixed by measuring the delay in each playback environment and controlling a buffer to increase or decrease the delay so that the playback timing is correctly synchronized to T_p in each playback environment, thereby facilitating synchronized playback even in different playback environments.

4.2 Reduced latency of video transmission

The buffer that holds the decoded video until time T_p is needed for synchronization of video playback timings, but when this buffer contains more space than necessary, the video is delayed longer than necessary, resulting in additional latency. To avoid this, the amount of buffered video can be measured during playback, and by adjusting the buffer to eliminate excess capacity, we can achieve low-latency playback.

5. Future prospects

With the aim of implementing truly immersive public viewing services, NTT Service Evolution Laboratories is conducting research and development aimed at enhancing the real-time high-resolution wide-angle video stitching technologies we have developed so far and implementing functions with lighter system overheads and enhanced video processing to expand the application fields.

Reference

[1] T. Nakachi, T. Yamaguchi, Y. Tonomura, and T. Fujii, "Next-generation Media Transport MMT for 4K/8K Video Transmission," NTT Technical Review, Vol. 12, No. 5, 2014.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201405fa6.html

Takako Sato
Senior Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

She received a B.S. in mathematics from Hirosaki University, Aomori, in 1994. She joined NTT Multimedia Systems Department in 1994 and developed video transmission systems. She was also involved in the development of IP-based broadcasting systems. She moved to NTT Service Evolution Laboratories in 2012 and is developing new services based on R&D products.

Koji Namba
Senior Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in electronic engineering from Osaka University in 1999. He joined NTT in 1999 and developed digital rights management systems. From 2002 to 2015, he worked at NTT WEST, where he was involved in network engineering projects and the development of consumer services. He moved to NTT Service Evolution Laboratories in 2015, where he has been developing immersive telepresence technology.

Masato Ono
Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in information engineering from University of Tsukuba, Ibaraki, in 2006. He joined NTT EAST in 2006 and engaged in network engineering, customer service management, and creating business-to-business services. He moved to NTT Service Evolution Laboratories in 2016 and is developing new services based on immersive telepresence technology.

Yumi Kikuchi
Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

She received a B.E. and M.E. in electrical engineering from Tokyo University of Science in 2000 and 2002. Since joining NTT in 2002, she has been researching and developing content configuration methods and content delivery systems. Her present research is focused on feature point extraction of images and matching using the extracted feature points.

Tetsuya Yamaguchi
Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received a B.E., M.E., and Ph.D. in information engineering from Osaka University in 1997, 1999, and 2008. Since joining NTT in 1999, he has been researching and developing content distribution and navigation systems. His current interests are advanced media transport and ultra-realistic communications.

Akira Ono
Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in computer engineering from Waseda University, Tokyo, in 1992. He joined NTT in 1992 and engaged in research and development of video communication systems. From 1999 to 2010, he worked at NTT Communications, where he was involved in network engineering and creating consumer services. He moved to NTT Cyber Solution Laboratories (now, NTT Service Evolution Laboratories) in 2010. Since 2015, he has been studying the immersive telepresence technology called “Kirari!”.