Audio-visual Technology for Enhancing Sense of Presence in Watching Sports Events

Dan Mikami, Yutaka Kunita, Yutaka Kamamoto, Shinya Shimizu, Kenta Niwa, and Keisuke Kinoshita

Abstract

The number of people who experience sporting events remotely by viewing via television or the Internet far outnumber those who see the event in person. In recent years, there has been a diversification in individual preferences and viewing styles, and viewers want to enjoy the event in the way they prefer. This article introduces NTT's efforts concerning video and audio technology that enables high-sense-of-presence viewing. This work involves achieving both a high-presence remote viewing experience, which closely reproduces the actual on-site experience, and an ultrahigh sense of presence, which provides an experience that exceeds the on-site experience.

Keywords: High sense of presence, sports viewing, video and audio

1. Introduction

The dictionary defines *sense of presence* as, "the feeling that one is actually in a distant place." However, is that all that is needed to enjoy high-sense-of-presence sporting events? Sense of presence has two main aspects. One is the feeling that you are actually at that location, which we call *high sense of presence*. The other is the feeling that you see or know more than you would if you were actually at the location, which we call *ultrahigh sense of presence*. Sports events require both of these aspects. Many people want to have an experience in their own homes that is like being in the stands or even on the field of the event. There are also many viewers that want an even higher sense of presence that includes video that cannot be seen from the stands or conventional television (TV) such as video from the athlete's point of view, and sounds that cannot usually be heard such as talking among the players, which can provide an even richer experience than is available from the spectator's seat. This article describes elemental technology that NTT is working on to implement ultrahigh–sense-of-presence viewing.

2. Interactive distribution technology for omnidirectional video

In recent years, inexpensive head-mounted displays (HMDs) and cameras that can capture images that have a field of view of close to 360° have appeared on the market. Such devices have stimulated wider active interest in virtual reality viewing, which has previously been limited to some specialists and enthusiasts. NTT Media Intelligence Laboratories has been moving forward with research on interactive panorama distribution technology [1] that enables users to view in any direction they prefer. A specific application of this technology is interactive distribution technology for omnidirectional video viewing, which separates omnidirectional (360-degree) video into a number of regions and feeds the data to high-quality encoders. Then, high-quality video is selectively distributed according to the direction in which the user is looking (**Fig. 1**). Being able to view high-quality video for only the direction of viewing makes

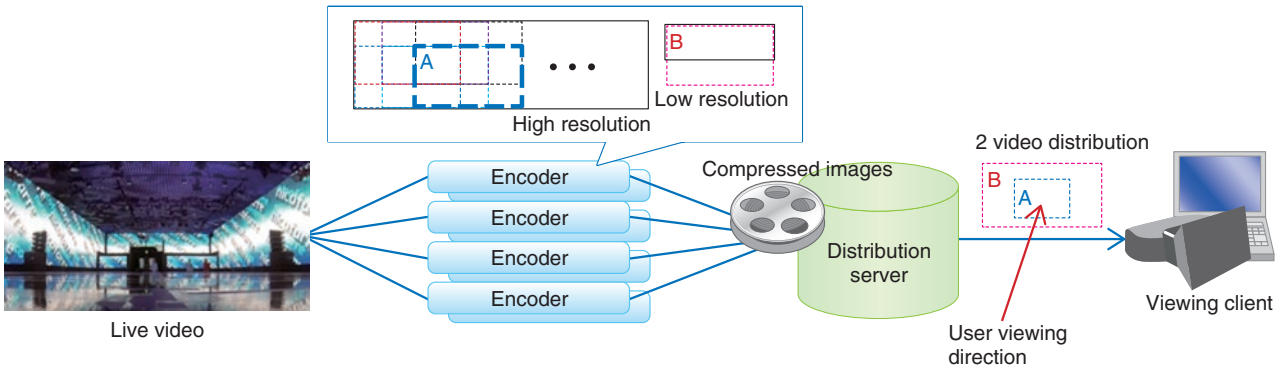


Fig. 1. Overall configuration of interactive distribution system for omnidirectional video.

it possible to deliver the video with less bandwidth than would be needed to deliver the omnidirectional video at high quality.

When selective distribution technology that is based on interactive panorama distribution technology is applied in omnidirectional distribution, the user's viewing experience varies greatly in terms of the features listed below.

(1) Wide field of view: Video that covers the entire field of view can give the user the sensation of being in a space (immersion). The visual field of humans has higher spatial resolution closer to the center and low spatial resolution at the periphery. That feature can be used to provide a high sense of presence with a limited bandwidth by transmitting and displaying video with high resolution only in the central area of the field of view.

(2) Head tracking: The HMD has sensors for acceleration and position that can be used to detect movement of the user's head. This makes it possible to present video to both eyes according to head motion, creating the feeling of looking around in a space. In contrast to viewing with a tablet or other conventional means, the user does not need to consciously select what part to view, so the viewing experience is more intuitive.

The space that can be experienced in current games and attractions at amusement parks is mostly produced by computer graphics. The technology we describe here, however, does not use computer graphics yet has been confirmed to deliver a sense of presence using live video from music performances and other events. We expect that applying this technology to the viewing of sports events can provide users with the exciting atmosphere of being in any spectator's seat in the stadium and convey the effect of being on the playing field.

3. Lossless audio encoding

Audio compression technology such as MP3 (MPEG*-1/2 Audio Layer 3) and AAC (Advanced Audio Coding), which are used in portable audio players and digital broadcasting, is widely used to provide audio at reasonable quality under the constraints of transmission bandwidth and memory capacity. To achieve a high sense of presence, however, transmission of sound with fidelity to the original source is needed. NTT has been participating in the standardization of MPEG–4 ALS (Audio Lossless Coding) and working to expand the use of lossless audio encoding [2].

Lossless coding makes it possible to completely reproduce the original sound waveform, even with compression, so audio data can be transmitted with no degradation of sound quality and with efficient use of network resources. A video and lossless audio system that we developed jointly with NTT Network Innovation Laboratories was used in trials of high-sense-of-presence live audio distribution conducted by NTT WEST and others. In those trials, users experienced a much stronger sense of being part of the scene than with conventional distribution methods, including joining in naturally with spectator applause and cheering [3, 4]. This increase in sound quality has also influenced the broadcasting of 4K and 8K video. In a survey of opinions on ultrahigh-definition TV conducted by the Ministry of Internal Affairs and Communications (MIC) of Japan in the spring of

* Moving Picture Experts Group

2014, nearly half of the responses concerned higher sound quality, and many of those were requests for use of lossless audio coding [5]. As a result, the MIC issued a Ministerial Ordinance for the capability of using MPEG-4 ALS in 4K/8K broadcasting, which was standardized by ARIB (Association of Radio Industries and Businesses) as ARIB STD-B32 in the summer of 2014.

We can thus see that there is a demand for higher sound quality to increase the sense of presence. In response to this demand, we have also moved forward with implementation of lossless audio coding for tablet terminals and set-top boxes and conducted verification testing for efficient use of the radio frequency band. In the future, wider use of lossless audio coding can be expected to improve the sense of presence for TV broadcasting and content distribution. Compression by lossless audio coding will also enable efficient transmission of the audio data acquired by the zoom microphone technology that is described later in this article, meaning that we are approaching the day when control of reverberation according to the listening environment will allow users to enjoy high-sense-of-presence content.

4. Distribution and encoding for arbitrary viewpoint video

Arbitrary viewpoint video allows the viewing of cuts from any position or orientation, regardless of the position or orientation of the camera that captured the scene. This technology is intended to provide a sense-of-presence video experience that is not possible with conventional video technology. That higher sense of presence is achieved by providing video from locations where ordinary cameras cannot be placed, such as the line of sight of players or the ball itself in a soccer match.

Arbitrary viewpoint video is created by using multiple-viewpoint video images taken simultaneously in different locations and orientations in a scene. The number of cameras necessary for taking the videos depends on the degrees of freedom of the viewpoints and the quality of the video to be created, but generally, many cameras are needed. However, video photography using many cameras, and the storage and transmission of the large volume of resulting video data are difficult. One method of creating arbitrary viewpoint video with less video data is to use depth mapping, which represents the distance of objects from the camera. We describe here our work on arbitrary viewpoint video using depth mapping together with video taken from multiple viewpoints.

Progress in sensor technology in recent years has made it possible to obtain depth maps directly by using depth cameras or rangefinders. However, the depth maps obtained in this way have low spatial resolution and contain a lot of noise. Therefore, the quality of arbitrary viewpoint video created with this technology is not high. To solve this problem, we have developed noise reduction processing and depth map up-sampling processing that uses the correlation between video and depth maps and the consistency of depth maps between viewpoints. Furthermore, we achieved real-time composition of arbitrary viewpoint video from the multi-viewpoint video and the depth maps obtained from the sensors by implementing the processing with a GPU (graphics processing unit).

Multiple viewpoint video and depth maps are a compact representation of arbitrary viewpoint video, but the amount of data is still huge compared to ordinary video. Therefore, efficient compression technology is essential for actual distribution of arbitrary viewpoint video. We previously developed a number of techniques for encoding arbitrary viewpoint video, including the use of viewpoint synthesis prediction and palette-based prediction. Viewpoint synthesis prediction is a technique applied in the synthesis of arbitrary viewpoint video to achieve efficient prediction between points of view by synthesizing predicted images using viewpoint video and depth maps that have already been encoded. Palette-based prediction is a method for generating predicted images by using the depth map feature, the value of which varies greatly between objects in the scene but varies little within a single object. In addition to achieving highly accurate prediction, this technique can also prevent degradation of performance in the synthesis of arbitrary viewpoint video by depth map encoding. These techniques that we have developed have been adopted in the 3D-HEVC (High Efficiency Video Coding) standard, which is an extension of the most recent HEVC international standard for video encoding [6].

In addition to the technology that we have described so far, implementation of arbitrary viewpoint video requires a lot of technology for elements ranging from imaging to the display and the user interface. With the current arbitrary viewpoint video synthesis technology using depth maps, the degree of freedom in moving the viewpoint and the quality of the synthesized image are limited. In the future, we plan to continue developing technology for realizing arbitrary viewpoint video that provides a video experience

Fig. 2. Zoom microphone system.

for larger scene spaces such as moving down into the playing field at sporting events.

5. Zoom microphone system

To improve the experience of sports events viewed via broadcasting or via telecommunications methods such as the Internet, we are developing technology for generating video that gives the viewer the feeling of being on the playing field. The zoom microphone system makes it possible to pick up distant sound sources clearly. It would be an elemental audio technology that is required for producing such video.

Our research started with a simple question: If a camera can zoom in on a target object, why can't we *zoom in* on distant sound sources in order to pick them up clearly? If distant sound sources were clearly recorded, it would be possible to provide audio that gives the viewer an impression of being on the playing field in the future.

The zoom microphone system consists of the two technologies shown in **Fig. 2**.

(1) Microphone array design for segregating sound sources from each other

We previously proposed a basic principle concerning how spatial signals should be captured with multiple microphones used to separate sound sources [7]. We defined the mutual information between sound sources and multiple microphones based on the information theory. To maximize that information, we placed microphones at optimum positions in front of parabolic reflectors. The implemented system shown in Fig. 2 includes 96 microphones and 12 parabolic reflectors.

(2) Noise suppression processing with less output degradation

We developed a signal processing algorithm for segregating sound sources arriving from target beamspace from other noise. With our algorithm, the output noise level would be reduced while maintaining the output signal quality. By utilizing phase/amplitude differences between microphones, a spectral filter that reduces the noise output power by as much as a factor of 1/10,000 can be generated [8]. So far, we have established principles for clearly picking up sound sources and have confirmed accurate estimation of sound sources at arbitrary locations 20 m away.

In the future, we will investigate the performance on an actual field and make technical improvements to reduce the number of microphones needed.

6. Reverberation removal and control

Hearing the cheers of the audience is an important element of sense of presence in viewing sporting events. Although being surrounded by cheering can greatly add to the sense of presence, suppression of that cheering may allow viewing that is more analytical. NTT has been working on reverberation removal and control technology, which plays an important role in controlling the sense of presence. The major application of this technology is in sound recording at concerts, so we describe it here in that context.

Recordings of dynamic and memorable performances and music from the past are available all around us in the form of compact discs (CDs), records, and other formats. We might wonder whether a sense of presence such as that obtained when hearing the music in the acoustic field of the site where it was recorded could be restored if those recordings could be played back in stereo, but that is not

necessarily the case. The reason is that it is difficult to reproduce the acoustic environment that existed at the time the performance was recorded when the recording is played back.

When we listen to music from a seat in a concert hall, two types of sound arrive at our ears from the stage. One is direct sound, which comes straight to our ears from the stage in front of us, and the other is reverberation, which comes to our ears indirectly as reflections from the walls and ceiling in four directions. Recordings on media such as CDs generally record a mixture of direct sound and reverberation that is picked up at a location near the audience seats. Thus, ordinary stereo recordings cannot reproduce the original acoustic environment that existed at the time of the recording.

We have developed the world's first technology for separating direct sound from the reverberation component in an audio signal, a technique we call *reverberation control*. We can enhance the sense of presence by applying this technique to separate the music signal into direct sound and reverberation components, and then create an acoustic environment that is similar to the acoustic environment at the time of recording by playing the direct sound component through the front speakers of a surround-sound system and playing the reverberation component through the front and rear/surround speakers [9]. Application of this technique to the past work of famous international artists and to consumer audio products has been well received thus far. In the future, we will continue our basic research with the objectives of applying it to broadcasting and achieving more accurate reverberation control processing.

7. Future development

We have described here some elemental technologies for achieving a high sense of presence in the viewing of sports events. Achieving high-sense-of-presence viewing of sports is a complex task that involves a variety of audio and video elements, including recording, encoding, distribution, processing, and the viewing system. In the future, we hope to continue creating viewing experiences that provide an even higher sense of presence in remote locations and at meeting sites by further integrating elements from the wide range of research done by NTT laboratories.

References

[1] Y. Tanaka and D. Ochi, "Interactive Distribution Technologies for 4K Live Video," NTT Technical Review, Vol. 12, No. 5, 2014. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201405fa5.html

[2] T. Moriya, N. Harada, and Y. Kamamoto, "Scope of Research on High-quality Audio Signal Processing and Coding,"
Y. Kamamoto, T. Moriya, N. Harada, and C. Kos, "Enhancement of MPEG-4 ALS Lossless Audio Coding,"
N. Harada, T. Moriya, and Y. Kamamoto, "MPEG-4 ALS: Performance, Applications, and Related Standardization Activities,"
in Selected Papers: Research Activities in Laboratories of New Fellows Part I, NTT Technical Review, Vol. 5, No. 12, 2007. https://www.ntt-review.jp/archive/2007/200712.html

[3] H. Yamane, A. Yamashita, K. Kamatani, M. Morisaki, T. Mitsunari, and A. Omoto, "High-presence Audio Live Distribution Trial," NTT Technical Review, Vol. 9, No. 10, 2011. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201110fa4.html

[4] Y. Kamamoto, N. Harada, T. Moriya, S. Kim, T. Yamaguchi, M. Ogawara, and T. Fujii, "Multichannel Audio Transmission over IP Network by MPEG-4 ALS and Audio Rate Oriented Adaptive Bit-rate Video Codec," NTT Technical Review, Vol. 11, No. 7, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307ra1.html

[5] MIC, Results of Appeal for Opinions on Draft Report of Broadcasting System Subcommittee (Technical requirements for ultra-high-definition television broadcasting system), Mar. 2014 (in Japanese). http://www.soumu.go.jp/main_content/000283104.pdf

[6] S. Shimizu, "Recent Progress of 3D Video Coding Standardization in JCT-3V," Journal of the Institute of Image Information and Television Engineers, Vol. 67, No. 7, pp. 557–561, 2013.

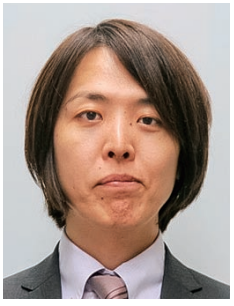
[7] K. Niwa, Y. Hioka, K. Furuya, and Y. Haneda, "Diffused Sensing for Sharp Directive Beamforming," IEEE Transactions on Audio, Speech and Language Processing, Vol. 21, No. 11, pp. 2346–2355, 2013.

[8] K. Niwa, Y. Hioka, and K. Kobayashi, "Post-filter Design for Speech Enhancement in Various Noisy Environments," Proc. of IWAENC 2014 (14th International Workshop on Acoustic Signal Enhancement), pp. 35–39, Juan-les-Pins, France, Sept. 2014.

[9] K. Kinoshita, "Enhancing Speech Quality and Music Experience with Reverberation Control Technology," NTT Technical Review, Vol. 12, No. 11, 2014. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201411fa3_s.html

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

He received the B.E and M.E in engineering from Keio University, Kanagawa, in 2000 and 2002, respectively, and the Ph.D. in engineering from University of Tsukuba, Ibaraki, in 2012. He joined NTT in 2002. His current research activities are mainly focused on multimedia content handling. He was awarded the Meeting on Image Recognition and Understanding 2009 Excellent Paper Award, the IEICE Best Paper Award 2010, the Institute of Electronics, Information and Communication Engineers (IEICE) KIYASU-Zen'iti Award 2010, and the IPSJ SIG (Special Interest Group)-CDS (Consumer Devices & Systems) Excellent Paper Award 2013. He is a member of IEICE, the Information Processing Society of Japan (IPSJ), and the Institute of Electrical and Electronics Engineers (IEEE).



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

He received the B.E., M.E., and Ph.D. in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. He joined NTT Cyber Space Laboratories in 2001 (now, NTT Media Intelligence Laboratories) and has been studying high-reality visual communication. He is a member of the Association for Computing Machinery, IEICE, and the Virtual Reality Society of Japan.

Yutaka Kamamoto
Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received the B.S. in applied physics and physico-informatics from Keio University, Kanagawa, in 2003 and the M.S. and Ph.D. in information physics and computing from the University of Tokyo in 2005 and 2012, respectively. Since joining NTT Communication Science Laboratories in 2005, he has been studying signal processing and information theory, particularly lossless coding of time-domain signals. He was also with NTT Network Innovation Laboratories, where he developed the audio-visual codec for ODS (other digital stuff / online digital sources) from 2009 to 2011. He has contributed to the standardization of coding schemes for MPEG-4 ALS, ITU-T (International Telecommunication Union, Telecommunication Standardization Sector) Recommendation G.711.0, and 3GPP (3rd Generation Partnership Project) Enhanced Voice Services (EVS). He received the Telecom System Student Award from the Telecommunications Advancement Foundation (TAF) in 2006, the IPSJ Best Paper Award from IPSJ in 2006, the Telecom System Encouragement Award from TAF in 2007, and the Awaya Prize Young Researcher's Award from the Acoustical Society of Japan (ASJ) in 2011. He is a member of IPSJ, ASJ, IEICE, and IEEE.

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

He received the B.Eng. and M.Info. in social informatics from Kyoto University and the Ph.D. in electrical engineering from Nagoya University, Aichi, in 2002, 2004, and 2012, respectively. He joined NTT in 2004 and has been involved in research and development of 3D video coding algorithms and standardization in MPEG and ITU-T. His research interests include signal processing for free viewpoint video, light fields, computer vision, and computational photography. He is a member of IEICE.

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

He received the B.E., M.E., and Ph.D. in information science from Nagoya University, Aichi, in 2006, 2008, and 2014, respectively. Since joining NTT in 2008, he has been engaged in research on microphone array signal processing. He was awarded the Awaya Prize from ASJ in 2010. He is a member of IEEE, ASJ, and IEICE.

Keisuke Kinoshita
Senior Research Engineer, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the M.Eng. and Ph.D. from Sophia University, Tokyo, in 2003 and 2010, respectively. Since joining NTT in 2003, he has been researching various types of speech, audio, and music signal processing, including speech enhancement such as 1ch/multi-channel blind dereverberation, noise reduction, distributed microphone array processing, and robust speech recognition. He received the 2006 IEICE Paper Award, the 2009 ASJ Outstanding Technical Development Prize, the 2011 ASJ Awaya Prize, and the 2012 Japan Audio Society Award.