ITUPublications

International Telecommunication Union
Radiocommunication Sector

Report ITU-R BT.2420-9

(03/2026)

BT Series: Broadcasting service (television)

Collection of usage scenarios of advanced immersive sensory media systems

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from https://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at https://www.itu.int/publ/R-REP/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	**Broadcasting service (television)**
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radio-wave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management
TF	Time signals and frequency standards emissions

***Note**: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2026

© ITU 2026

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

REPORT ITU-R BT.2420-9

Collection of usage scenarios of advanced immersive[1] sensory media systems

(Question ITU-R 143/6)

(2018-2020-03/2021-11/2021-03/2022-09/2022-2024-2025-2026)

TABLE OF CONTENTS

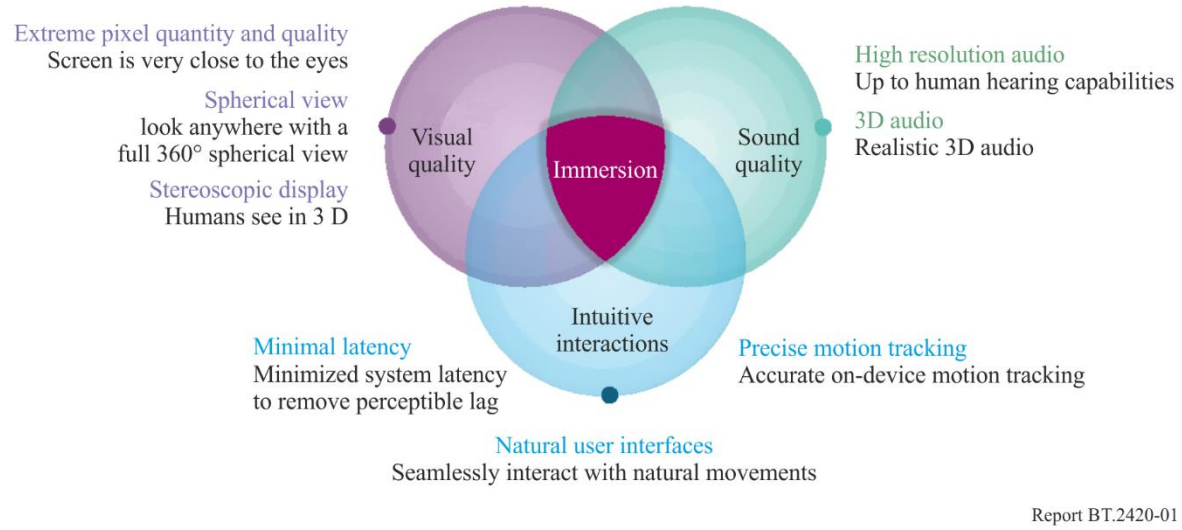
[1] The term ‘immersive’ in the context of this Report is deemed to include any format or medium or platform that offers or engages an audience by employing sensory based technologies such as audio, video, or haptic) and enables any form of interaction or control of the content presentation.

1 Introduction

Advanced immersive sensory media (AISM) systems allow a user to have immersive experiences with an unprecedented degree of presence including the advanced immersive audio-visual systems. By tricking the perceptual systems of the user's brain, AISM systems can make the user believe to be somewhere else and/or somebody else. This is achieved by (re)creating audio-visual realities and allowing the user to naturally interact with these virtual environments. Figure 1 depicts the three primary quality aspects in AISM systems that contribute to immersion and presence. The sense of immersion breaks down if the information presented to these modalities does not work properly together. In some instances, users may even experience sensory sickness (see below).

FIGURE 1

The three primary quality aspects in AISM systems



This ITU-R Report is intended to describe a brief technical background and important definitions used for AISM systems, use cases for broadcasting of AISM programme material, and other challenges that have emerged through those production trials.

2 Technical background and glossary

2.1 Overview

This section provides a brief overview of technical terms and concepts. For more detailed information, the interested reader is invited to study the following guides and primers [1] [2] [3] [4].

2.2 Forms of AISM Systems

Virtual Reality (VR): A technology that replicates an environment, real or imagined, and simulates a user's physical presence and environment to allow for user interaction. Virtual reality artificially creates a sensory experience, which in principle can include sight, touch, hearing, and smell. The current VR devices primarily present content to the visual and auditory systems. On occasion, haptics information is also included.

Augmented Reality (AR): The addition of images or enhanced digital content overlaying the physical world. This can be introduced in the visual field or another sense such as audition and sound. More developed applications lead to a fusion of the physical and virtual worlds into one reality which can be experienced via an HMD as defined in § 2.3. Augmented Reality can be experienced via an HMD or a plain screen. Examples are Microsoft's HoloLens1 and 2, Magic Leap 1, Bose Frames, Google Glass, Pokémon Go and Yelp Monocle.

2.3 Presentation modes

HMD: A head mounted display (HMD) is a display worn on the body that fits over a user's head. It has small display optics in front of the eyes and is usually equipped with additional sensors to track the viewer's head motions such as coordinate positions, pitch, roll, and yaw. In some instances, the position of the user's gaze is also captured. HMDs for AR allow the user to passively view the contextual physical world (e.g. Microsoft HoloLens), whereas HMDs for VR occlude perception of the contextual physical environment. For VR applications, some HMDs enable event-driven user integration of mobile devices to act as system displays and/or processors for the device.

Magic window: This presentation mode enables exploration of 360° video content to be accessible without use of an HMD on: mobile devices, desktop computers, televisions or large theatrical screens.

Using a mouse pointer, gamepad or finger gesture, the user drags and rotates the 360° image on the screen to see a portion of the 360° video scene. Depending on the resolution of the 360° video, some applications may also allow the user to zoom into the scene. Further, the motion and position sensors on mobile devices allow a user to steer the mobile device in a desired direction to see a select region of the 360° video through the screen. In all cases, the accompanying sounds and acoustic scene should adapt accordingly. This mode of content presentation does not provide full immersion, but it does enable an extended mode of content interaction and consumption that has low risk for sensory sickness.

Second screen: An AISM second screen could offer specific VR vantage points that accompany the regular television programme. These experiences do not necessarily have to replace the current television viewing paradigm, but, rather, may complement TV programmes by offering second screen services synchronized to the TV broadcasting.

Sound reproduction device: Sound is reproduced using headphones, soundbars or loudspeakers. Headphones that support head-related transfer-function (HRTF) processing allow users to perceive and interact with sounds in all directions, aligning with HMD-based viewing. Soundbars can simulate a space that users can navigate virtually. Playback using three-dimensional loudspeaker configurations can be paired with Magic-window viewing on desktop computers, televisions or large theatrical screens. Advanced sound system renderers can be used where the centre of loudspeaker configuration is the reference position. However, if an immersive 6DoF audio renderer, (such as the MPEG-I immersive audio renderer), supports "tracked loudspeaker rendering", then the fully immersive user experience is not restricted to the centre of the loudspeaker configuration but is extended to a very wide area inside the setup (with minimum distance of 0.5 m to the closest loudspeaker).

Haptic interface: A haptic interface is a device that allows a user to interact with a system by receiving tactile feedback. Through the haptic device, the user receives feedback in the form of haptic stimuli, which include vibrations and changes in pressure and temperature. Haptic devices are expected to serve as interfaces for enhancing viewer immersion in broadcast programmes by presenting viewers with haptic stimuli linked with the broadcast content.

2.4 Modes of interactivity

Three Degrees of Freedom (3DoF): Programme material in which the user can freely look around in any direction (yaw, pitch, and roll). A typical use case is a user sitting in a chair looking at 3D VR/360° content on an HMD as shown in Fig. 2.

Three Degrees of Freedom Plus (3DoF+): Programme material in which the user is free to look in any direction (yaw, pitch, and roll), plus limited translation movements due to the head movements not being centred on the optical and acoustical centre. This provides support for perceptual effects such as motion parallax which strengthen the sense of immersion. A typical use case is a user sitting in a chair looking at 3D VR/360° content on an HMD with the capability to move his head slightly up/down, left/right, and forward/backward as shown in Fig. 2.

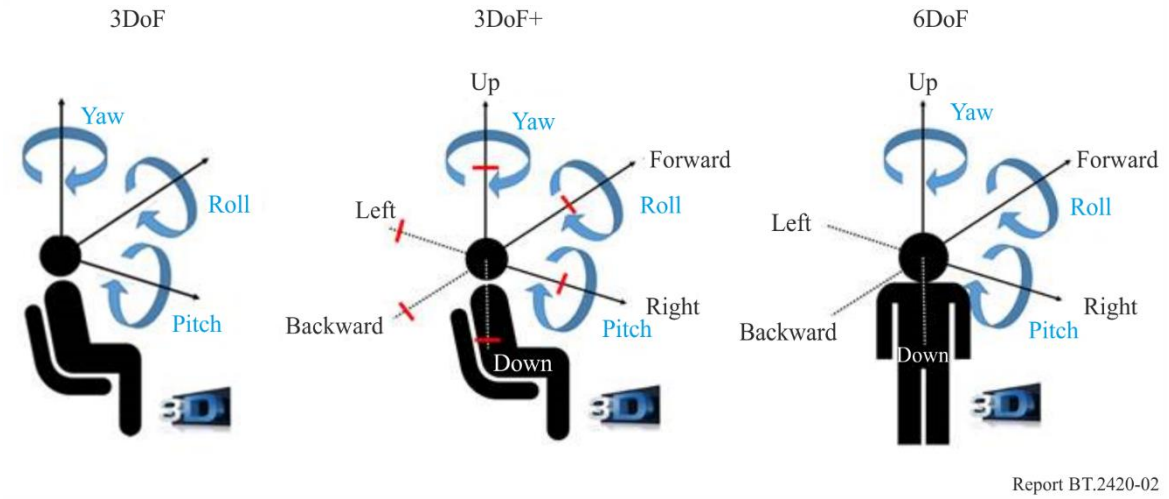
Multiple vantage points: While the definitions for 3DoF and 3DoF+ are based around a single point of observation, these concepts can be extended to a scenario where users may experience a scene from multiple discrete vantage points.

Six Degrees of Freedom (6DoF): Programme material in which the user can freely navigate in a physical space. The self-motion is captured by sensors or an input controller. Both rotation (yaw, pitch, and roll) and translation (x, y, z translation) interactions are possible. A typical use case is a user freely walking through 3D VR/360° content (physically or via dedicated user input means) displayed on an HMD as shown in Fig. 2.

It is currently expected that the majority of VR experiences to be deployed in the near term are going to be 3DoF. Mass market consumer devices and services supporting 6DoF can be expected to be widely available by 2020 (see [5], [6]).

FIGURE 2

Freedom of movement in a three-dimensional space



Note to Fig. 2: This Figure is taken from ISO/IEC JTC1/SC29/WG11 N17264 − Working Draft 0.4 of Technical Report on Architectures for Immersive Media.

Diegetic and non-diegetic elements: A diegetic audio or video element is a stable element in the virtual scene. It is spatially rendered to be dependent on the position and movements of the user's head as recorded by the HMD sensors. The perceived position a diegetic element is unaffected by head motion. In contrast, a non-diegetic audio or video element is spatially rendered independent of the virtual scene and changes with movements of the user's head. The perceived position of the element is not constant and updates as the HMD receives user-driven information updates from the device sensors. An example of a diegetic audio element could be a person talking in the background

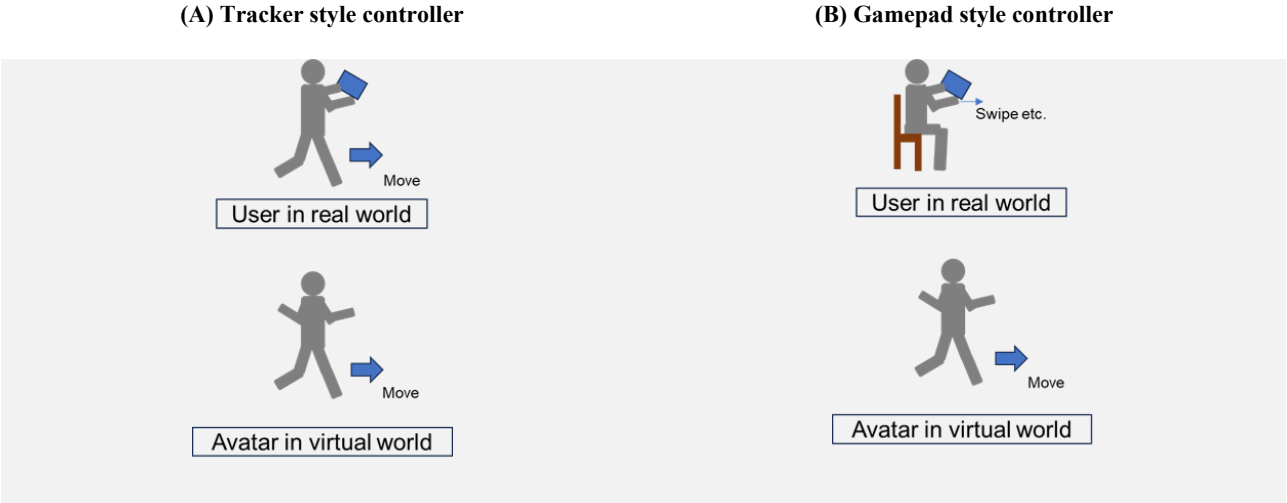
that is not present in the virtual scene. An example of a non-diegetic video element could be a graphical overlay, such as end credits.

Linear narrative vs. non-linear interactive content: A Linear narrative programme moves sequentially through time and does not allow a user to modify how the content sequence is presented in time. Linear programme material is the primary content type used in broadcasting. In contrast, non-linear narrative programmes allow a user to interact with the content. This allows a user to modify how the content sequence appears in time. Non-linear programme material is common in the gaming community. Example non-linear AISM programme experiences may enable a user to walk anywhere within the scene. The user's motion and behavioural interactions with the content will directly influence how the programme material appears in time. Because it is not clear how such non-linear interactive content could be delivered via broadcast, this Report will primarily focus on linear narrative content.

Controller for user's position and rotation in the virtual space: User's position and rotation in the 6DoF contents can be controlled by at least two types of controllers. Tracker-like controller can directly connect a user's position and rotation to a virtual camera or an avatar. Then, a user moves through their real space as well as a virtual camera or an avatar (see Fig. 3A). A user may operate position and rotation of a virtual camera or an avatar using Gamepad-like controller. Then, a user usually sits on a chair or sofa such as if watching a TV (Fig. 3B).

FIGURE 3

Control devices for 6DoF content



3 ITU-R related use cases

3.1 Overview

This section lists a number of use cases that are related to broadcast systems. Use cases less relevant to broadcasting (i.e. unicast) are studied elsewhere, e.g. at 3GPP (see [3, section 5]).

3.2 The infinite seat VR broadcast

Infinite Seat VR Broadcast is a method for capture of live-VR content. At the event capture site, omnidirectional camera and microphone rigs can be placed at certain seating/viewing locations. Each audio and video capture rig delivers a distinct experience, corresponding to the unique seating locations at the event site. The infinite seating experience can be further augmented with additional

audio elements (e.g. commentator voice) or visual elements (e.g. player statistics) and with techniques such as action replay, or cuts from different vantage points.

The viewer might be able to select between different seats individually, while the orientation at each seat is defined by the user's rotation (yaw, pitch, and roll).

3.3 Linear narrative cinematic VR broadcast

During Linear Narrative Cinematic VR Broadcast, a consumer watches a VR movie from a fixed point in the scene (3DoF or 3DoF+). The consumer can freely turn and move his head to observe details in the scene and may follow the story by listening and watching the actors.

3.4 Free viewpoint television

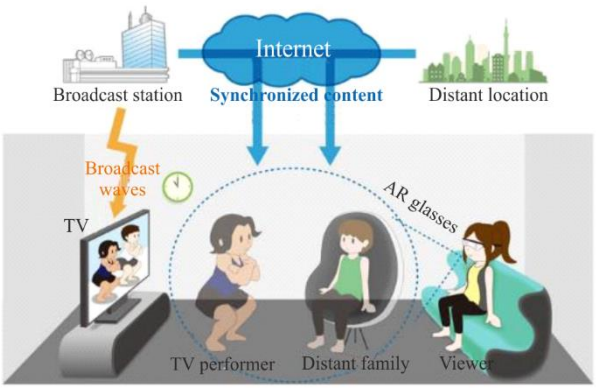
Free Viewpoint Television is a visual media that allows users to view a three-dimensionally recorded scene by freely changing their point of observation. By changing the vantage point of the broadcasted scene, the reproduced sound scene and audio elements will adapt accordingly. Free viewpoint images may also be presented simultaneously with a normal television programme by delivering additional 3D objects linked to the TV programme in real-time through broadband Internet.

3.5 Integration of TV and AR

AR has the potential to enhance the experience of TV viewing. One new viewing style is 'virtual space sharing', where viewers simultaneously experience AR of six degrees of freedom (6DoF) on their handheld devices or glasses equipped with AR technology while watching a TV programme on television [7]. Figure 4 illustrates such a concept of the integration of TV and AR, where three-dimensional objects of TV performers or family members and friends in different locations are combined and displayed through the AR device. Viewers can feel as though they are sharing the same space as TV performers, family, and friends while watching TV. With AR glasses, performers or persons are displayed in their actual size to provide an increased sense of reality.

FIGURE 4

Concept of the integration of TV and AR



Report BT.2420-03

3.6 Haptic AR broadcast

The term 'haptic' refers to the sense of touch. While broadcasting is traditionally a medium for conveying mainly visual and/or audio information to the audience, the addition of haptic information would enable the provision of broadcast services with new sensory experiences that could greatly enhance qualitative experiences. In a sporting event, for example, vibrating a haptic device when a

ball is in motion or bouncing would provide viewers with an enhanced immersive experience as if they were participating in the event rather than simply spectating. A haptic interface could also help visually or hearing-impaired members to intuitively understand the broadcast content.

3.7 HMD-based content consumption

3.7.1 Private VR television

Private VR Television presents conventional 2D television programme material to a viewer in a virtual environment. The 2D television programme material is rendered directly in front of the user on a virtual rectilinear screen within the virtual environment. The virtual scene may adapt the audio and video to the user's head motion. In a different mode, the virtual screen may be head-locked, meaning that the screen is always displayed in front of the user independent of their head motions.

3.7.2 Enriched VR television

During an Enriched VR television viewing experience, the television programme is rendered on a virtual 2D-rectilinear screen inside the 3D-virtual environment. A 360-degree scene covers the background of the spherical environment. The user can activate the contextual menu for displaying additional information, e.g. sport statistics related to the current game, additional movie information, the electronic programme guide, and a selection of different viewing angles. Also, a stereo camera can be attached to the HMD that captures the scene and depth around the viewer enabling reproduction of a mixture of real and virtual images on the HMD. In this experience, the real scene near the viewer (including the viewer's hands and body) is displayed to the user along with presentation of the virtual world (virtually represented environment and elements) seamlessly according to depth.

3.8 Supporting visually or hearing-impaired audience members

VR/AR technology could enable an improved broadcast content experience for visually or hearing-impaired audience members. A recent article [8] demonstrated that VR/AR glasses may be used to help the visually impaired recognize objects. Haptic information is also useful for visually or hearing-impaired users to understand the content. In this way, it may be possible to use VR/AR devices to provide visually or hearing-impaired consumers with specific enhancements to broadcast content that enables a markedly improved programmed experience.

3.9 Collective viewing of VR content

A key attraction of VR content is that it offers a highly personalised experience within a virtual space. However, delivering such experiences typically requires individual viewing and listening devices, such as HMDs and headphones. Consequently, promoting VR content at exhibitions or public events faces challenges in providing sufficient opportunities for large audiences, which often diminishes the overall promotional impact. To address this issue, an alternative method involves presenting VR content using a large screen with a multichannel sound system, allowing many people to experience the content simultaneously (see Fig. 5). For this purpose, VR content that supports 6DoF should be used because it is designed to enable personalised and immersive experiences.

Two presentation approaches can be considered for collective viewing:

– Live navigation: A single participant acts as the representative user, who navigates the VR content in 6DoF. On the large screen with its multichannel sound system, the other participants simultaneously follow the representative's experiences.

– Pre-recorded navigation: A 6DoF session performed by a user is pre-recorded. The recorded performance is then played back on the large screen with its multichannel sound system, allowing the participants to relive the session together.

Although these methods do not provide a true 6DoF interaction tied to each participant's physical movements, an essential aspect of the VR experience, they still offer a valuable opportunity to engage in a pseudo-interactive form that can help broaden the audience and increase their familiarity with VR experiences.

FIGURE 5

Collective viewing of VR content using a large screen with a multichannel sound system

4 Broadcaster VR productions and trials

4.1 Overview

There is significant interest in using both advanced audio and video technologies for VR production and programme applications from programme creators, broadcasters, and media consumers. Many TV broadcasters are undertaking production trials to get familiar with the production workflow and to evaluate the feasibility of VR programme production and delivery. A representation of these broadcasters includes BBC, Sky, ZDF, Arte, Canal+, RAI, NBC, CBS, DirecTV, Telemundo, Turner Sports, Eurosport, NHK, Nippon TV, TV Asahi, Fuji TV, and TBS-TV. The DVB authored a report referenced in § 6, that includes an overview of these engagements with technical details of 32 production trials [2, section 6]. Both experienced Video FX companies, as well as new specialized startups with proprietary VR equipment and production tools, are collaborating with broadcasters in VR productions (e.g. NextVR, Jaunt, Within, Felix & Paul). Other agencies are specializing in creation of advertisements for VR (e.g. OmniVirt, Advrtas, VirtualSKY). A recent case study on 360 degrees advertisement [9] suggests that the 360-degree ad formats can trigger 85% engagement on mobile and 33% engagement on desktop respectively.

For distribution, many broadcasters have developed their own mobile device VR applications. Typically, these apps developed by broadcaster do not support content streaming. Consequently, it is necessary for users to download the VR content in its entirety prior to viewing. Recently, some broadcasters began to provide their own on-demand VR distribution channels (e.g. ABC News VR, Discovery VR). These VR distribution channels are hosted on the content platforms of HMD providers (Oculus, Viveport, Samsung VR, PlayStation VR), at streaming services (YouTube 360, Facebook 360, Twitter, dailymotion, Vimeo), or accessible via websites using WebVR, Flash, or other HTML5 extensions.

4.2 Content types in VR/AR trials

4.2.1 Examples content type of VR/AR trials

This section provides a selection of resources of typical broadcast content produced for VR/AR consumption. Most of these examples have a duration of less than 15 minutes.

In addition, media types outside the broadcast sector, such as: newspapers, comic books, radio, and musicals are using VR and AR to engage user beyond the initial content consumption endpoint or production to provide users with additional footage (such as behind-the-scene reports), bonus material, or marketing content (e.g. The New York Times, The New Yorker, The Guardian, USA Today, Huffington Post, National Geographic, Madefire Comics, CBC Radio, Dali Museum, School of Rock musical).

4.2.2 Sport and sport highlights

– **Summer Olympics**
BBC: http://www.bbc.com/sport/36883859

– **Basketball**
http://www.recode.net/2016/10/20/13341408/nba-virtual-reality-games-nextvr

– **American Football**
BTN: http://btn.com/2016/11/10/btn-to-become-first-college-sports-network-to-produce-live-football-game-in-virtual-reality
https://www.cnet.com/news/nfl-nextvr-highlights-virtual-reality-super-bowl/

– **Hockey**
http://venturebeat.com/2015/02/26/nhl-streams-a-hockey-game-in-360-degree-virtual-reality
https://www.nhl.com/news/nhl-introduces-virtual-reality-experiences/c-279085566

– **Boxing**
Fox: http://fortune.com/2016/01/21/fox-sports-nextvr-team-on-boxing
DirecTV: http://variety.com/2015/digital/news/directvs-first-virtual-reality-app-takes-boxing-fans-ringside-1201613503

– **Golf**
Fox: http://www.sportsvideo.org/2016/06/14/fox-sports-nextvr-drive-virtual-reality-experience-at-u-s-open

– **Tennis**
France Television: http://advanced-television.com/2016/05/20/france-televisions-airs-roland-garros-in-4k-vr
https://www.prolificnorth.co.uk/2016/07/laduma-films-wimbledon-in-360-degrees

– **Racing**
NBC, Horse Racing: http://www.sportsvideo.org/2016/05/06/live-virtual-reality-hits-the-track-with-nbcs-first-ever-kentucky-derby-vr-production/
Fox, Car Racing: http://fortune.com/2016/02/18/fox-sports-daytona-500-virtual-reality

– **Extreme sport**
Red Bull: https://www.redbull.com/gb-en/events/red-bull-air-race-vr-experience

– **E-sport**
https://www.pastemagazine.com/articles/2016/11/e-sports-in-vr.html

4.2.3 News

ABC News VR: http://abcnews.go.com/US/fullpage/abc-news-vr-virtual-reality-news-stories-33768357

The Big Picture – News In VR: https://www.youtube.com/watch?v=C5qbR5SQleY

The Economist: http://visualise.com/case-study/economist-vr-app

4.2.4 Documentaries

BBC: http://www.bbc.co.uk/taster/projects/invisible-italy

Doctors Without Borders: http://visualise.com/case-study/msf-doctors-without-borders-forced-home

4.2.5 Television shows

NBC, Saturday Night Live: https://www.youtube.com/watch?v=6HS9h4xFRww

ABC, Dancing with the stars: http://abc.go.com/shows/dancing-with-the-stars/news/updates/vr-05022016

Competitive Cooking shows: https://www.youtube.com/watch?v=JpAdLz3iDPE

4.2.6 TV series, episodic

http://www.hollywoodreporter.com/behind-screen/virtual-reality-tests-episodic-story-940425

http://variety.com/2016/digital/news/hulu-ryot-virtual-reality-news-comedy-show-1201866110

4.2.7 Animation

Spotlight Stories: http://www.polygon.com/2017/1/24/14370892/virtual-reality-first-oscar-nominated-short-film-pearl

BBC: http://www.bbc.co.uk/taster/projects/turning-forest

4.2.8 Music videos

Reeps One: https://www.youtube.com/watch?v=OMLgliKYqaI

Muse: https://www.youtube.com/watch?v=91fQTXrSRZE

The Who: http://www.billboard.com/articles/6312181/the-who-new-app-greatest-hits-virtual-reality

4.2.9 Concert experiences

Kasabian: http://visualise.com/case-study/kasabian-o2

4.2.10 Special event content

NBC, US presidential debates: http://fortune.com/2016/09/21/presidential-debate-virtual-reality

BBC, London New Year's Eve Fireworks: http://www.bbc.co.uk/taster/projects/new-years-eve-fireworks-360

4.2.11 Features films or promo teaser

BBC, Planet Earth II: http://www.bbc.co.uk/programmes/articles/365zWpz7HypS4MxYmd0sS36/planet-earth-ii-in-360

ZDF, TEMPEL: http://visualise.com/case-study/360-trailer-zdfs-tv-series-tempel

HBO, Game of Thrones interactive VR experience: https://www.framestore.com/work/defend-wall

20th Century Fox, Wild: http://www.roadtovr.com/ces-2015-fox-debut-wild-vr-360-movie-experience-starring-reese-witherspoon

Sony, Ghostbusters: http://www.theverge.com/2016/6/29/12060066/ghostbusters-dimension-the-void-times-square-madame-tussauds-vr

4.3 VR trials

4.3.1 VR in collaboration with TV programmes and events

NHK conducted VR and AR projects including "NHK VR NEWS", "Panorama Tour", "8K VR Theatre", "Augmented TV", and others in collaboration with existing TV programmes and events.

VR content is distributed on the NHK website.

NHK VR × AR): https://www.nhk.or.jp/vr/

In June 2017, NHK broadcast a programme titled "BS1 Special – Real Trump World: The World That Created the New President Explored with a 360° Camera". It was filmed using a 360° camera, and broadcast as a programme for a normal-sized screen. The 360° video was delivered simultaneously over the Internet in sync with the broadcast. It was the world's first experiment allowing the viewer to freely look at images 'outside' the frame that were not visible on the television screen by moving their mobile device up or down and to the left or right. The question arose as to whether viewers would find 360° video delivered in sync with television broadcasts appealing.

Public broadcaster efforts to engage with VR like the "Trump World" broadcast have only just begun and many issues remain. There is an open question regarding what public broadcasters should do when it comes to basic telecommunications services looking toward the 2020 Tokyo Olympics, Paralympics, and beyond. VR will offer an important perspective in keeping with the aspirations to become a public service media. The research report entitled "The Meaning of VR: Delivering 360 Degree Videos by Public Service Broadcasters Towards 2020 and Beyond" is available at http://www.nhk.or.jp/bunken/english/reports/pdf/report_17121201.pdf.

4.3.2 360-degree VR image system for comparative views

4.3.2.1 Overview

Situations in disaster areas are reported by distributing 360-degree images from the scene via network services such as NHK VR News. It is informative for viewers to be able to see not only how the sites are damaged but also how they are being reconstructed. NHK has developed a VR system that can show comparative views captured right after a disaster and captured several months after the disaster at the same position by displaying both images side-by-side [10]. The user can watch any direction of the 360-degree scene and change the border of the two images horizontally by a user interaction.

4.3.2.2 Capturing 360-degree images

In 2018, 360-degree images of areas severely damaged by a typhoon were captured at the quasi-same position in three cities in western Japan at two different times, initially captured immediately after

the disaster and a second captured several months later (Fig. 6). Insta360 Pro was used to capture 360-degree images with the resolution of 3 840 × 1 920 pixels for each eye.

FIGURE 6

Capturing 360° images at a disaster site

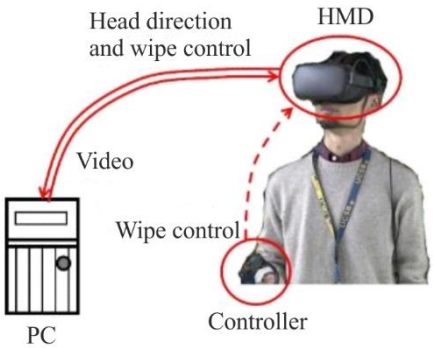
Report BT.2420-04

4.3.2.3 Displaying comparative views

The display system for comparative views consists of a workstation, an HMD, and a controller, as shown in Fig. 7. Captured images are rendered by the workstation and fed to an HMD. The centre of a sphere is set to the position of the camera in a virtual space to render the scenes. As shown in Fig. 8, two 360-degree images, one captured right after the disaster and the other several months later, were mapped to the sphere. A moveable white border that the user can control between the two images indicates where the image transitions between the past and more current scenes. Fig. 9 shows images of comparative views displayed on the HMD. The user is free to watch any direction of the 360-degree scene.

FIGURE 7

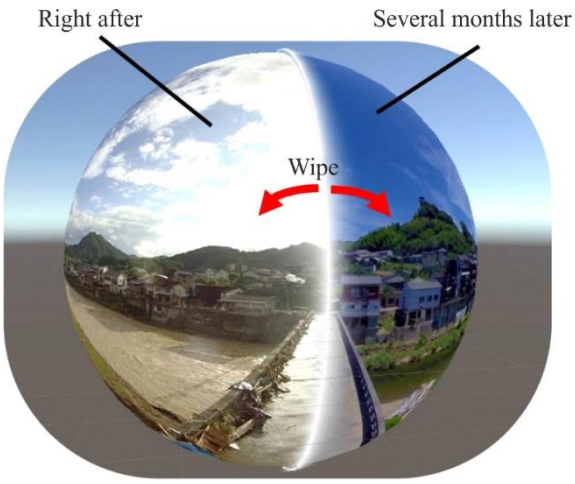
Hardware configuration



Report BT.2420-05

FIGURE 8

Rendering a scene



Report BT.2420-06

FIGURE 9

Images of comparative views displayed on an HMD

Report BT.2420-07

4.3.3 360° VR image system with 8K display

4.3.3.1 Overview

There are two major quality-related problems to address in 360° VR imaging to make viewing high-quality VR images comfortable; 1) the resolution of 360° images, and 2) the resolution of head mounted displays (HMDs). To prevent users from perceiving the pixel structure of a display when viewing part of a 360° image, future HMDs must include a wider field of view than currently available as well as having significantly higher spatial resolution. To support this, it is necessary for 360° images to have a much higher spatial resolution the typically captured. To cope with these problems,

a prototype HMD using an OLED panel with a spatial resolution of 8K × 4K was developed, and a 360° image with a spatial resolution of 30K × 15K was captured.

4.3.3.2 HMD for VR with 8K resolution

An OLED panel with a spatial resolution of 8K × 4K was used for the HMD. Table 1 shows the specifications of the display.

TABLE 1

Specifications of 8K OLED panel used for HMD

Screen size	8.33-inch diagonal
Spatial resolution	7 680 × 4 320 for R, G, and B each
Pixel pitch	1 058 ppi (24 μm)
Frame frequency	60 Hz
Developer	Semiconductor Energy Laboratory Co., Ltd.

The HMD consisted of the OLED panel, optical components, a motion sensor unit, and image processing circuits. The motion sensor unit consisted of a 3-degree acceleration sensor, a 3-degree angular speed sensor, and a 3-degree geomagnetic sensor, and could detect the viewing direction of a user three-dimensionally in real time.

The dimensions of the panel were 103.68 × 184.32 mm, and the size of both the left and right images was 103.68 × 92.16 mm. Designing optical components so that the field of view is about 100 degrees is ideal with an 8K display, and this is achieved with a focal length of 38.67 mm. Lenses with a focal length of 50 mm were used due to easy availability. This resulted in a decreased field of view of 85.33 degrees.

It is ideal that the distance between the centres of the right and left lenses correspond to the average pupillary distance, which is about 65 mm. However, the size of the OLED panel was a little bit larger. To best use the 8K panel, the optics were designed by using an optical beam shift.

Moreover, as shown in Fig. 10, the prototyped HMD was not a goggles-type one but a hand-held one mounted to a support arm.

FIGURE 10

HMD for 8K VR

Report BT.2420-08

4.3.3.3 Capturing 360-degree spherical images

360-degree spherical images with a significantly high resolution were obtained from multiple sub-images captured by a still camera with a 5 472 × 3 648 resolution by using a robotic camera mount for automated panorama shooting. A total of 144 sub-images (12 × 12) was captured for a 360° spherical image except for the area at the foot of the camera mount. The 144 images were then stitched into a rectangular spherical image of 55 184 × 21 524 (55K × 22K) pixels by using the equirectangular projection (ERP)format. The ERP format is likely to be adopted for the MPEG-I Part 2: Omnidirectional Media Format (OMAF) for representing 360-degree video on a 2D plane.

The 55K × 22K image was scaled down to 30 720×15 360 (30K × 15K) pixels, which is sufficient to represent the 360-degree sphere, and a black bar was inserted at the bottom as shown in Fig. 11.

FIGURE 11

Stitched spherical image with 30 720 × 15 360 pixels

Captured at Tsukuba Space Center, Japan Aerospace Exploration Agency (JAXA)

Report BT.2420-09

4.3.3.4 Presenting 8K × 4K VR images

To present 8K × 4K VR images on the HMD, the rectangular spherical image was first re-mapped to a dome-shaped spherical image. The direction in which the user is facing was detected every 10 ms by the motion sensor attached to the HMD. In accordance with the direction, the corresponding area, which was at a size of 3 840 × 4 320 pixels (4K × 4K), was clipped from the re-mapped spherical image. The clipped image was then corrected to compensate for lens distortion and displayed side-by-side on the HMD.

4.3.4 Capture/display system for VR images with resolution beyond 8K

4.3.4.1 Overview

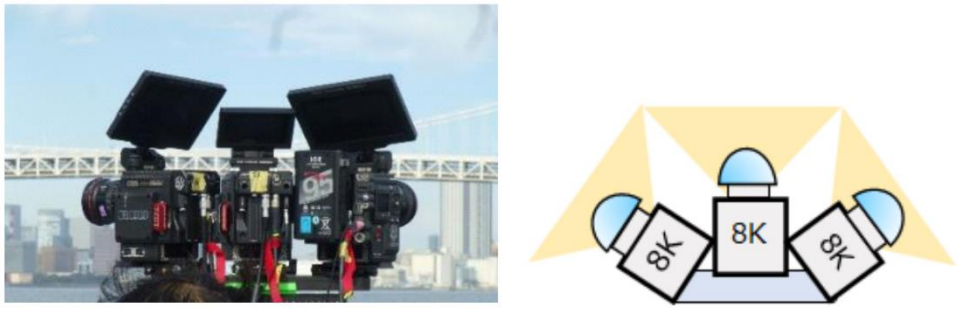
Virtual Reality with image resolutions beyond 8K is expected to provide highly immersive experience with a sense of presence and reality. NHK set up a display system that projects images of over-8K resolution to a large cylindrical screen that provides a horizontal field of view of approximately 180° by using eight 4K projectors [11].

4.3.4.2 Capture system

A camera array consisting of three 8K cameras was set up to capture VR images covering a 180-degree field of view, as shown in Fig. 12 and Table 2. The three 8K cameras were aligned radially to capture synchronized images. VR images with the equirectangular format were produced through post-production including stitching and grading, as shown in Fig. 13.

FIGURE 12

Camera array consisting of three 8K cameras



Report BT.2420-10

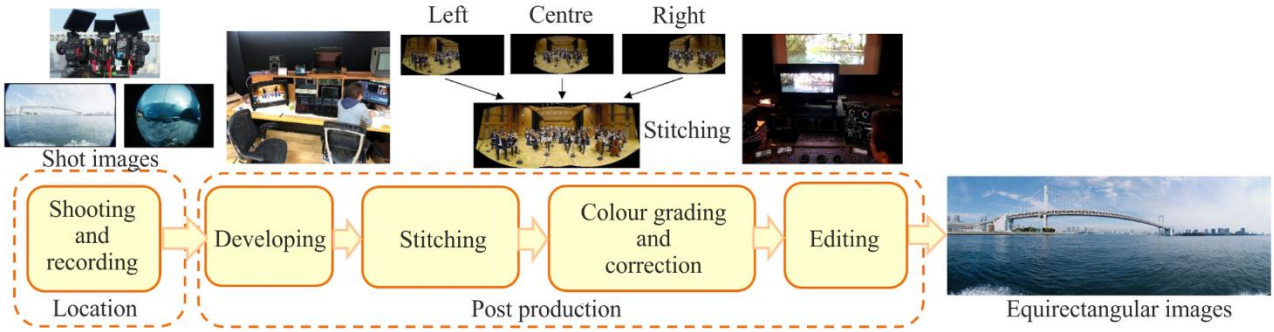
TABLE 2

Specifications of the camera array

Camera	RED / 8K MONSTRO × 3 units
Spatial resolution	8 192 × 4 320 (each camera)
Frame frequency	59.94 Hz

FIGURE 13

Workflow of high-resolution VR



Report BT.2420-11

4.3.4.3 Display system

A projector array consisting of eight 4K laser projectors was constructed to display the over-8K resolution images on a cylindrical screen, as shown in Fig. 14 and Table 3. Each projector was fixed in a portrait orientation. The array was placed 4 m above the floor to minimize the shadow of viewers on the screen. The diameter and the height of the screen were 11 m and 4 m, respectively. The bottom part of the screen was a round shape to widen the vertical field and minimize image distortion due to changes of a viewing position. The playout system geometrically converted the equirectangular images for each projector using the 3D shape model of the screen to display cylindrical views.

FIGURE 14

Display system

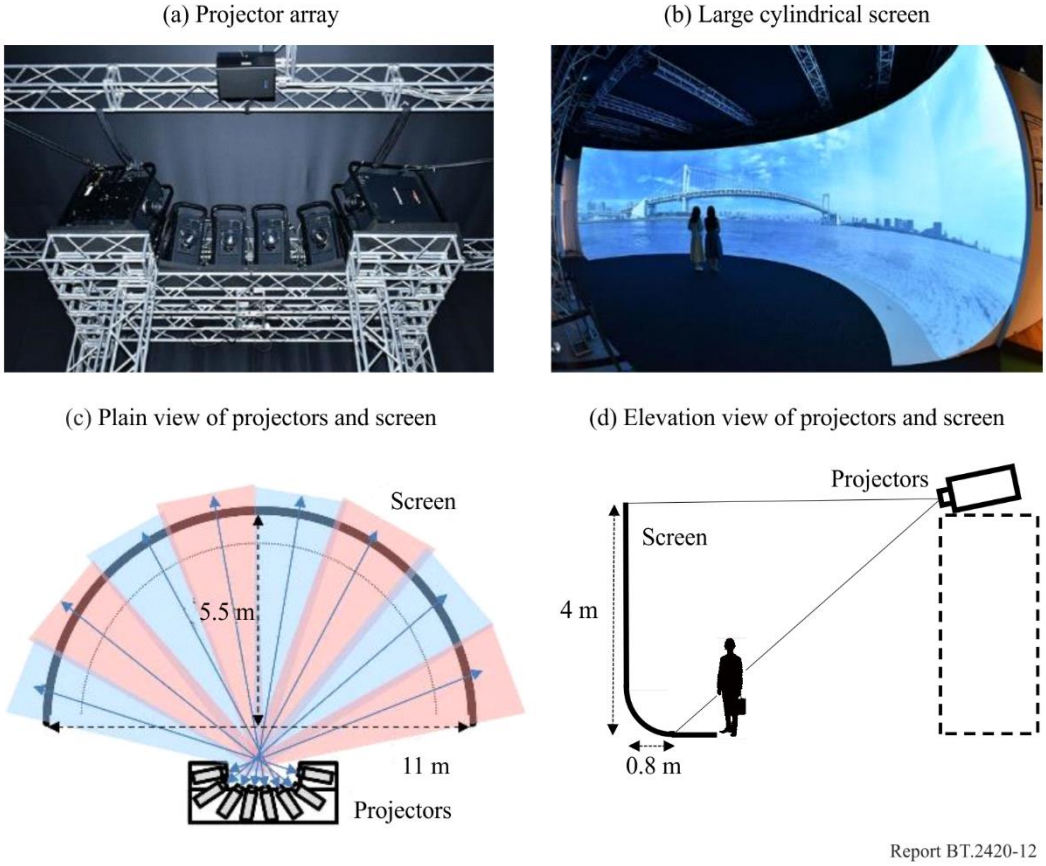


TABLE 3

Specifications of the display system

Projector array	4K laser projectors × 8 (Panasonic / PT-RQ13KJ)
Spatial resolution	3 840 × 2 160 (each projector)
	About 12K × 4K (displayed images)
Frame frequency	59.94 Hz
Screen	About 180° horizontal field of view Cylindrical (diameter: 11 m, height: 4 m) 19-face polyhedron

4.3.5 Immersive content to be displayed on a large flat screen

Nippon TV produced 9 984 × 2 160-pixel (10K × 2K) content of a prestigious Japanese temple garden through all four seasons displayed on a large flat screen at an exhibition in Tokyo National Museum.

4.3.5.1 Capture system

The images were captured with four arrayed 4K cameras covering a 180-degree field of view, as shown in Fig. 15 and Table 4. Panoramic VR images were produced through post-production including colour balancing between cameras, stitching, cropping, colour grading, and editing. Figure 16 shows the production workflow and image composition.

FIGURE 15

Camera array consisting of four 4K cameras

Report BT.2420-13

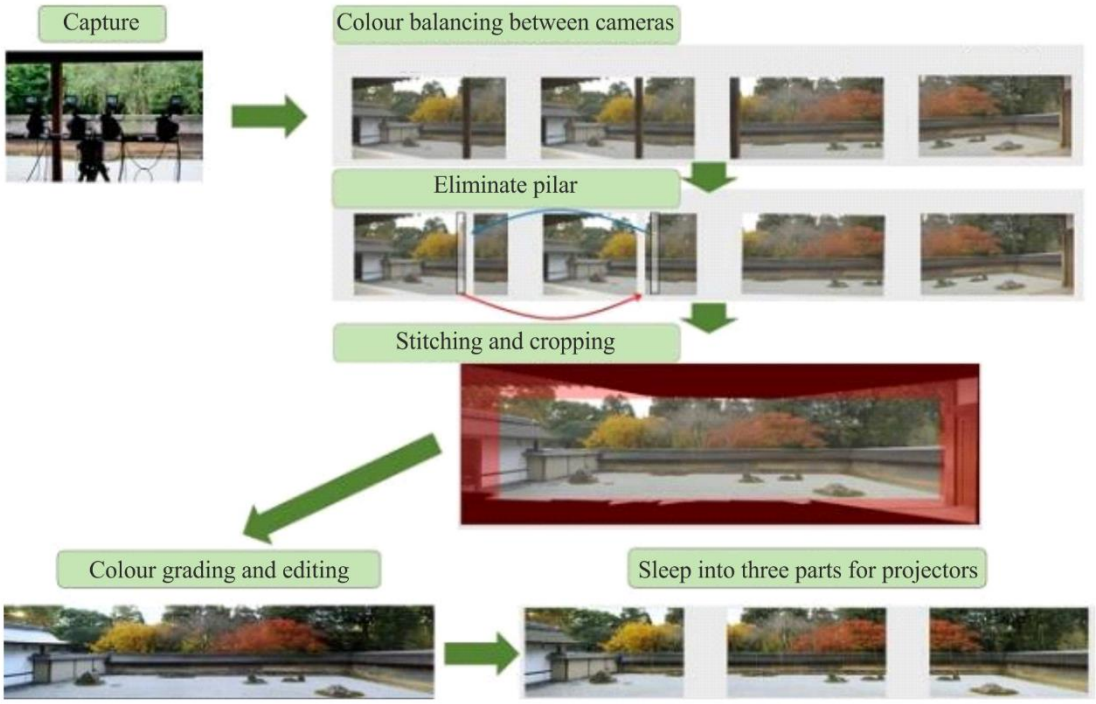
TABLE 4

Specifications of camera array

Camera	Canon Cinema EOS C500 EF × 4 units
Lens	Canon CN-E24mm T1.5LF
Spatial resolution	4 096 × 2 160 (each camera)
Frame frequency	29.97 Hz (progressive)

FIGURE 16

Production workflow and image composition



Report BT.2420-14

4.3.5.2 Display system

Three 4K projectors aligned horizontally were used to display the panoramic VR images on a large flat landscape screen of 15.6 × 3.4 m together with 5.1 channel surround sound, as shown in Fig. 17 and Table 5. Each projector was placed in a landscape orientation. Optical blending and signal-level blending were conducted to make the luminance of overlapping projection areas on the screen uniform.

FIGURE 17

Display system

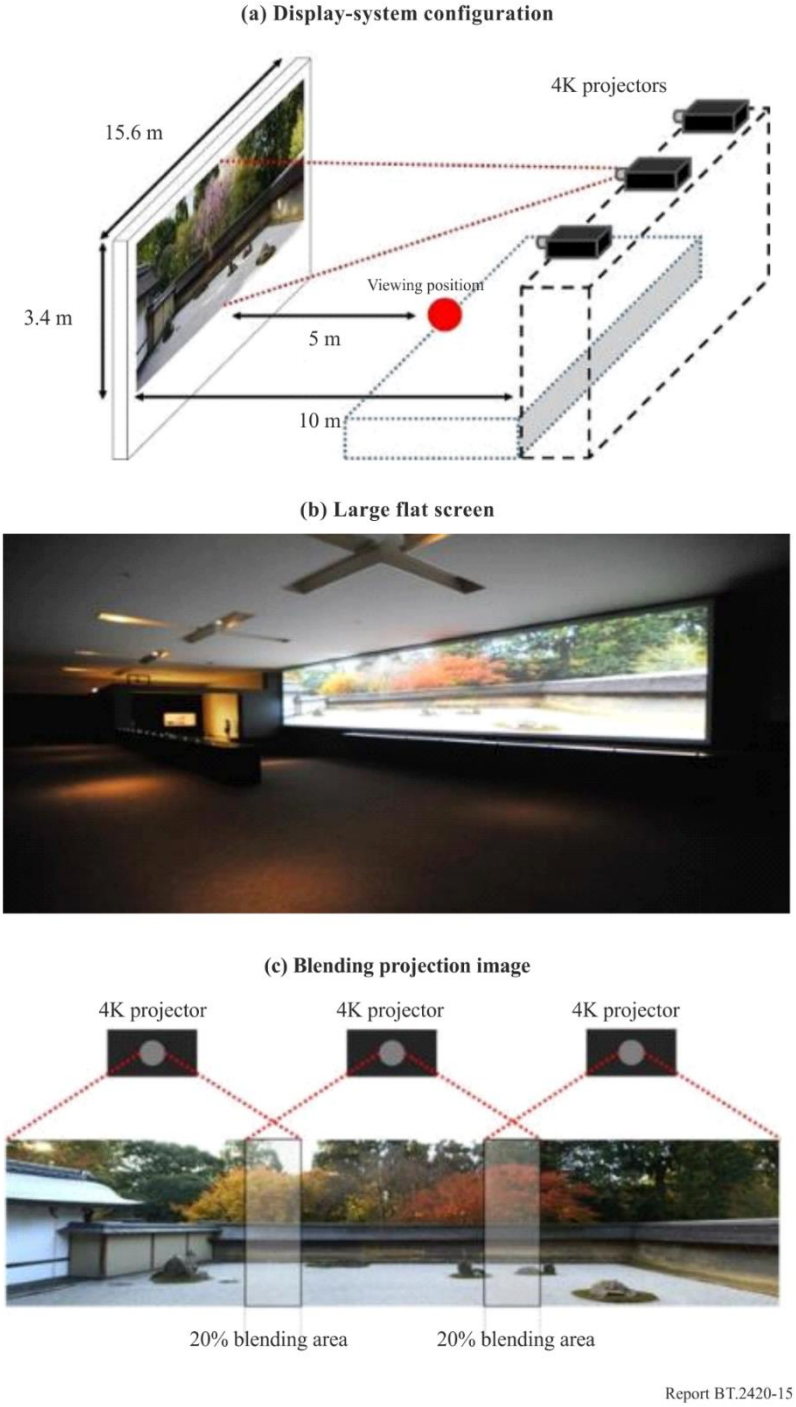


TABLE 5

Specifications of display system

Projector	4K projectors × 3 (JVCKENWOOD/DLA-SH7NL)
Spatial resolution	4 096 × 2 160 (each projector)
	About 10K × 2K (displayed images)
Frame frequency	29.97 Hz
Screen	Flat white screen (width: 15.6 m, height: 3.4 m)

4.3.6 Immersive VR display

NHK developed immersive curved VR displays (see Table 6 and Fig. 18) which surround the viewer's head. Since the video covers virtually the viewer's entire field of view, the immersive feeling for individual viewers is greatly enhanced. The curved displays are a thin, lightweight, and bendable with organic light emitting diodes (OLEDs) that were fabricated on a thin film substrate.

To produce a highly immersive personal viewing display, three 30-inch 4K flexible OLED displays are installed in a continuous curved shape. This provides a total horizontal resolution of 6K and vertical resolution of 4K at a pixel pitch of 0.173 mm, which makes the pixel structure barely perceptible at a viewing distance of around 37 cm, and covers a field of view of approximately 180°around the viewer's head. By combining it with a chair that vibrates to match the video and audio (chair haptic device, see § 4.4.7), it is possible to further increase the sensation of realism.

TABLE 6

Specifications of immersive VR display

Screen size	51.3-inch diagonal
Spatial resolution	6 480 × 3 340 for R, G, and B each
Pixel pitch	0.173 mm
Radius	Approximately 37 cm

FIGURE 18

Immersive VR displays and haptic chair device

Report BT.2420-16

4.3.7 30K 360-degree video capture system and 15K hemispherical display system

4.3.7.1 Overview

To provide an unprecedented immersive media experience, NHK has developed a 360-degree video capture system and a hemispherical display system. The video format adopted is compliant with Recommendation ITU-R BT.2123.

4.3.7.2 30K 360-degree video capture system

To develop a 30K 360-degree video capture system compliant with Recommendation ITU-R BT.2123, the system adopts a pentagonal prism configuration. Industrial cameras (9 344 × 7 000 pixels, with a 3.2 μm pixel pitch) are used as the element cameras.

Figure 19 shows the view-angle design of the pentagonal prism configuration. As shown in Fig. 19 (left), the basic configuration comprises seven cameras in total, five horizontally arranged cameras, one zenith camera, and one nadir camera. Figure 19 (right) illustrates the expected coverage using the 9K × 7K cameras. In the equatorial direction, five horizontal cameras provide 6,144 pixels × 5 = 30 720 pixels. In the meridional direction, the system covers 6 508 pixels × 2 at the top and bottom and 8 852 pixels × 2 vertically, giving a total of 30 720 pixels. In the actual implementation, the downward view camera was omitted.

The developed video capture system, therefore, comprises one zenith camera and five horizontal cameras. The pixel densities are 73.25 pixels/degree for the zenith camera, and 95.59 and 85.33 pixels/degree vertically and horizontally, respectively, for the horizontal cameras. These densities are comparable to the standard value of 76.8 pixels/degree for 8K resolution (7 680 pixels over a 100-degree field of view). Custom horizontal and zenith lenses were fabricated and evaluated, confirming that the modulation transfer function (MTF) modulation remained above 0.2 within the field of view.

To reduce stitching errors in the creation of 360-degree images, the distance between the principal points of the lenses was minimised. A folded-lens design was adopted for the horizontal cameras for this purpose, achieving a significant reduction in principal point distance. Figure 20 shows the folded lenses and zenith lens. The effective field-of-view (FOV) of the horizontal camera lens is 103.75-degree with an F-number of 4.5. The zenith lens was designed to be as thin and elongated as

possible while maintaining the required optical path length, achieving an effective FOV of 99.4-degree with an F-number of 4.5.

Figure 21 shows the assembled camera head and its camera rig. A dedicated rig was developed to assemble the camera head, allowing the horizontal cameras to be positioned at a tilt angle of 26.5 degrees so that they can be placed more closely than in standard configurations. The rig also enables horizontal cameras to be adjusted along the X- and Y-axes, while the zenith cameras is vertically adjusted. This rig design reduced the lens principal-point distance to only 6 cm.

Figure 22 and Table 7 show the overall configuration of the video capture system, which consists of the camera head, a transmitter, and a recorder. The recorder contains six workstations. The six element cameras and the workstations are connected via a 100 GbE optical interface. Each workstation is equipped with a redundant array of independent disks (RAID) comprising four U.2 solid-state drives (SSDs) to support high-speed data recording. In addition, a fan-out cable was developed to connect the cameras to the workstations, and a transmission system was implemented to carry signals over a single optical multicable up to 150 m in length.

FIGURE 19

View-angle design of pentagonal prism configuration

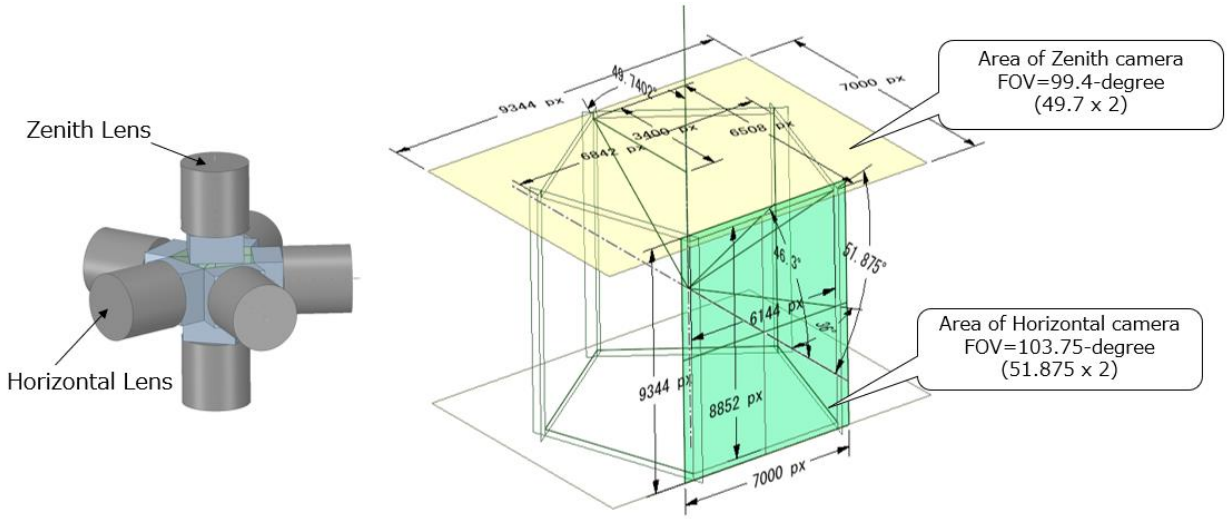


FIGURE 20

Folded lens for horizontal and zenith lens

FIGURE 21

Assembled camera head with camera rig (left: aerial view, right: top view)

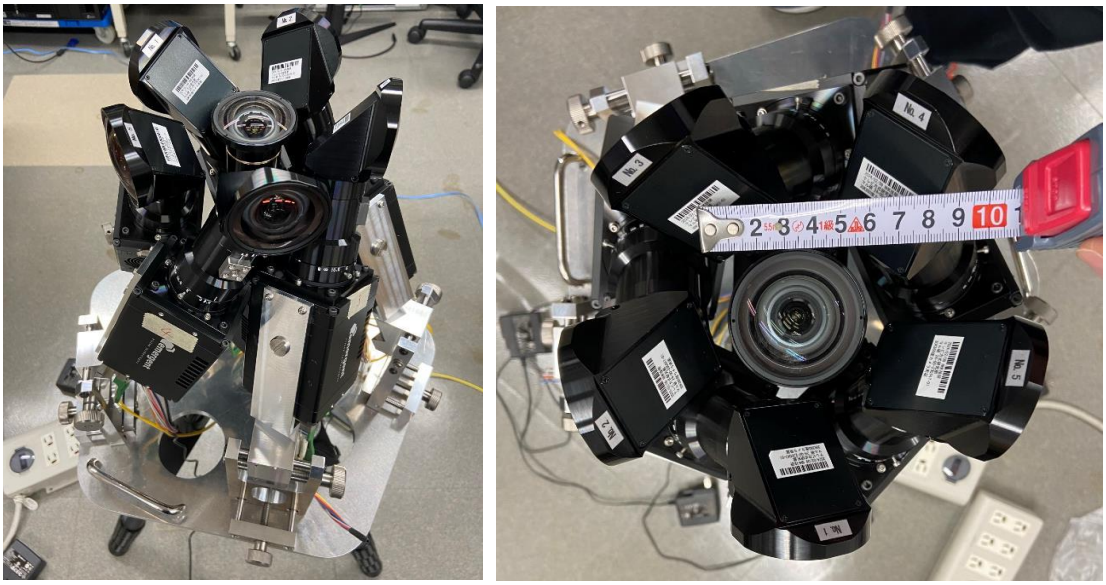


FIGURE 22

Configuration of the video capture system

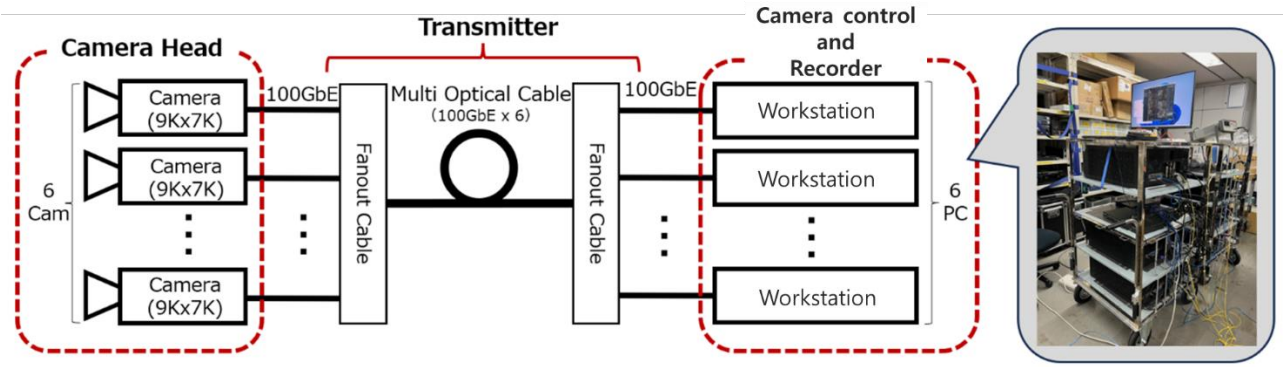


TABLE 7

Specification of the 30K video capture system

Number of pixels	30 720 (horizontal)
FOV	360-degree (No downward angle of view)
Modulation of Lens MTF	Over 0.2 at 156.25 LP/mm
Frame rate	60 Hz
Bit depth	10 bits
Distance between optical principal points	Approx. 6 cm
Camera head size	429 mm (W) × 413 mm (D) × 364 mm (H)
Camera head weight	Approx. 14 kg

4.3.7.3 15K hemispherical display system

Figure 23 shows the appearance of the 15K hemispherical display, and Fig. 24 shows a schematic diagram of the display system. The display comprises a hemispherical screen with a diameter of 3 m and a 15K-equivalent projector installed above the screen (Fig. 25). The horizontal and vertical viewing angles at the centre position are 180 degrees and 100 degrees, respectively.

A 15K-equivalent projector with a native resolution of 7 680 × 4 320 pixels displays twice its resolution (15 360 × 8 640 pixels) using a wobbling method that shifts the projected image diagonally by half a pixel in each frame. Although the vertical resolution is limited, it corresponds to a 100-degree vertical field of view, aligning with the FOV of the hemispherical screen.

Table 8 lists the projector specifications. The projector is equipped with three 1.21-inch reflective LCD panels (each with 7 680 × 4 320 pixels) for red, green, and blue channels. It receives two phase-shifted 8K videos sampled from 15K videos. The displayed image is shifted diagonally by half a pixel every 1/120 s, achieving a 15K-equivalent resolution for each 1/60 second frame.

The output light from the projector is reflected at an angle by a mirror and directed into a specially developed fisheye projection lens. Figure 26 shows the fisheye lens designed to project video across the entire hemispherical screen. This configuration prevents viewer shadows from being reflected on screen, even when viewers stand near the centre. The modulation at the centre of the projection lens is 286 lp/mm, indicating that the modulation is not null with respect to the number of 15K pixels. The projector produces 4,500 lumens, and the hemispherical screen (gain = 1.0) achieves a maximum luminance of approximately 100 cd/m2.

A signal processing chain consisting of a signal processor and geometric correction units was developed to deliver video signals to the projector. Due to the extremely large data volume of 30K 360-degree video, real-time processing is not currently feasible. Therefore, only the front 180 degrees (15 360 × 15 360 pixels) is provided to the signal processor. The input video is divided into eight streams (each 7 680 × 3 840 pixels) and transmitted via eight HDMI 2.1 interfaces. The signal processor performs two functions. First, it extracts a 15 360 × 8 640-pixels region from the 15 360 × 15 360 pixels input. Second, it generates two 8K streams (7 680 × 4 320 pixels) with diagonal phase shifts for use in the projector's "wobbling method." Each 8K video is then sent to a geometric correction unit, where geometric distortions are corrected prior to projection. Two units process each 8K stream in real time and synchronise with one another. The corrected signals are transmitted via eight DisplayPort interfaces to the projector.

The system also supports the input of 15K 360-degree video. Using this capability, viewers can interactively select any portion of the 360-degree scene for display by providing instructions to the geometric correction units.

FIGURE 23

Hemispherical display

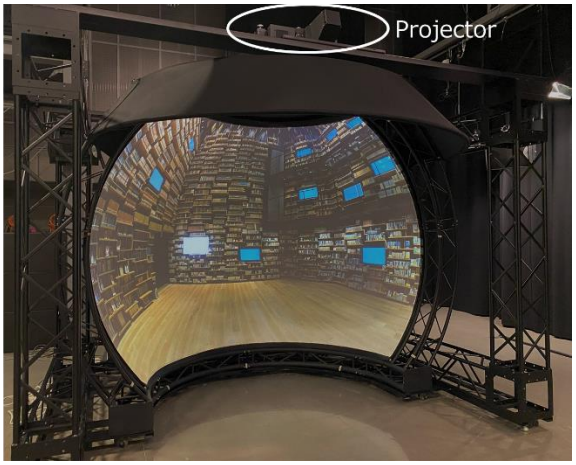


FIGURE 24

Schematic diagram of signal processing

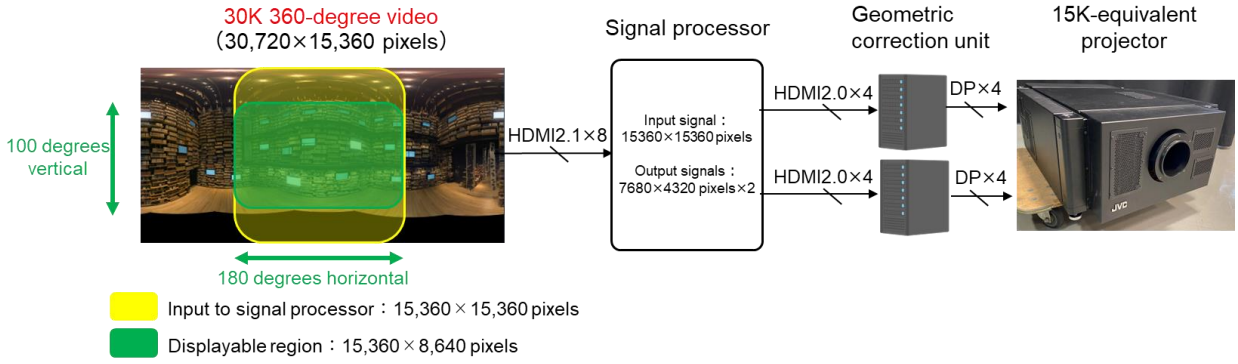


FIGURE 25

15K-equivalent projector

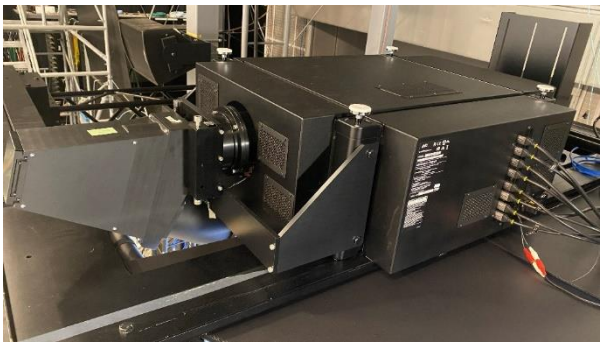


TABLE 8

Specifications of 15K-equivalent projector

Method	Shifting the projected position diagonally by half a pixel in each frame
Device	Three 1.21-inch reflective LCD panels (7 680×4 320 pixels)
Frame rate	59.94 / 60 Hz
Bit depth	12s bit
Light output	4 500 lumen
Size	747 mm (W) × 1017 mm (D) × 350 mm (H)
Weight	86 kg

FIGURE 26

Projection fisheye lens

4.3.7.4 Content production and demonstration

Location shootings were conducted using the 30K 360-degree capture system at four sites in the Tokyo area: from a sightseeing bus, from a cruise ship, at an autumn foliage location in Hakone, and in an indoor environment surrounded by bookshelves (Fig. 27). A 3.5-minutes video was produced to fully exploit the 30K 360-degree resolution, incorporating scenes captured from moving vehicles as well as indoor scenes requiring fine detail representation.

This video was demonstrated using the 15K hemispherical display system as shown in Fig. 28. When standing at the centre of the display, viewers' fields of view are almost completely filled with high-resolution imagery, demonstrating that the system enables an immersive viewing experience without the need for special viewing devices.

FIGURE 27

30K 360-degree content location shooting

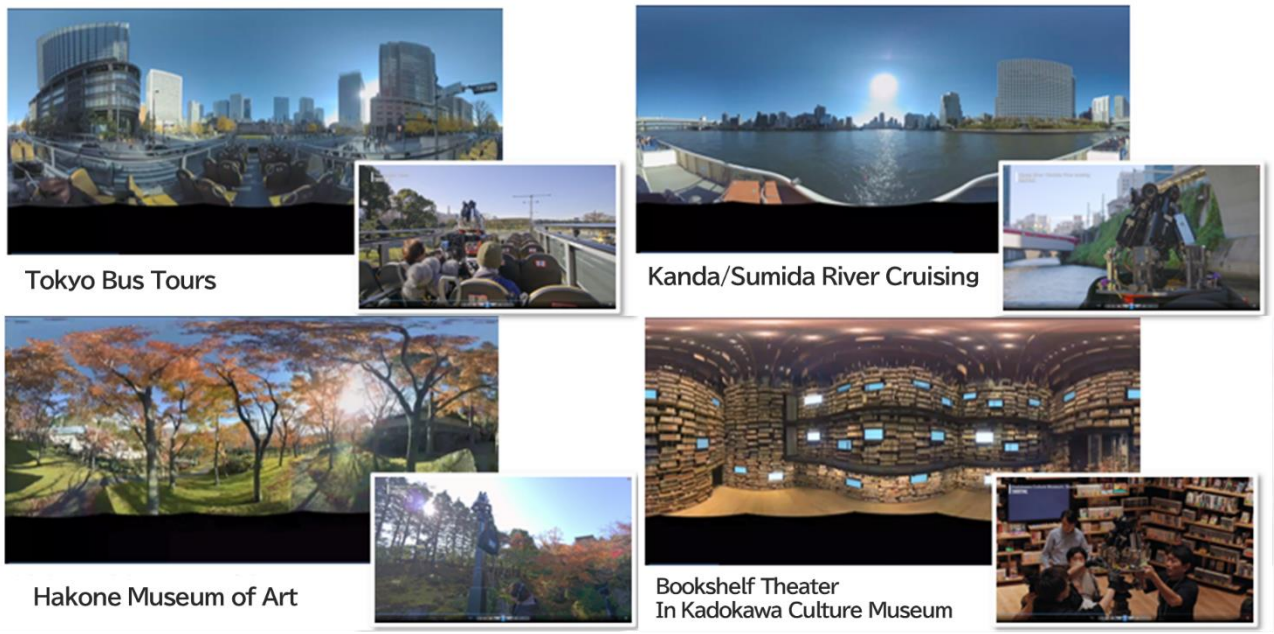


FIGURE 28

Demonstration using the hemispherical display

4.3.8 Immersive sound content for VR images

4.3.8.1 Overview

NHK has produced sound content for 3DoF VR images using audio-related metadata based on the Audio Definition Model (ADM) [12]. VR images were provided as animation of computer graphics data rendered by Unity. A user can walk through a jungle and encounter animals. VR images rendered into 1 920 × 1 080 or 3 840 × 2 160 are displayed on a projector, an LCD monitor or a head mount display (HMD) (see Fig. 29). The sound content consists of background sound, narrations, audio description and near-field sound as shown in Table 9. The main sound content rendered into 24 sound

signals or their binaural stereo signals are reproduced using 24 loudspeakers of sound system H or headphones.

FIGURE 29

Reproduction system

Report BT.2420-17

TABLE 9

Specifications of sound content

Audio programmes (1)	60 programmes (3 viewpoints × 10 languages × 2 reproduction systems (main and second devices))
Audio objects (2)	55 audio objects
Audio channels (3)	128 channels
Background sound	72 channels (9+10+3 (sound system H [13]) × 3 viewpoints), 3 audio objects (3 viewpoints)
Music	36 channels (4+7+0 (sound system J [13]) × 3 viewpoints), 3 audio objects of 7.1.4 (3 viewpoints)
Narrations	10 channels (mono × 10 languages), 30 audio objects of mono (3 viewpoints × 10 languages)
Audio description	4 channels (2 channels × 2 languages), 16 audio objects of mono (4 objects per signal × 2 channels × 2 languages)
Near-field sound	6 channels (2 channels (stereo)) × 3 viewpoints), 3 audio objects of stereo (3 viewpoints)

(1) An audio programme contains all audio contents including narration and background music to make the complete mix. The audio programme has a single set of parameters such as language. This is specified as audioProgramme element in Recommendation ITU-R BS.2076.

(2) An audio content refers to an audio object such as narration, background music and sound effects. The audio object contains a set of the actual audio signals with the format including a loudspeaker layout or reproduced positions. This is specified as audioObject element the in Recommendation ITU-R BS.2076.

(3) An audio channel is an actual PCM audio signal. This is specified as audioTrackUID and audioChannelFormat elements in Recommendation ITU-R BS.2076.

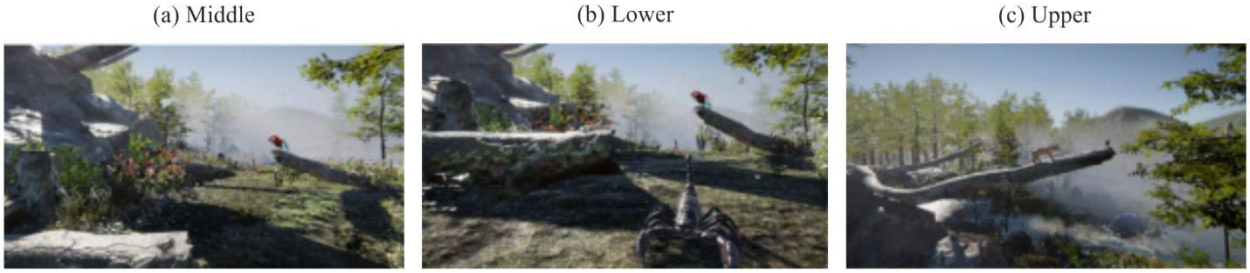
4.3.8.2 Background sound for different viewpoints

Background sounds including music and roars of animals from three viewpoints were produced as channel-based sound signals of the sound systems H (9+10+3) and J (4+7+0) in Recommendation ITU-R BS.2051. Users can change a viewpoint by selecting an audio object for background sound. Three viewpoints at different heights of middle (human's-eye view), lower (mouse's-eye view) and

upper (bird's-eye view) were provided (see Fig. 30). The area of visibility was limited by audio-related metadata to match visual contents with narration and audio description.

FIGURE 30

Viewpoints at different heights



Report BT.2420-18

4.3.8.3 Narration and audio description

The main narration was provided as a static audio object of a monophonic sound signal located in front of the viewer. Multilingual narration was provided by audio objects for a narration (Japanese and English). A user can adjust the level balance between narration and background sound. The range of level adjustment was limited by the audio-related metadata. The user can also turn off the narration.

Audio description for individual animals was provided as static audio objects of a monophonic sound signal. The sound signal temporally recorded in the renderer is reproduced upon the user's request. A single sound signal conveys multiple audio description objects along the timeline.

4.3.8.4 Near-field sound or second sound device

The near-field sound of stereo signals including roars of animals was provided for a second sound device, a wearable neck loudspeaker. The audio programme for near-field sound is automatically switched in conjunction with the audio programme of background sound according to the viewpoint. A second device is connected via Bluetooth. The wearable neck loudspeaker system equipped with a vibrator makes the user experience sound signals as haptic stimuli.

4.3.8.5 User interface

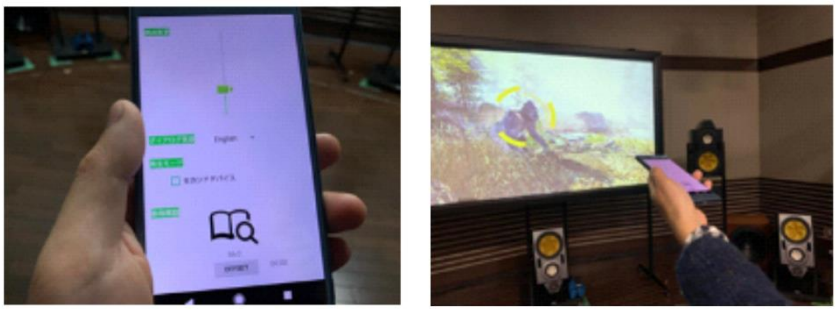
The Open Sound Control (OSC) protocol [14] was used for the user interface to control audio and video renderers. The OSC message including IDs of the audio programme and audio object for the user's action and position was conveyed via the User Datagram Protocol (UDP). A user interface was developed as an application of mobile phone and a user can select reproduction conditions and point to an audio description object using it (see Fig. 31).

FIGURE 31

User interface

(a) Application installed in mobile phone

(b) Pointing out audio description object



Report BT.2420-19

4.3.9 Photography-based VR content

Fuji Television experimentally created 360° VR content from old landscape photographs of Tokyo shot 120 years ago, consisting of 13 photographs covering a 360-degree scene. The 360-degree VR content was produced for a specific HMD (Pimax Vision 8K Plus, see Table 10).

TABLE 10

Specifications of HMD

VR format	StreamVR 1.0
Display panels	3 840 × 2 160 pixels for each eye
Field of view	200°diagonal
Refresh rate	90 Hz maximum

The technical problems solved during the production are described below (see Figs 32 and 33):

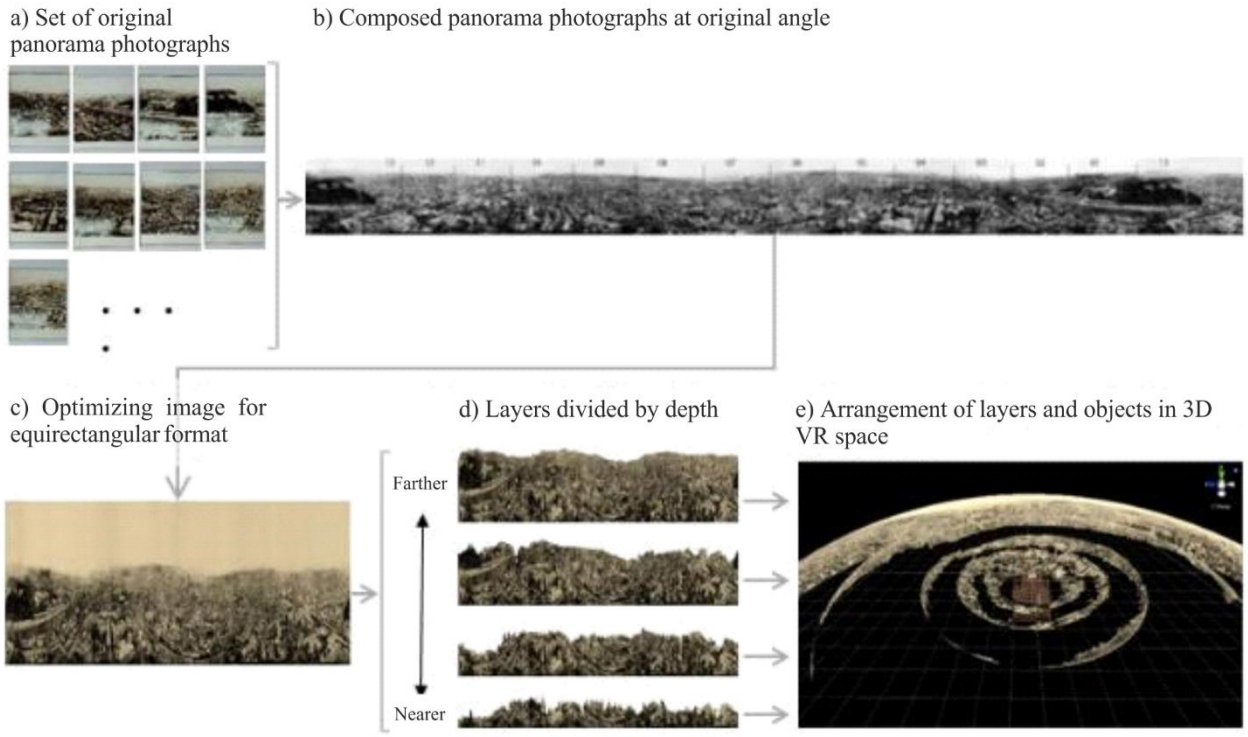
1) The original photographs were not shot in a spherical format but each at a 30-degree angle and needed to be transformed into the equirectangular projection (ERP) format. This required geometric adjustments of the transformation to mitigate visual discomfort and unnaturalness.

2) The original photographs do not contain depth information. To present degrees of immersion, the photographs were manually divided into five layers from near-view to distant-view. The different layers were arranged spatially in the VR space.

3) The original photographs were not shot from a fixed camera position by panning the camera but by slightly changing the position. This means there is a mismatch between the camera position and the centre of the 360-degree image, which could cause visual discomfort when a viewer looks around the scene. To mitigate this problem, the content was designed for the viewer to walk around in the scene rather than look around from a fixed viewing position. Additionally, to reduce visual discomfort the viewing angle was limited by masking the outer peripheral area.

FIGURE 32

Working process of photography-based VR content



Report BT.2420-20

FIGURE 33

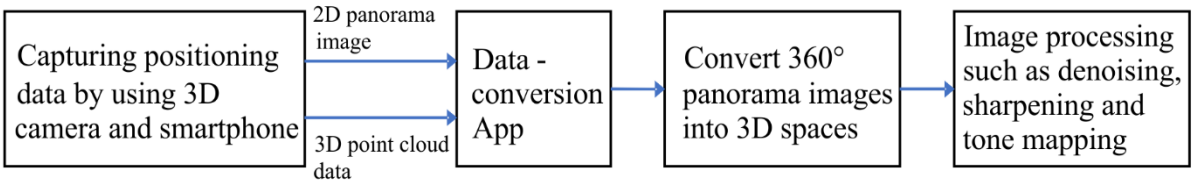
Image of walk around experience of content displayed on HMD

Report BT.2420-21

TV Asahi has produced 3D content that replicates a studio set from photographs. 360-degree panoramic photos are converted to 3D VR content with positional information captured using devices such as 3D cameras and LiDAR cameras (mounted on iPhones and iPads) as 3D sensors (see Fig. 34). This conversion uses an AI engine with deep learning from an architectural perspective to recreate a more realistic space. The 3D content can give viewers an experience of being inside a TV programme production. TV Asahi is considering developing it as interactive 3D content for TV broadcasting.

FIGURE 34

Process flow for cleating 3D content



Report BT.2420-22

4.3.10 Light-field HMD system

4.3.10.1 Overview

When viewing stereoscopic 3D images with conventional head-mounted displays (HMDs), the discrepancy between the convergence distance and the accommodation distance is considered to cause eye strain. NHK has developed a light-field HMD system that utilizes light-field technology to reproduce light rays from objects by designing the optics of the HMD, manufacturing the housing, and developing a program to generate elemental images in real time. The 3D images that can be viewed by a light-field HMD are considered to be consistent with convergence and accommodation distance because the eyes are able to focus in accordance with the depth position of the 3D images.

4.3.10.2 Optical design of light-field HMD

A light-field HMD consists of a display, lens array, and eyepiece. Table 11 shows the specifications of the display and optical elements, and Fig. 35 shows the arrangement of the optical system. The display shows elemental images that contain information of luminance and colour of the 3D images. After passing through the lens array, the light rays from each pixel intersect in three dimensions to form an intermediate image. A viewer can see the magnified 3D image through the eyepiece. By placing the distance between the display and the lens array closer than the focal length of the lens array, the intermediate image is formed at the back of the display as a virtual image, reducing the depth of the display.

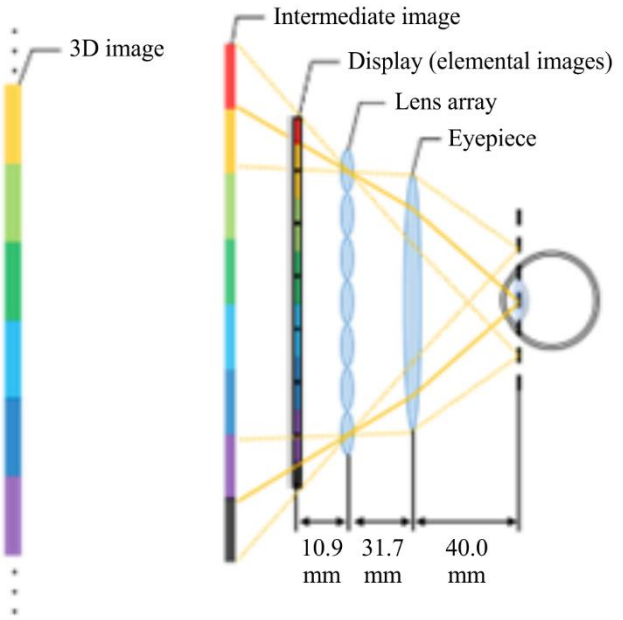
TABLE 11

Specifications of the display and optical elements

Display (per eye)	Resolution	1 440 × 1 440
	Frame rate	60 Hz
	Size	2.9-inch diagonal
Lens array	Pitch	3.0 mm
	Focal length	15.0 mm
Eyepiece	Focal length	77.0 mm

FIGURE 35

Arrangement of the optical system



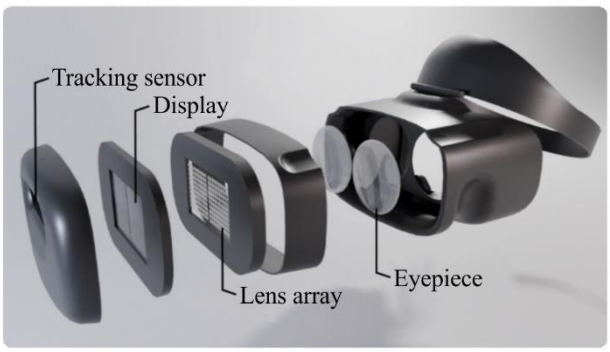
Report BT.2420-23

The designed optical system and a tracking sensor to obtain information of the direction the viewer is facing are mounted in the housing as shown in Fig. 36. Table 12 shows the specifications of the HMD.

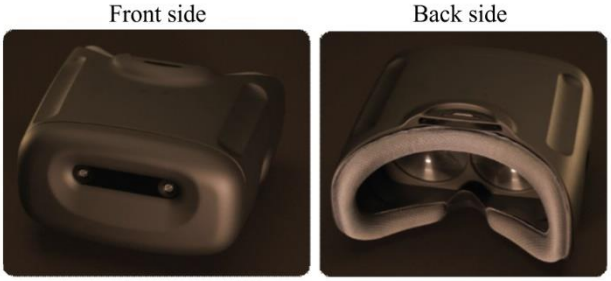
FIGURE 36

Developed light-field HMD

a) Configuration of light-field HMD



b) Photograph



Report BT.2420-24

TABLE 12

Specifications of the light-field HMD

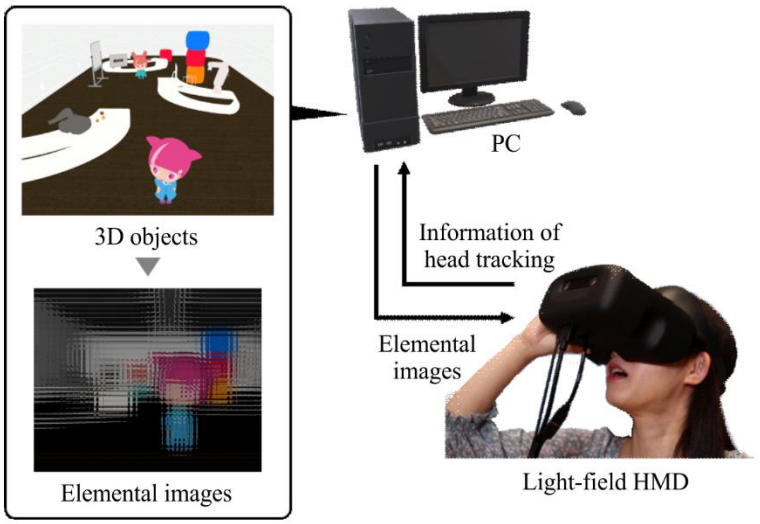
Size	Width 18.2 cm, height 11.3 cm, depth 12.7 cm
Field of view	44° horizontal, 44° vertical
Weight	513 g (without headband)
Tracking sensor	Intel RealSense T265

4.3.10.3 Light-field HMD system

Figure 37 shows the configuration of the light-field HMD system. To generate elemental images from a 3D object, rays emitted from each pixel and passing through the lens array and eyepiece are calculated, and those rays are tracked in the opposite direction to determine their contact with the 3D object and assign the acquired colour to the pixel. 3D VR images can be viewed by having elemental images rendered in real time by ray tracing in accordance with the head movement, as shown in Fig. 38.

FIGURE 37

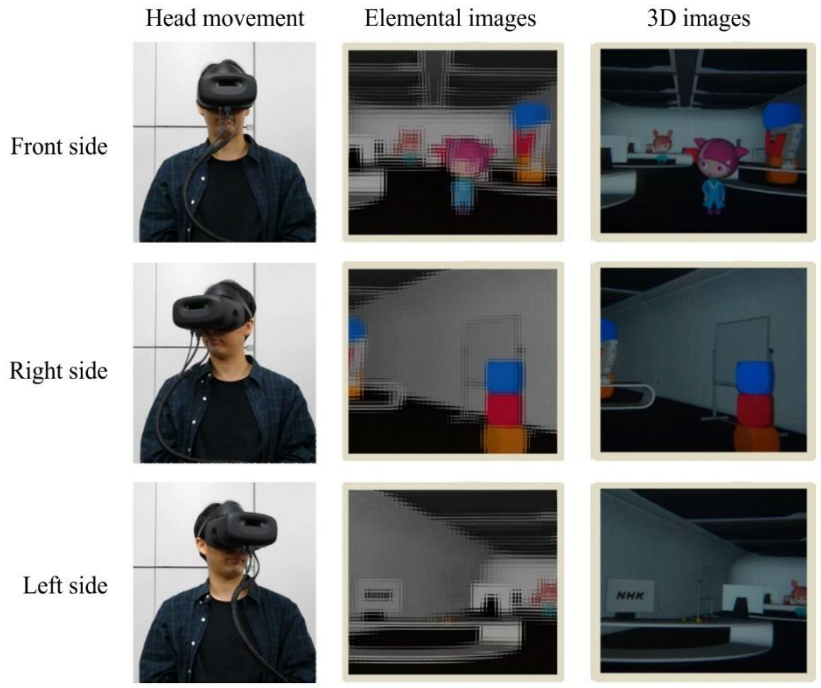
Light-field HMD system



Report BT.2420-25

FIGURE 38

Light-field HMD and 3D images



Report BT.2420-26

4.3.11 Portable interactive 3D display

NHK has developed a portable interactive 3D display (Fig. 39) that enables a natural 3D image to be viewed without the need for special glasses in accordance with the viewing position and operation of the device of a viewer. To realize interactive and high-quality 3D image viewing on a portable device, generating an elemental image array in real time and displaying a large number of light rays are necessary. However, generating an elemental image array in real time on a portable device is difficult owing to it having lower processing power than a high-performance computer. In the case of displaying a 3D image on an existing portable device, a limited number of pixels display results in a low-quality 3D image.

To address the above problems, generating the elemental image array by a server (high-performance computer) instead of a portable device was proposed. Moreover, an eye-tracking function was introduced to the 3D display of the portable device to improve the quality of the 3D image.

The display system consists of a portable device, a server, and a router (Fig. 40). The portable device acquires the viewer's viewing position and operation of the device, and the server generates an elemental image array on the basis of this information. This enables the viewer to view 3D images in accordance with his/her viewing position and operation of the device.

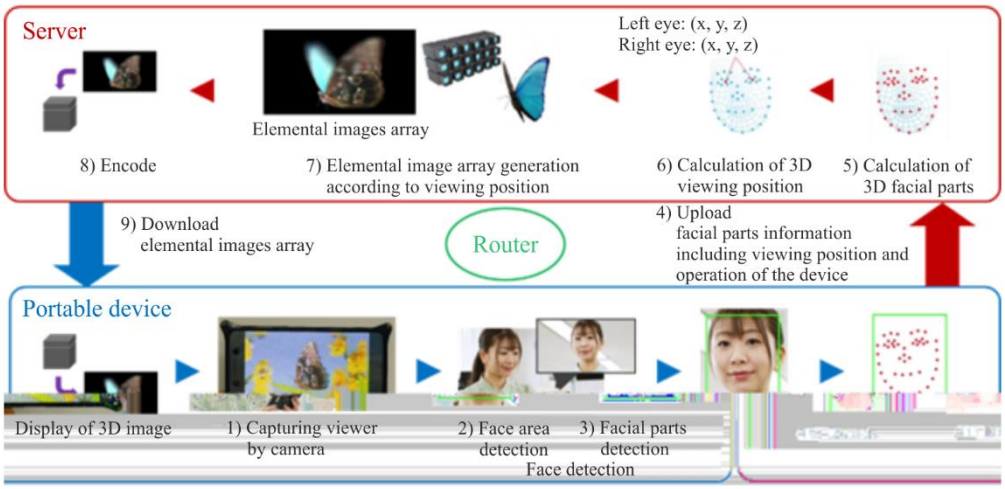
FIGURE 39

Interactive viewing with 3D display

Report BT.2420-27

FIGURE 40

Portable interactive 3D display system



Report BT.2420-28

4.3.12 Audio metadata and production tools for 6DoF audio content

4.3.12.1 Overview

NHK developed ADM-based 6DoF audio metadata and production tools for creating 6DoF audio content. The Audio Definition Model (ADM), a standard audio-related metadata for 3DoF specified in Recommendation ITU-R BS.2076, has been extended to support 6DoF audio content. The production tools include an audio metadata production editor, an audio metadata viewer, and an audio renderer, all of which are compatible with ADM-based metadata. These tools were used to produce 6DoF audio content with embedded metadata.

4.3.12.2 ADM-based 6DoF audio metadata

The ADM metadata were extended with the additional elements necessary to support 6DoF user movement in a 6DoF environment. These extensions include parameters for object location, orientation, radiation characteristics, and distance attenuation. The metadata description followed XML Schema version 1.1.

– **Location and orientation**: ADM originally supports a coordinate system normalised to loudspeaker positions with the user position as the origin. However, in 6DoF audio content, the user's position is dynamic, requiring absolute positioning of the audio objects. Therefore, new elements were added to define the origin within the 6DoF content space and specify the coordinate units.

– **Radiation characteristics**: Directional- and frequency-dependent gain values can be specified to express sound radiation from various directions within a 6DoF space.

– **Distance attenuation characteristics**: Attenuation parameters, including constants and coefficients, can be defined to express the energy attenuation based on the distance between the sound source and user.

4.3.12.3 6DoF audio metadata production tools

An audio metadata editor and an audio metadata viewer were developed to facilitate the creation of 6DoF audio metadata. Because 6DoF metadata involve more parameters and a more complex structure than 3DoF, it is essential to verify both the correctness of the metadata descriptions and alignment with the creator's intent.

– **Audio metadata editor**: A text-based editor developed using open-source software (Microsoft VS Code and Red Hat XML extension) equips validation functions based on the XML schema. It detects and displays error messages for incorrect metadata descriptions in real time (see Fig. 41).

– **Audio metadata viewer**: This tool visualises the locations and orientations of audio objects described in the metadata. It provides top and side views while allowing creators to edit the parameters and acoustic characteristics directly within the viewer (see Fig. 42).

FIGURE 41

Audio metadata editor displaying ADM-based 6DoF audio metadata

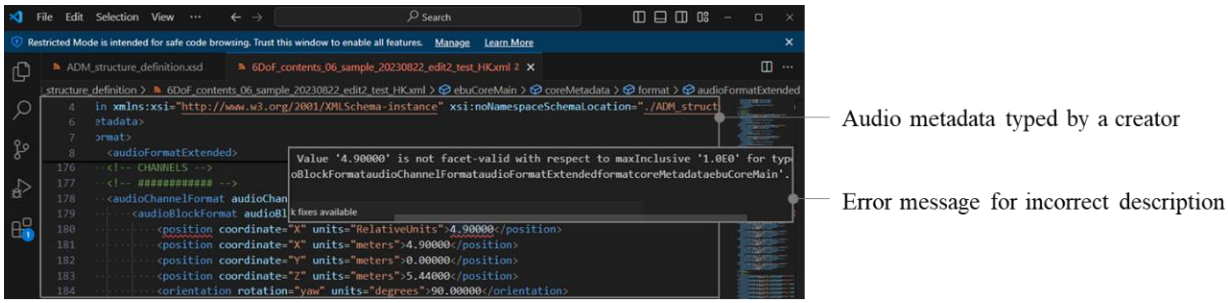
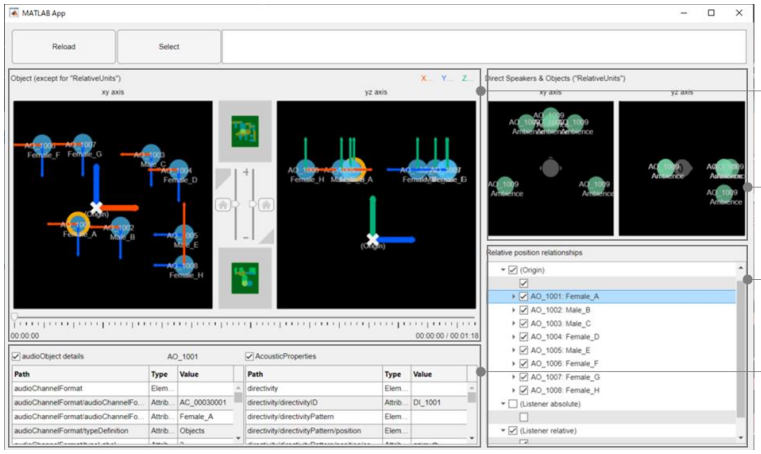


FIGURE 42

Audio metadata viewer loading ADM-based 6DoF audio metadata



4.3.12.4 Audio renderer for ADM-based 6DoF audio metadata

An audio renderer was developed for production. It can generate binaural signals by processing the input audio signals and ADM-based metadata according to the user's position. The renderer also monitors the input and output signals using a bar meter (see Fig. 43). The specifications are listed in Table 13.

FIGURE 43

Audio renderer for 6DoF content

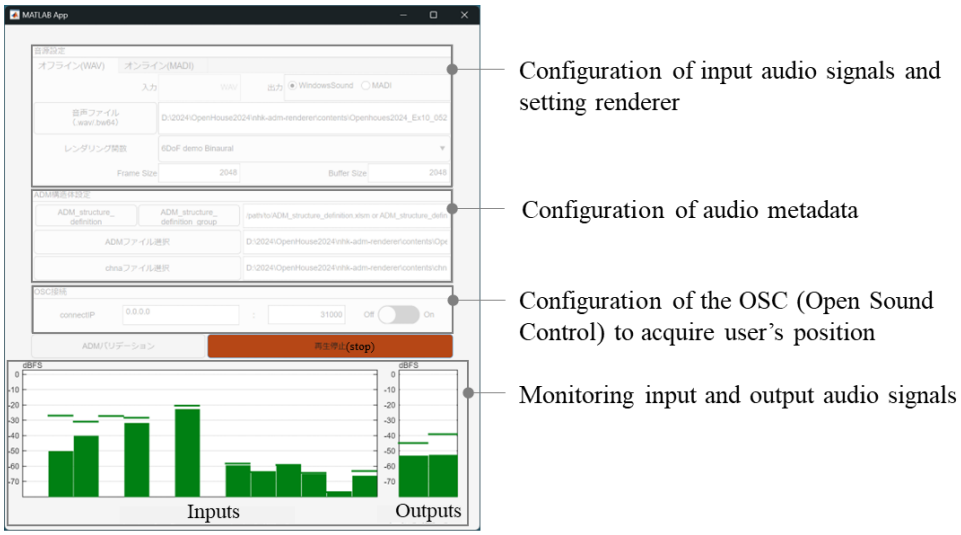


TABLE 13

Specifications of audio renderer for 6DoF content

Audio signal inputs	Up to 64 channels
Audio metadata input	ADM-based audio metadata
Audio signal outputs	2 channels (binaural audio)
Rendering functions	Based on the positional relationships between the user and audio objects, the following processes are applied: – Application of radiation characteristics to each audio object – Application of distance attenuation characteristics to each audio object – Binaural rendering of both channel-based and object-based audio using these characteristics

4.3.12.5 New functionality to maintain the creator's intent

Producing 6DoF audio content generally involves designing a unified acoustic space based on physical acoustic principles and placing audio objects within that space. It is equally important to reflect the intent of the creator in the acoustic design. For instance, different sound sources may be assigned to different acoustic spaces within the content. To support such creative requirements, using ADM-based 6DoF audio metadata allows creators to assign unique acoustic characteristics to each audio object.

4.3.12.6 Example of 6DoF audio content

Sample 6DoF audio content was produced using ADM-based 6DoF audio metadata and associated production tools. This mock 6DoF VR content was showcased using cardboard cutouts instead of actual VR video playback (see Fig. 44). Audio was rendered based on the position of a representative user and shared simultaneously among multiple listeners. The content comprised eight mono-audio objects and one background sound object. Background sounds were produced as channel-based audio signals using Sound System B (0+5+0), as specified in Recommendation ITU-R BS.2051.

Metadata was produced using the audio metadata editor and viewer, with output monitored via the renderer. The audio object positions were set such that the sounds appeared to originate from the cutouts. All the objects shared the same radiation characteristics. However, the distance attenuation characteristics varied depending on the creative intent. Objects #1 and #2 (see Fig. 44), which were intended as the main characters, were designed to be audible from any location. A more gradual distance-attenuation curve was applied to these objects by setting a smaller attenuation constant in the metadata. The other objects were not intended to be prominent, particularly in distant positions. Therefore, a steeper attenuation curve was obtained using a large attenuation constant. By applying different attenuation curves, the renderer was able to emphasise specific objects without reducing the overall sound levels, thereby preserving the creator's intent even when multiple acoustic characteristics coexisted within a single content space (as shown in Fig. 45).

FIGURE 44

Overview of the 6DoF audio content setup

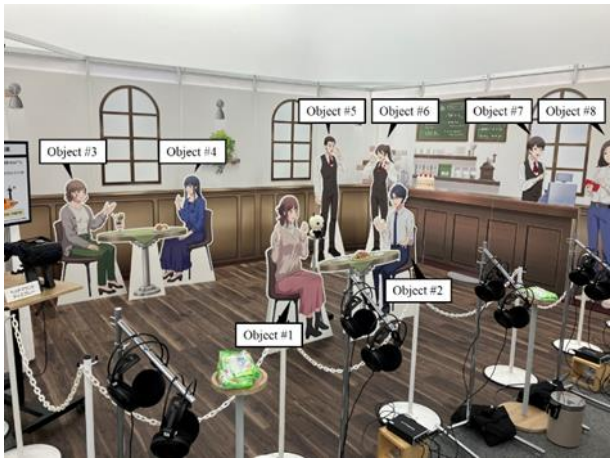
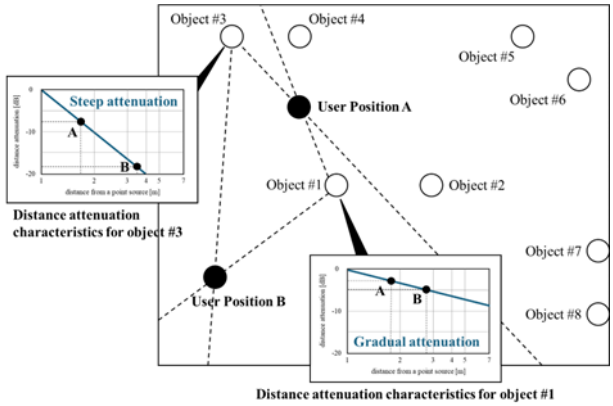


FIGURE 45

Multiple distance-attenuation characteristics in the content space



4.3.13 Harmonised operation between audio and video for presenting immersive content

4.3.13.1 Overview

NHK developed technology that enables harmonised operations between video and audio, allowing users to watch videos and listen to audio presented on different devices from their preferred positions within a virtual space.

4.3.13.2 Technology for harmonised operation between video and audio

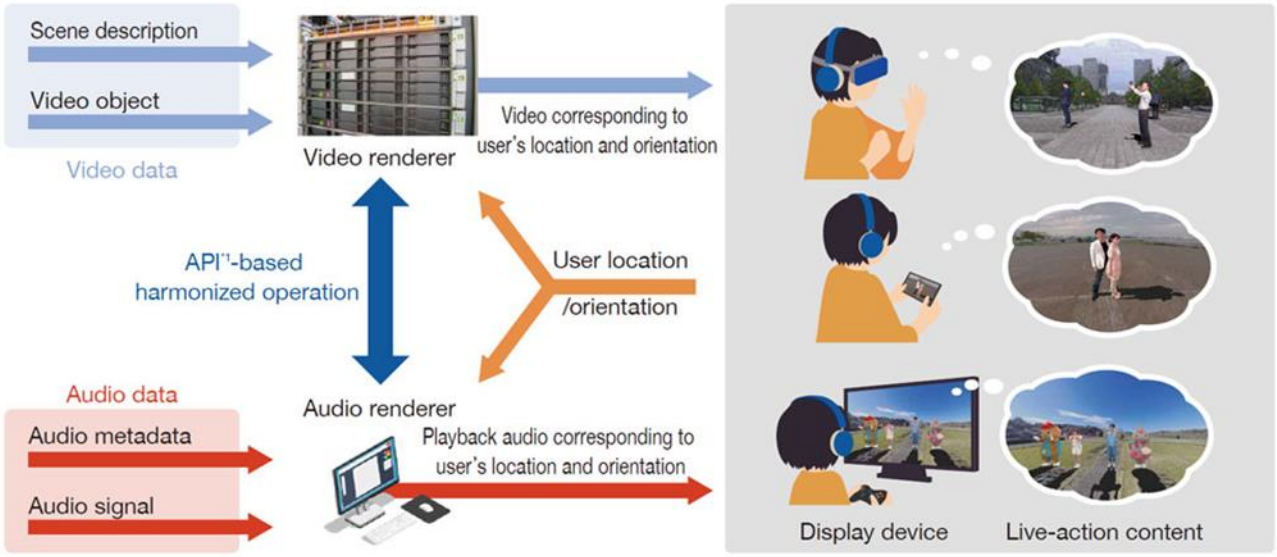
To harmonise video and audio renderers, an application programming interface (API)-based mechanism was developed that distributes the user's positional information to both renderers. Figure 46 shows an overview of the proposed system, in which the video system is based on the high-level system architecture specified in Recommendation ITU-R BT.2154. Video and audio renderers are implemented independently, with video and audio signals rendered on separate PCs. These signals are then presented simultaneously through a coordinated operation enabled by API communication between the renderers.

The API facilitates shared control between the renderers. In this system, the cloud-based video renderer is the primary device, whereas the local audio renderer is a secondary device. The renderers are connected via the Internet. The audio renderer listens to commands from the video renderer, such as play and stop, and immediately responds upon receiving them. This mechanism ensures synchronised video and audio play back.

The video renderer generates 2D images based on the user's viewing position (location and orientation), using 3D data from the virtual space. The audio renderer reproduces audio content created from recorded signals and metadata, such as the sound source position and acoustic characteristics, which are rendered according to the user's position. The renderers have been developed by independently extending MPEG-I standards, including ISO/IEC 12113 "Runtime 3D asset delivery format - Khronos gITF™ 2.0", 23090-4 "Coded representation of immersive media Part 4: MPEG-I immersive audio" (under development) and 23090-14 "Coded representation of immersive media Part 14: Scene description".

FIGURE 46

Overview of the system



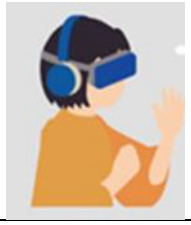
4.3.13.3 Variety of devices

Table 14 compares the different video devices available to users. HMDs provide the highest level of immersion, but are head-mounted, which can be inconvenient. In addition, 6DoF content allows users to move freely within their physical space, often requiring a large operating area. Other challenges include sensory sickness (see § 5.4.2) and restrictions on use by children. The concept of "Magic windows" (see § 2.3) addresses these issues, making it desirable to offer multiple device options for users to choose based on their environment and preferences. Accordingly, 2D displays with game controllers and tablets were prepared alongside the HMDs.

This option reproduces audio exclusively through headphones, as loudspeakers are unsuitable for personal use because of sound dispersion, which may disturb others.

TABLE 14

Variety of video devices

HMD	Magic window (see § 2.3)	
	2D display with game controller	Tablet
		

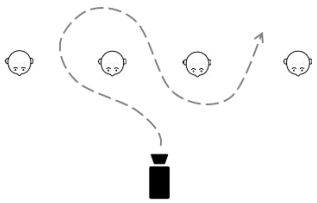
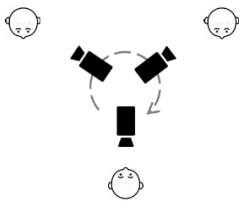
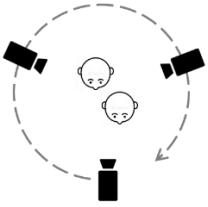
4.3.13.4 Variety and usage of content

Table 15 lists the components of the immersive content. All the content was designed to allow users free movement within a virtual space. Character videos were captured individually using a volumetric video system with 48 cameras, and background videos were captured using a 360-degree camera. The audio signals were produced using conventional methods, including standard microphones and sound libraries. Video and audio content were described using scene description formats independently extended from the MPEG-I standards.

The expected user perspectives are indicated by grey dashed lines, showing examples such as moving between objects, viewing the surroundings from a central location, and navigating around a specific object.

TABLE 15

Components of the contents

Content	Kids programme (dance and music)	Drama	Magic
Image			
Layout of objects and examples of perspectives (Grey dashed lines)			
Audio components	– Singing voices of the characters (mono × 4) – Hand clapping (mono × 4) – Ambience (5 ch) – Music (mono)	– Voices of the characters (mono × 3) – Footsteps (mono × 3) – Other SE (mono × 3) – Ambience (5 ch)	– Voices of the characters (mono × 1) – Finger snapping (mono × 3) – Music (5 ch)

4.3.14 Volumetric audio production system

4.3.14.1 Overview

NHK developed audio production technology that reproduces spatial auditory impressions, such as depth, volume, and directional perception, based on the listener's position and orientation. This was achieved by estimating the three-dimensional sound radiation characteristics of the audio objects (i.e. sound sources), as shown in Fig. 47. This technology, named "volumetric audio production", serves as the audio counterpart alongside volumetric video and is designed for integrated production. Figure 48 shows the workflow of the system.

FIGURE 47

Concept of volumetric audio production technology

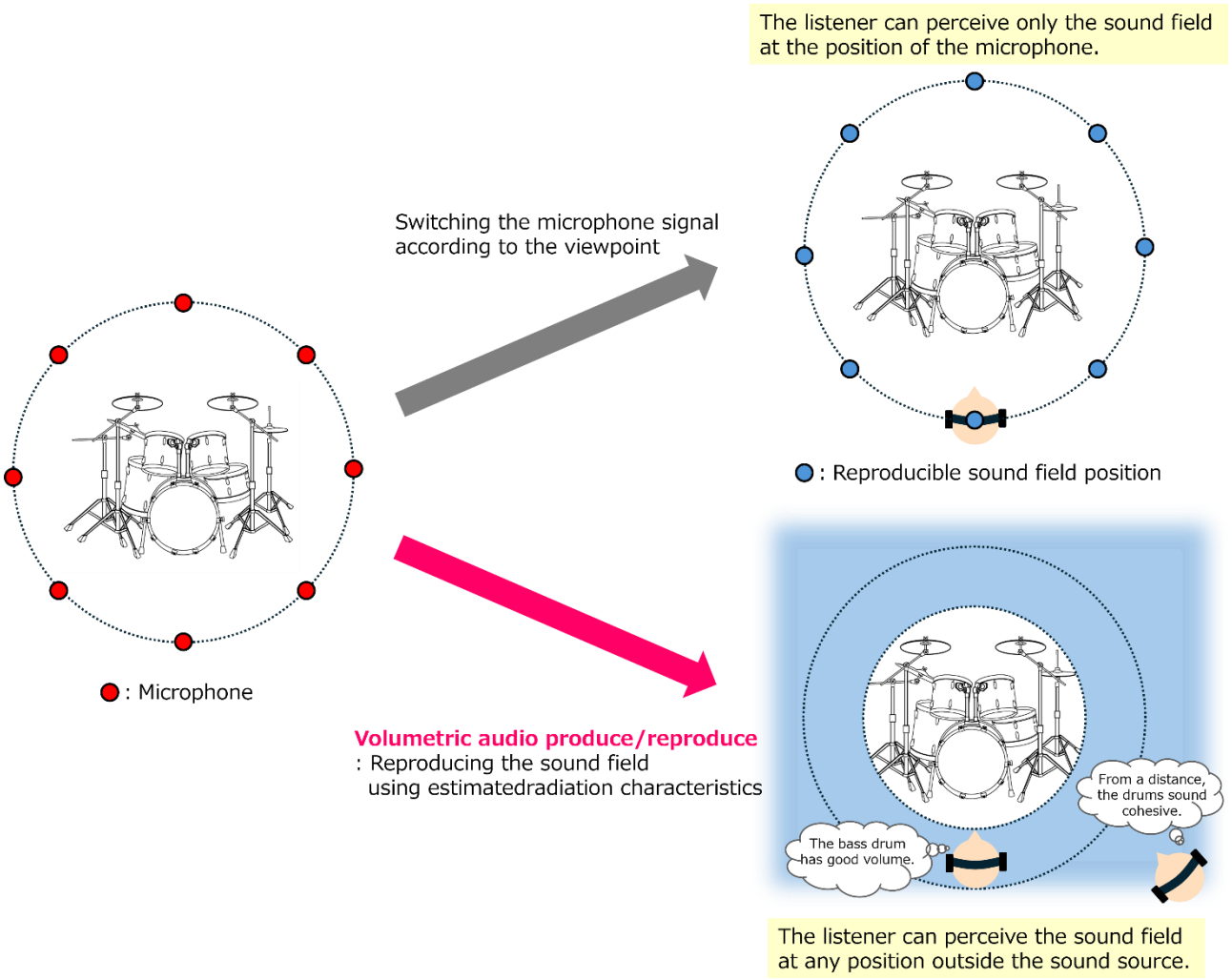
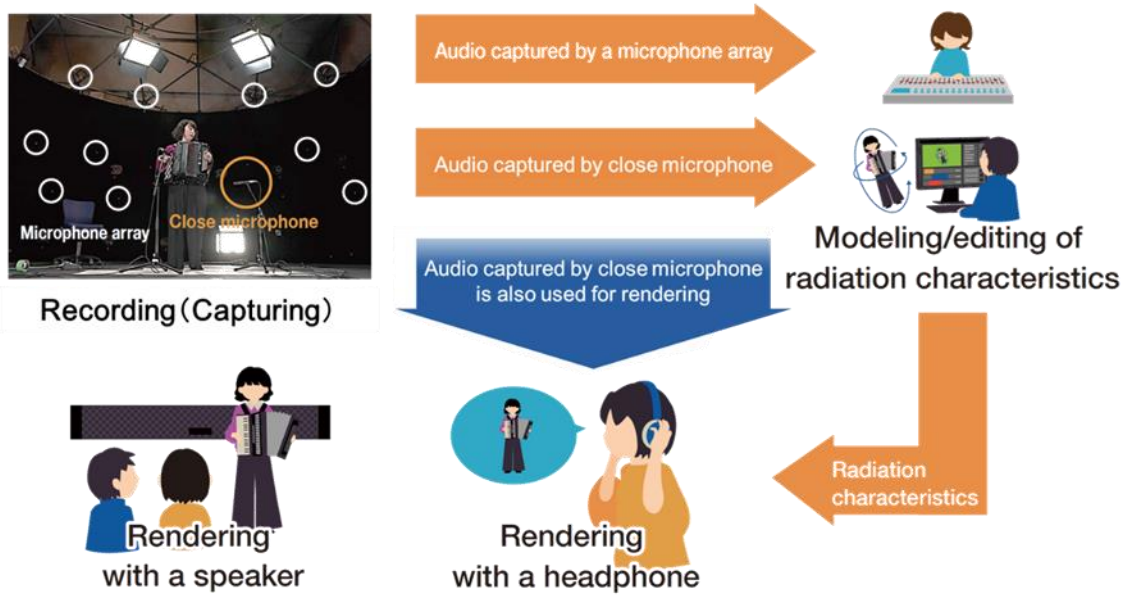


FIGURE 48

Workflow of volumetric audio production and reproduction



4.3.14.2 Recording system

A recording environment was established in Meta-Studio, a facility for capturing volumetric videos, as shown in the top-left section of Fig. 48.

The microphone array in the Meta-Studio setup comprises 82 microphones evenly arranged at the vertices of a hemispherical dome with a radius of 4 m. This array captures sound waves radiating from audio objects and propagates them in multiple directions. The specifications are listed in Table 16.

In conventional programme production, closed microphones are commonly used to capture clear sounds. Volumetric audio production incorporates both a microphone array and closed microphones to ensure comprehensive sound capture.

TABLE 16

Specifications of the microphone array in Meta-Studio

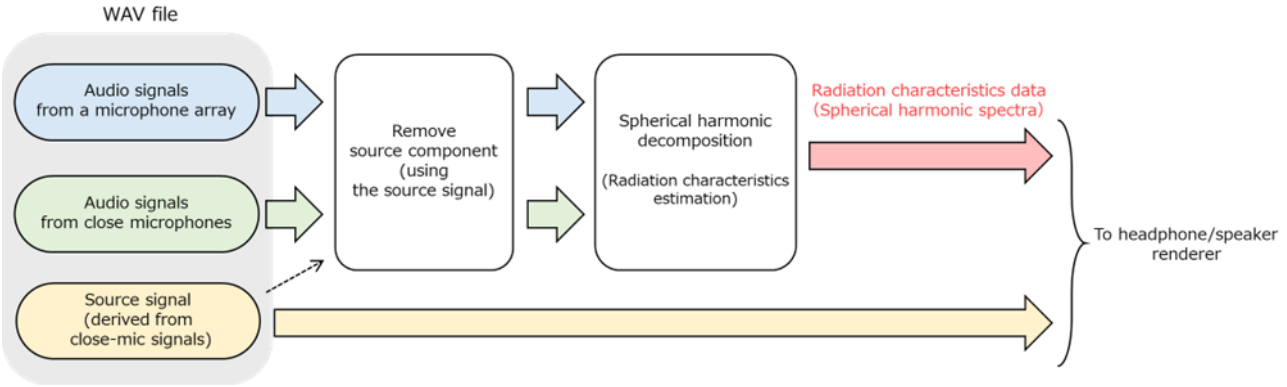
Transducer type	Condenser
Number of microphones	82
Polar pattern	Cardioid
Frequency range	20 Hz-20 kHz

4.3.14.3 Acquisition of radiation characteristics

Technology was developed to estimate the radiation characteristics of audio objects from captured audio signals. By analysing the signals from both the microphone array and nearby microphones using spherical harmonic decomposition, the radiation characteristics were estimated and represented as spherical harmonic spectra (see Fig. 49).

FIGURE 49

Acquisition of radiation characteristics



4.3.14.4 Volumetric audio rendering with headphones

A binaural rendering method has been developed for the headphone-based reproduction of volumetric audio. In this method, the radiation characteristics of audio objects are first converted into sound waves that arrive in the vicinity of the user from various directions (left side of Fig. 50). These sound waves are transformed into binaural signals using head-related transfer functions (HRTFs) (right side of Fig. 50). In this trial, a set of directional HRTFs measured from an exemplar dummy head were used to simulate direct sounds and reflected sounds spread from one volumetric sound source. This approach is expected to deliver more accurate spatial sound imaging than conventional binaural rendering by reproducing the sound field around the listener's head.

FIGURE 50

Concept of volumetric audio rendering

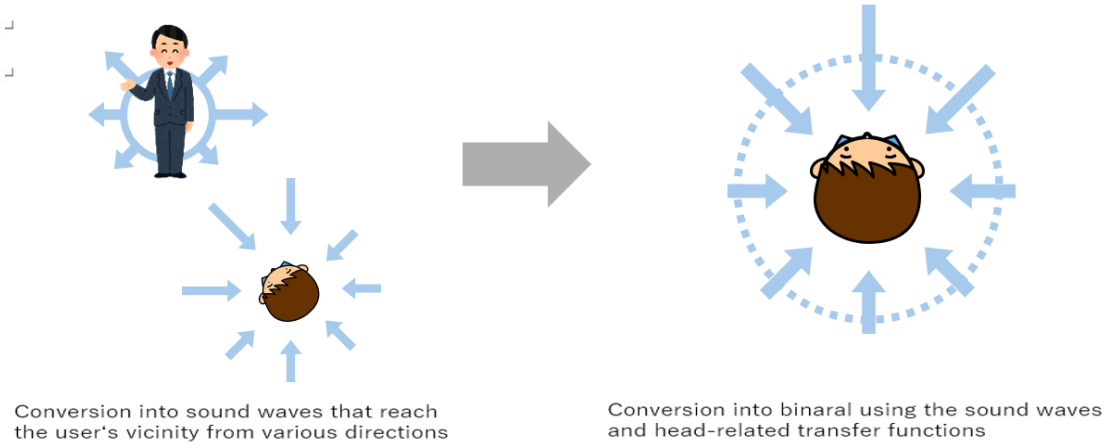


Figure 51 is extracted from a sample performance by an accordion trio. Each audio object was captured individually in Meta-Studio and freely placed within a virtual space, while users can enjoy the video's free camera movements. The audio is rendered based on the relative position and orientation of the audio objects and camera, that is, the user's viewing during the performance.

FIGURE 51

Content of an accordion trio performance

4.4 AR trials

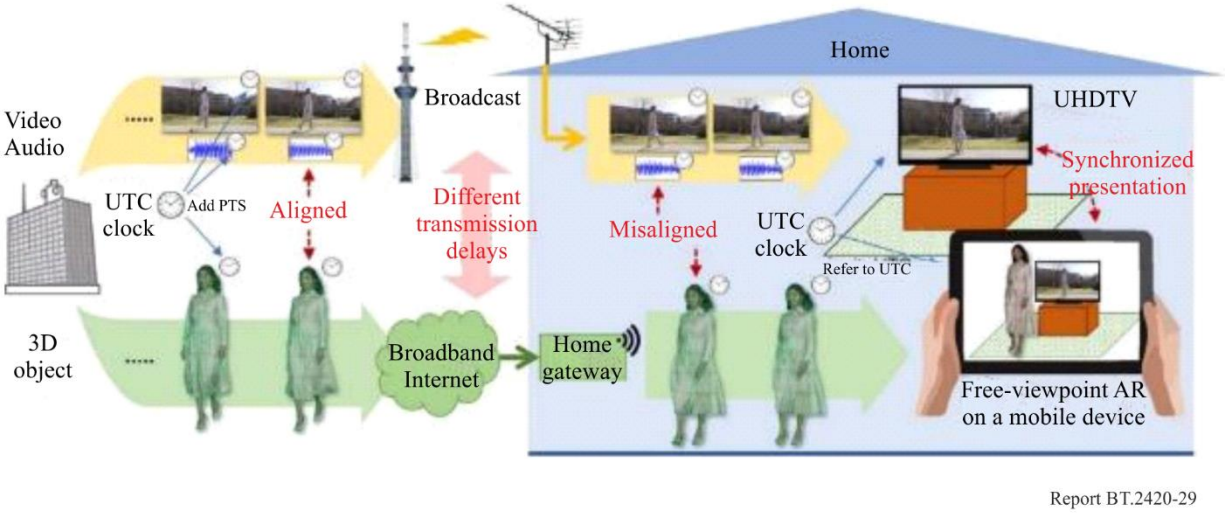
4.4.1 Free-viewpoint AR presentation synchronized with a TV programme

4.4.1.1 Overview

NHK developed a prototype for a content and distribution system based on the concept of integrated TV and AR (see § 3.5). Figure 52 shows a distribution model of a TV programme synchronized with AR content [15]. The TV programme and 6DoF AR content, which extends the world of the programme beyond the TV screen, share a common timeline and storyline. While TV video and audio are delivered over normal TV broadcasting, 3D objects linked to the TV programme are delivered in real-time through broadband Internet and rendered with free viewpoint in accordance with the manipulation of the AR device by the viewer. To enable synchronized presentation of AR content with a TV programme, presentation timestamp (PTS) based on Coordinated Universal Time (UTC) may be used, which is supported by MPEG media transport (MMT).

FIGURE 52

Distribution model of synchronized TV programme and AR content



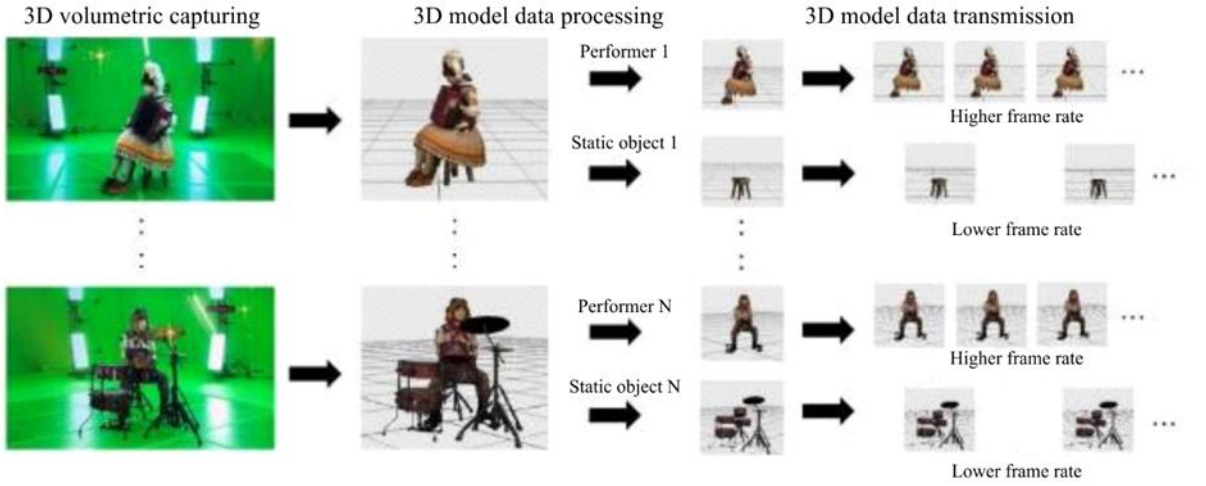
4.4.1.2 Production

Prototype content where performers were playing musical instruments, was created in a volumetric capture studio. Figure 53 shows the 3D volumetric capturing, and the data processing and transmission. Each performer's sequential volumetric data at a fixed frame rate are generated from

the images captured using multiple cameras surrounding a performer. Three-dimensional models are encoded to the 3D geometry of the 3D object using multiple polygons and 2D texture images containing the patterns to be displayed on the faces of these polygons mapped onto the model. The 3D models for the static objects, e.g. musical instruments and chairs, are separated from the moving objects in the scene to create separate 3D models. Wavefront object files (.obj) [16] are used to represent the uncompressed volumetric data consisting of polygonal geometry and Google Draco [17] used to compress the geometry during transmission. The texture images are compressed by JPEG. The number of vertices in the 3D model geometries and the resolutions of the texture images are adjusted to match the size of the AR display and bit rate available through the transmission channel to the viewing device. The audio objects are recorded for each instrument and placed at the positions of the video object.

FIGURE 53

3D volumetric capturing, data processing and transmission



4.4.1.3 Real-time streaming and reception of 3D model data

Three-dimensional objects that make up the content are separately encoded with each object given its own transport packet ID. The encoded objects are then multiplexed into a single IP packet flow. This makes it possible to change the interval of data transmission depending on the object on the sender side, to easily discard data for objects that do not need to be displayed, and customize the layout of objects by user operations on the receiver side.

Figure 54 shows a reception scene in which the performer in the TV programme displayed on a TV appears on the tablet in front of the TV. The user can freely walk around the world of a programme from any perspective by changing the position and direction of the tablet and can also enjoy the changes in the sounds linked to the viewpoint and timbre of the various musical instruments.

FIGURE 54

Prototype AR content synchronized with TV programme

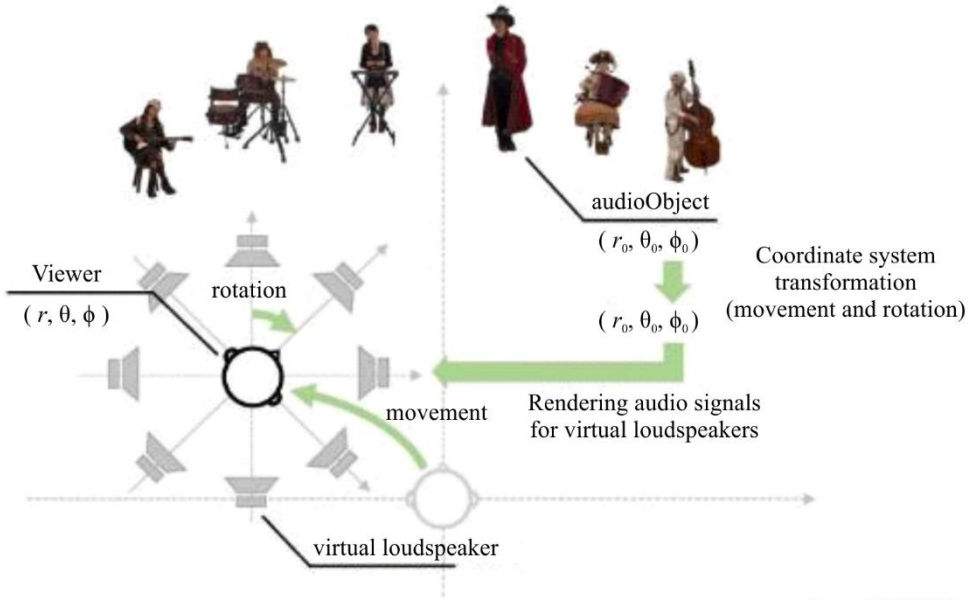
Report BT.2420-31

4.4.1.4 Audio system

Object-based audio (see § 4.3.8) was used in the prototype content, where audio objects coinciding with video objects were placed on the virtual space using the virtual loudspeakers of sound system H (9+10+3) [13] placed around the viewer. A renderer was developed that can reproduce the 6DoF sound according to the viewer's movement and rotation. The renderer transforms the position coordinates of audio objects based on the viewer's movement and rotation, remapping the audio objects to the transformed coordinate system around the viewer (see Fig. 55) and also reproducing the sound using binaural processer. These processes were performed in real time in accordance with the position and gaze direction sent from the tablet in a 0.1-second cycle.

FIGURE 55

Transformation of coordinate system around viewer



Report BT.2420-32

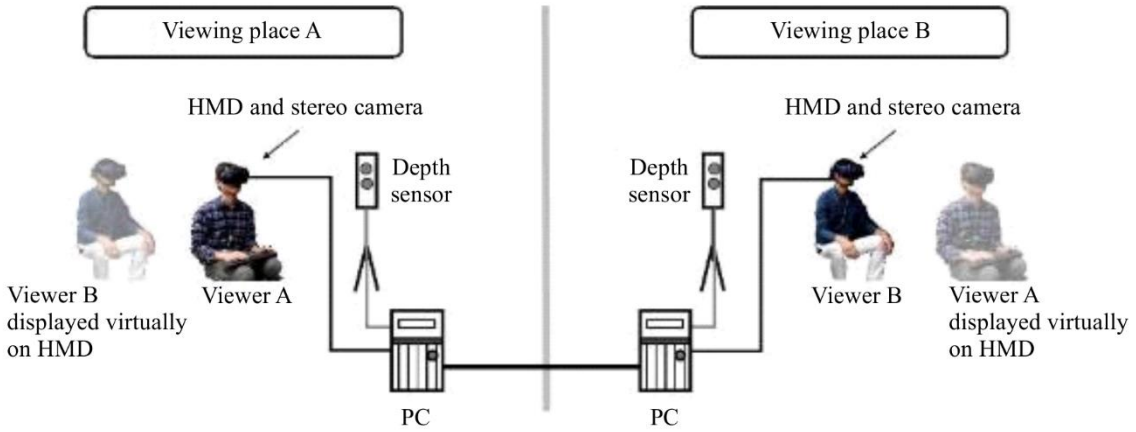
4.4.2 Watching VR/AR content with remote viewers

NHK developed an experimental system for 'virtual space sharing' (see § 3.5) in which three-dimensional objects of TV performers or family members and friends in different locations are

combined and displayed in their actual size to the viewer through the HMD [18]. Figure 56 shows the outline of an experiment where users at two different locations simultaneously watch VR/AR content by wearing an HMD. The three-dimensional position of the HMD is captured by sensors. A microphone attached to the HMD captures the voice of the viewer and a depth-sensor captures the volumetric images of the viewer. Information elements captured from one viewing location are transmitted in real-time to the second viewing location, and vice versa. The virtual images of the person at the remote location are reproduced on the HMD along with his/her voice, thus enabling aural communications as if both viewers are watching the content together from the same location.

FIGURE 56

Experimental system for 'virtual space sharing'

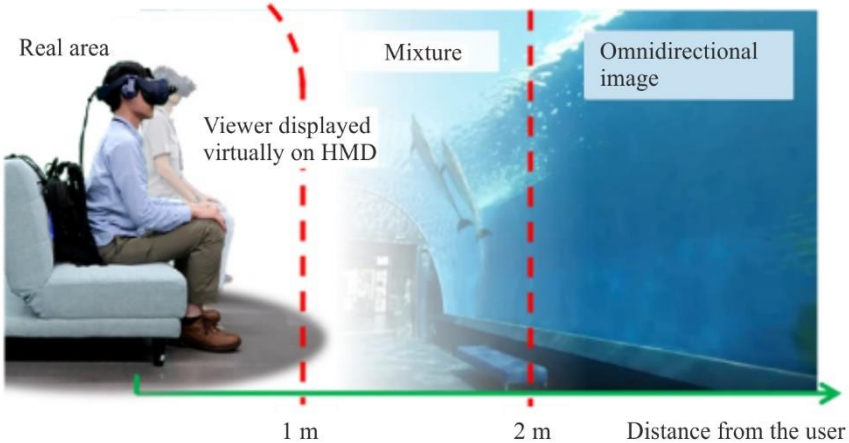


Report BT.2420-33

A stereo camera attached to the HMD captures the scene and depth around the viewer and enables reproduction of a mixture of real and virtual images on the HMD. In this experience, the real scene near the viewer (including the viewer's hands and body) is displayed to the user along with presentation of the VR content and virtual image of the remote user, as shown in Fig. 57.

FIGURE 57

Mixing real space and VR space according to distance



Report BT.2420-34

Figure 58 shows three prototype content types for watching on an HMD with remote viewers. The viewers can share the immersive visual experience in a 360° VR environment. ‘Aquarium’ lets users enjoy the 360-degree space with family and friends. ‘Aerial TV’ shows a 2D TV frame floating in the air with an arbitrary image size. ‘AR instructor’ is a life-sized volumetric AR element, where a performer is displayed with his/her feet on the ground.

FIGURE 58

Images of prototype contents

(a) Aquarium (b) Aerial TV (c) AR instructor

Report BT.2420-35

4.4.3 AR application with CG overlay

Nippon TV developed CG content that uses an AR system called “Nippon TV Mixed Reality”. This enables multiple viewers in a room to wear transparent smart glasses that display data and computer graphics outside the television screen. Characters and programme guests ‘pop’ right out of the screen, moving and speaking realistically in front of viewer’s eyes. During live sports broadcasts, viewers can bring athletes and useful data straight into their living room. Fans of musicians can also take the fun and involvement to a whole new dimension by "welcoming" the artists into their home to perform.

https://www.facebook.com/ntvmr/

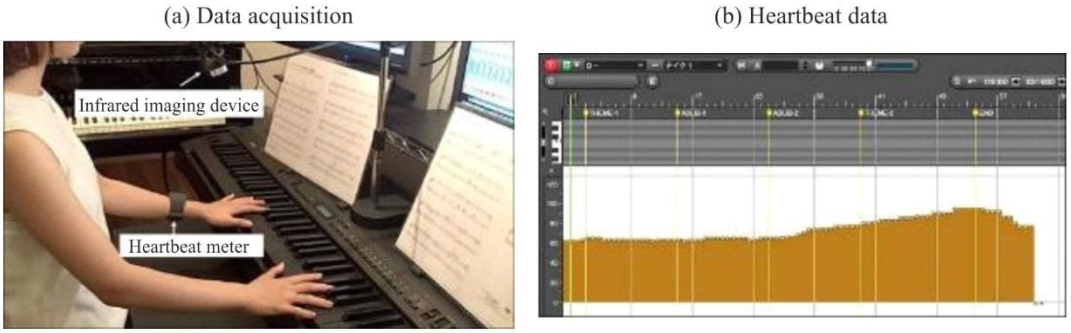
4.4.4 AR application with visualized emotions

Nippon TV has also developed an AR application that introduces mixed experiences of visualized finger movements and emotions of a professional piano player to viewers wearing optical see-through HMDs or watching other displays. This application can guide the viewer on how to play the piano by showing virtual finger movements of the professional player without having to be able to read music.

The player’s fingers and body movements are captured using infrared imaging devices. The finger-capture device estimates the detailed position and action of fingers including each joint. The body capture device estimates the geometry and colours of the body surface which are recorded as point cloud data. The emotion data is estimated from the time sequential sampling of user’s heartbeats and converted to MIDI format to synchronize with audio data performed by the player. Figure 59 shows a scene of data acquisition and heartbeat data.

FIGURE 59

Data acquisition and heartbeat data



Report BT.2420-36

The captured data is visualized as 3D CG virtual images and mixed with a real scene on the optical see-through HMD. To simulate time sensitive emotion, heart rate and synchronized touch intensity data is used. The higher the heart rate and stronger the touch intensity, the stronger or higher the emotion and vice versa. Emotions are depicted as coloured particles. Blue denotes tranquillity, while red denotes strong or high emotion. The number and size of particles are proportional to emotional strength. The shapes and sizes of fingers are modelled on the player's fingers, and movements are replicated by specifying finger-joint locations in the virtual space. The user wearing an optical see-through HMD can see the virtual image as shown in Fig. 60, together with the real scene.

FIGURE 60

Virtual image of fingers and emotion

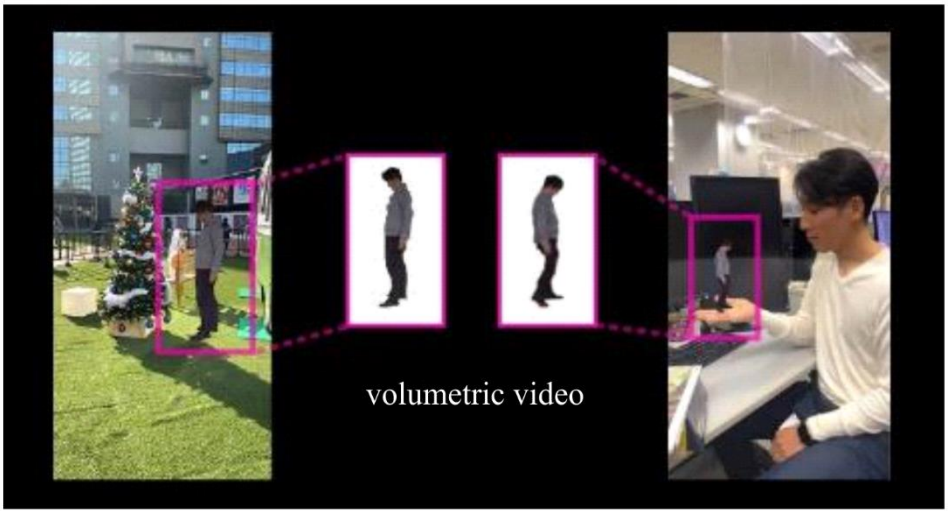
Report BT.2420-37

4.4.5 Smartphone application for volumetric 3D content

Tokyo Broadcasting System Television developed an AR application for smartphones. After installing the application and downloading the volumetric data, users can scale, move, and rotate the volumetric video overlaid on the picture shot using their smartphone cameras and even store the AR-displayed still image on their smartphones (see Fig. 61). While volumetric-video-data size depends on factors such as video length and video quality, most of the content produced and distributed is 10 seconds long with a data size of 50 to 100 MB.

FIGURE 61

AR-displayed image on smartphone screen



Report BT.2420-38

This application was first used in a music programme broadcast in March 2021. In the programme, some of the singers' volumetric video data was distributed with the broadcast, and the application controlled the presentation timing so that the viewer could experience the AR content during the broadcast programme.

4.4.6 Volumetric video format

The volumetric capture system 'Meta-Studio' described in § 5.2.6 captures various information for photo-realistic relighting. The descriptions and layers to be stored for each data point are listed in Table 17.

TABLE 17

Information captured by Meta-Studio

Name	Description	Layer
Position	3D position (x, y, z) for each point consisting of point cloud	Base
Normal	3D normal vector (n_x, n_y, n_z) of the surface at each point	Base
Diffuse	Diffuse reflection (K_d) of each point for R, G, and B	Base
Specular	Specular reflection (K_s) of each point for R, G, and B	Base
Ambient	Ambient reflection (K_a) of each point for R, G, and B	Base
Surface light field	Specific reflection patterns (rays) caused by illumination	Enhancement

A scalable volumetric video format comprising two layers, the base and enhancement layers, as illustrated in Fig. 62 is adopted in the data captured by Meta-Studio.

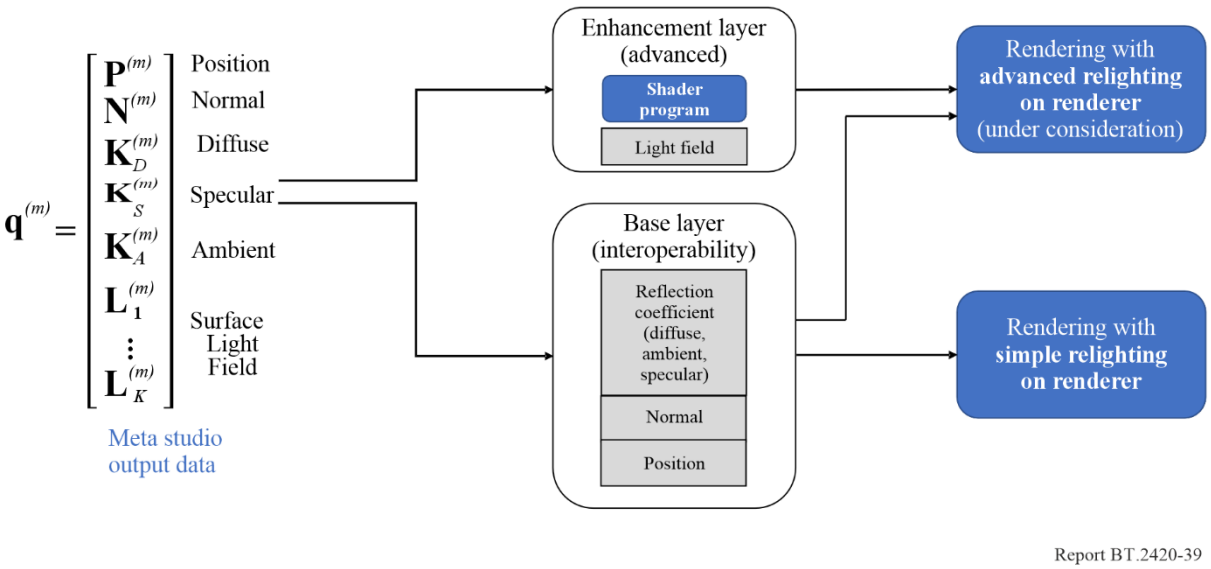
The base layer was designed for compatibility with the conventional format, rendering it a simple relighting method that seamlessly integrates with existing 3D applications. This is specified by partially extending the description of existing 3D file formats to both uncompressed and compressed formats.

Conversely, the enhancement layer contains the remaining components that are not included in the base layer, incorporating sophisticated rendering shading programs as extra components for photo-

realistic relighting. It is designed to enable comprehensive utilisation of advanced photo-realistic rendering methods to realise higher image quality reflecting the optical surface properties.

FIGURE 62

Scalable volumetric video format for Meta-Studio



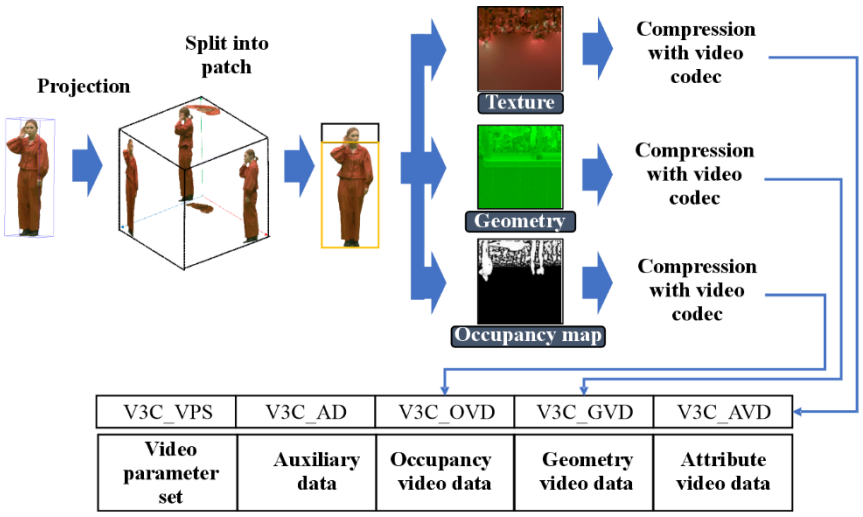
The uncompressed base-layer format was represented by the Stanford triangle format (.ply) and the compressed base layer is used for an extension of the MPEG volumetric video format called visual volumetric video-based coding (V3C). A detailed format for the enhancement layer is also being considered.

The V3C format treats point clouds as 2D textures by projecting them onto a surrounding axial bounding box. By compressing the projected 2D point cloud using existing video codecs such as AVC or HEVC, the bitrate can be significantly reduced. In the V3C sample stream format, as shown in Fig. 63, streams are constructed for each unit, called the V3C Unit. V3C_VPS represents the video parameter set for the following V3C units: V3C_AD demonstrates valid pixel areas and offset information in the subsequent 2D images, which are essential for restoring them to 3D images. The subsequent V3C_OVD, V3C_GVD, and V3C_AVD videos were compressed 2D videos encoded by video codecs such as AVC or HEVC, representing occupancy, geometry, and texture images, respectively. The texture image contained in the V3C_AVD usually assumes RGB values given in the ply format, and a syntax for the normal that allows for simple reflection is defined (Table 18). The reflectance information (ambient, specular, and diffuse reflections) contained the RGB values for each type of reflectance coefficient and vertex. Because the reflectance information is not defined in the syntax of V3C_AVD, the base-layer information is stored using an “unspecified ID”.

NHK developed a V3C encoder that multiplexes the V3C Units described previously by adding functions to the Test Model Category 2 (TMC2), the reference software used for standardisation. The encoder processes an input uncompressed base layer format to generate a compressed base layer stream. The encoder compresses the data in the base-layer format to generate a V3C stream with V3C_AVD units for the reflectance information as illustrated in Fig. 64.

FIGURE 63

V3C sample stream structure



Report BT.2420-40

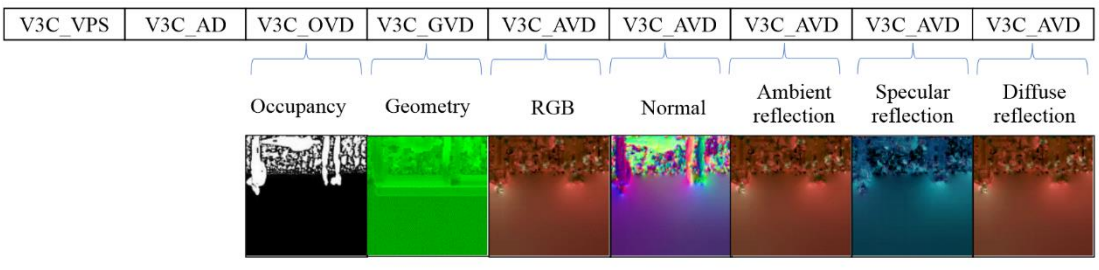
TABLE 18

Attributes type id syntax defined in V3C format in ISO/IEC 23090-5

ai_attribute_type_id[j][i]	Identifier	Attribute_type
0	ATTR_TEXTURE	Texture
1	ATTR_MATERIAL_ID	Material ID
2	ATTR_TRANSPARENCY	Transparency
3	ATTR_REFLECTANCE	Reflectance
4	ATTR_NORMAL	Normals
5-14	ATTR_RESERVED	Reserved
15	ATTR_UNSPECIFIED	Unspecified

FIGURE 64

Bitstream structure and example of a 2D projected point clouds frame in a compressed base layer



Report BT.2420-41

4.4.7 Haptic interfaces

4.4.7.1 Overview

Traditional content usually conveys only visual and audio-based information, but it would be more immersive by adding haptic information. The sense of presence when watching video could be enhanced through tactile stimuli from haptic devices, enabling video content to provide immersive experiences. A variety of haptic devices can be envisioned depending on the genre and content of a video programme.

4.4.7.2 Haptic presentation of impact stimuli and vibration

A ball-type haptic device envisioned for use mainly in sports programmes is shown in Fig. 65 [19]. This device is equipped with a vibrator and two servomotors, as shown in Fig. 65(b). It can convey the magnitude of an impact through vibration and the direction of a ball's movement by physically deforming the surface of the device. This ball-type haptic device enables the viewer to experience the impact applied to a ball or athlete and their movements as haptic stimuli.

A cube-type haptic device that can be used for diverse types of content in addition to sports is shown in Fig. 66 [20]. This device is equipped with a vibrator on each internal face of the cube, as shown on the left, right, top, and bottom faces in Fig. 66(b), so that each of the four faces can be vibrated independently. It can richly convey the 3D position and movement of the subject within the video space being shown. This device is also effective for showing subjects' salient motions to viewers of educational and animation programmes.

Figure 67 shows an overview of a haptic perception system for a sports broadcast programmes [21]. In this system, the types and timing of haptic stimuli, which are related to athletes' motions, are generated by analysing the broadcast video. The haptic devices are actuated with the haptic stimuli synchronized with audio and video. Two types of haptic devices, handheld and chair, were developed for the system. The handheld device mainly conveys the intensity of athletes' motions, and the chair device conveys the timing and impact of a specific action event such as throwing a trick and ground fighting in Judo matches. These haptic devices are expected to serve as interfaces for enhancing viewer immersion in broadcast content.

FIGURE 65

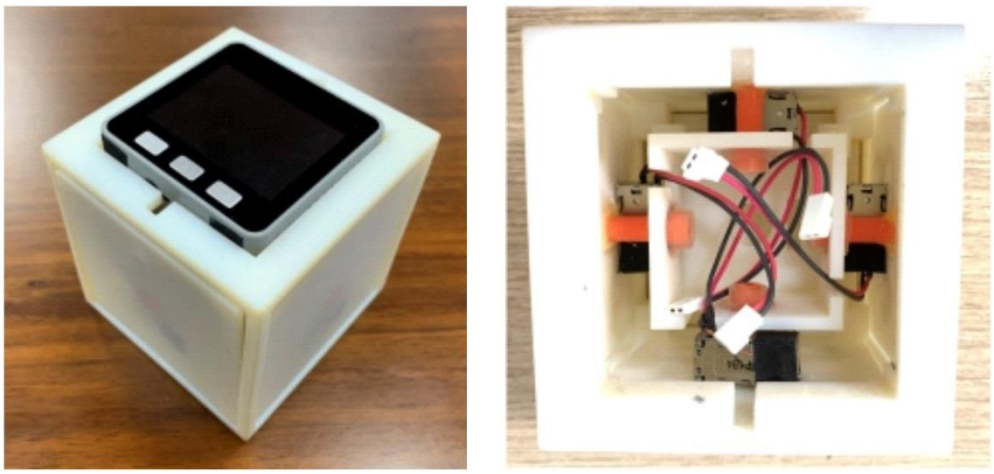
Ball-type haptic device

(a) External appearance (b) Internal structure

Report BT.2420-42

FIGURE 66

Cube-type haptic device

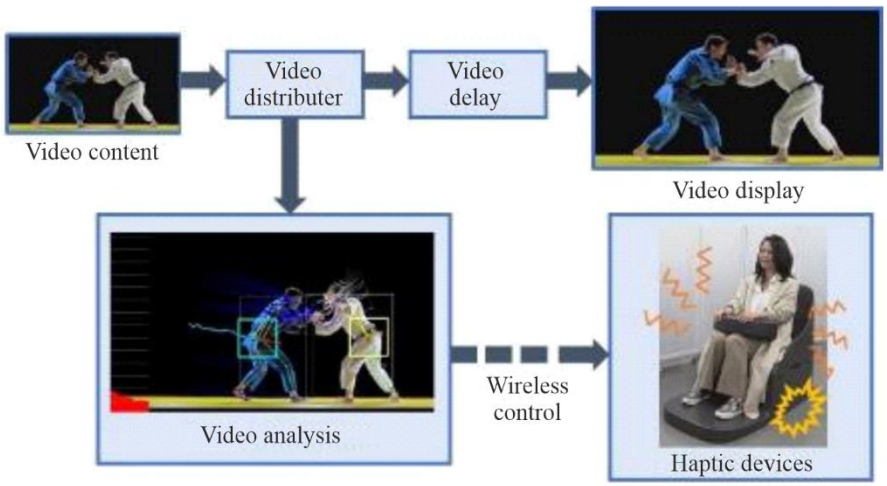


(a) External appearance (b) Internal structure

Report BT.2420-43

FIGURE 67

Haptic perception system for sports programmes

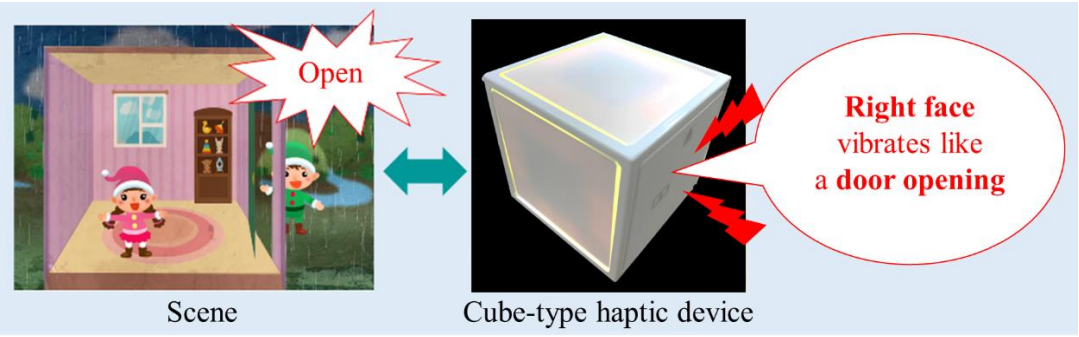


Report BT.2420-44

A linear resonant actuator (LRA) is suitable for the cube-type haptic device for conveying a more realistic sense of touch. The rapid and precise response provided by an LRA-type transducer allows touch feedback and presents tactile vibrations that match the visual image better. Figure 68 shows a schematic of a haptic presentation corresponding to a video scene, where users feel as if a miniature character is entering and leaving the cube-type haptic device [22]. In a scene where a character outside the house opens the door on the right side of the house and enters, the right side of the box vibrates, which corresponds to the sound of the door opening. The bottom then vibrates, which corresponds to the sound of the walking. The presentation of tactile stimuli in this manner allows users to feel as if the character has entered the box.

FIGURE 68

Schematic of a haptic presentation corresponding to a video scene



4.4.7.3 Haptic presentation of vibrotactile and warm/cold sensations

The presentation of thermal sensations in combination with vibrations provides a more immersive experience to users. People identify the characteristics of objects they touch not only by their shape, hardness, and texture but also through sensations of warmth and cold.

Technologies have been developed to provide warmth or coldness to users including temporal changes in temperature. Figure 69 presents an overview of a tablet-like device that presents vibrotactile and thermal sensations in a video scene. This device is equipped with two thermo-haptic feedback units that comprise a Peltier element and an LRA-type transducer as shown in Fig. 70. By using a material with high in-plane thermal conductivity on the surface that the user touches, vibrotactile and warm/cold sensations are simultaneously presented. When a cold carbonated beverage is poured into a glass in the video, the user's hands experience a cold, fizzy sensation. Furthermore, when hot tea is poured into a cup in the video, the users feel a warm, gurgling sensation in their hands.

FIGURE 69

Tablet-like device that presents vibrotactile and thermal sensations in a video scene

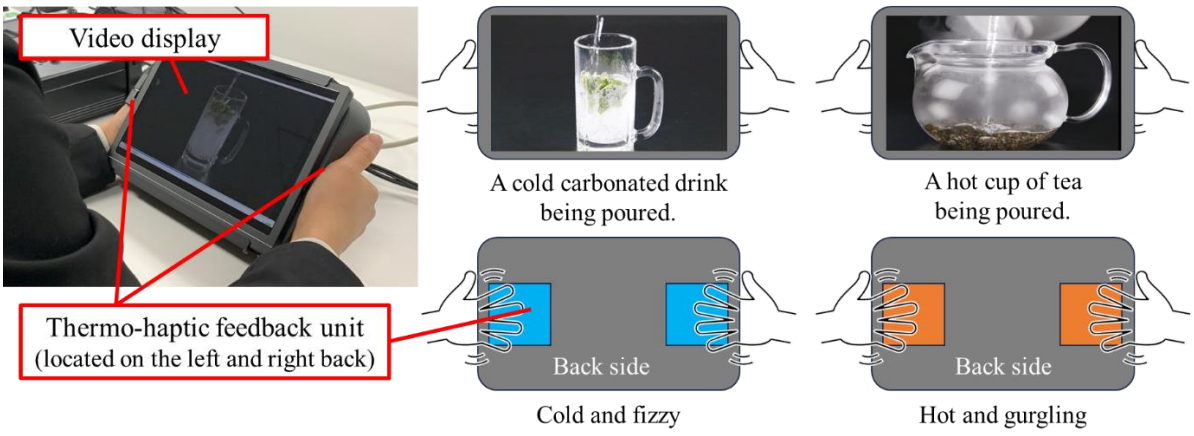
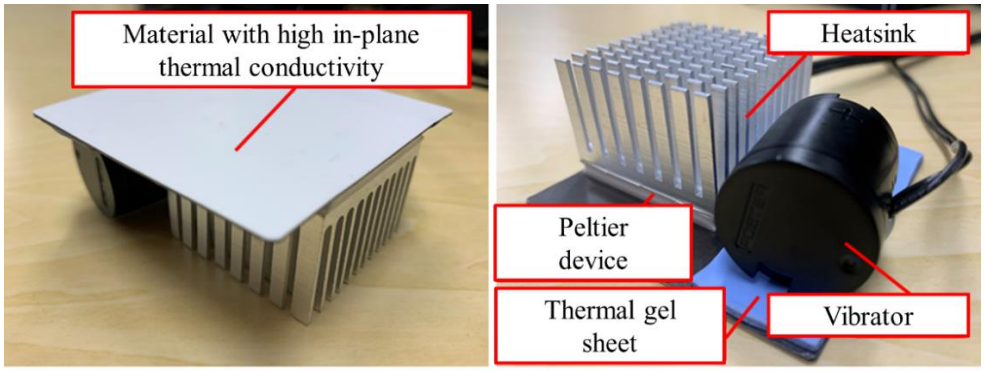


FIGURE 70

Thermo-haptic feedback unit



4.4.8 Linear loudspeaker array system with sound field synthesis technology

4.4.8.1 Overview

Sound field synthesis technology enables reproduction of a numerically described sound field from linear loudspeaker arrays. Synthesizing a sound field in which the sound source is located closer to the user than the loudspeaker array, allows the user to enjoy an auditory experience as if the sound source was actually there (Fig. 71). The virtual sound source with directivity also provides users with different ways of listening depending on their viewpoint. NHK developed a sound system to reproduce sound content in which virtual sound sources with directional characteristics appear to leap out from a screen and move around the user. The sound field reproduction technology called Spectral Division Method, which is based on the spatial Fourier representation of sound field in Cartesian coordinates, was used with modifications to reproduce virtual sound sources with directivity defined in spherical coordinates and to take into account the dynamic characteristics of the sound field with respect to the movement of the sound source.

The linear loudspeaker array consists of 64 loudspeakers arranged at 49 mm intervals, with two rows of loudspeakers installed in the front and rear, and the user listens to the content in between. The specifications of the linear loudspeaker array are shown in Table 19. The front array is used to synthesize a virtual sound source that leaps out in front of the user, and the rear array is used to synthesize a virtual sound source behind the user. The virtual sound sources move on the plane at the same height as the linear loudspeaker arrays.

FIGURE 71

Sound field synthesis technology using linear loudspeaker array

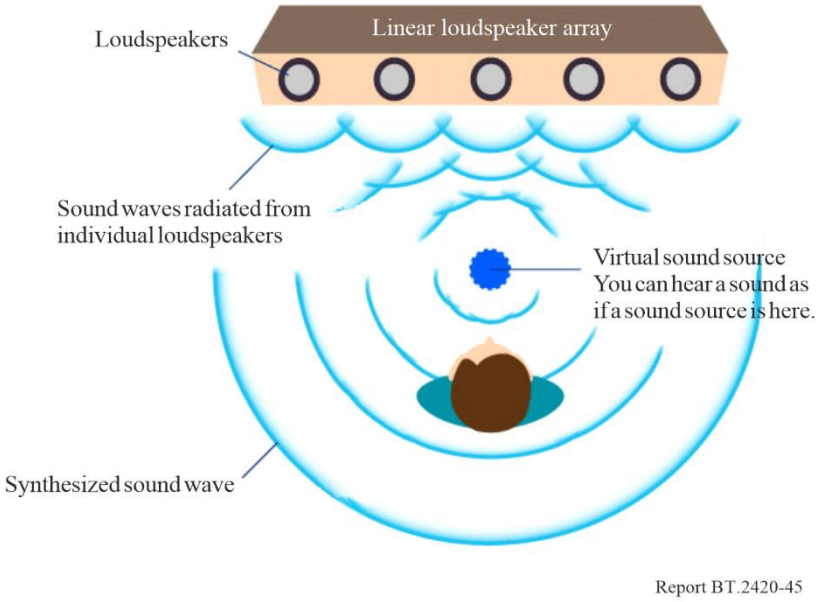


TABLE 19

Specifications of linear loudspeaker array

Diameter of a loudspeaker	1 inch
Number of loudspeakers	128 (64 at each side)
Intervals of each loudspeaker	49 mm
Length of the array	3.2 m

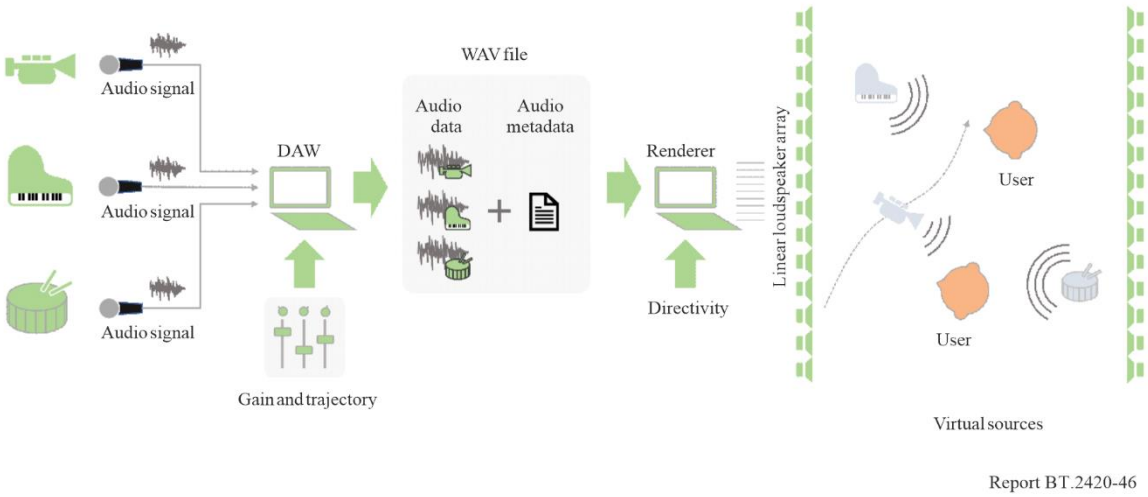
4.4.8.2 Content production and reproduction for moving virtual sound sources

Content with multiple virtual sound sources was produced where each sound source with complex directivity moves individually. The content consists of 28 audio objects, 18 audio objects move around users and 10 are equally spaced on the front and rear linear loudspeaker arrays as static virtual sources to play background music.

The sounds of static instruments were recorded as audio objects, and the trajectories of the audio objects were described using Audio Definition Model (ADM) metadata. The audio data and audio metadata were packed in Waveform Audio File Format (WAV) file. The sound field synthesis renderer generates the audio signals to be produced by the linear loudspeaker arrays so that the virtual sound sources are reproduced at the position described in the ADM for each audio object. The direction and directivity of the sound source were given to the renderer since ADM did not support them (see Figs 72 and 73).

FIGURE 72

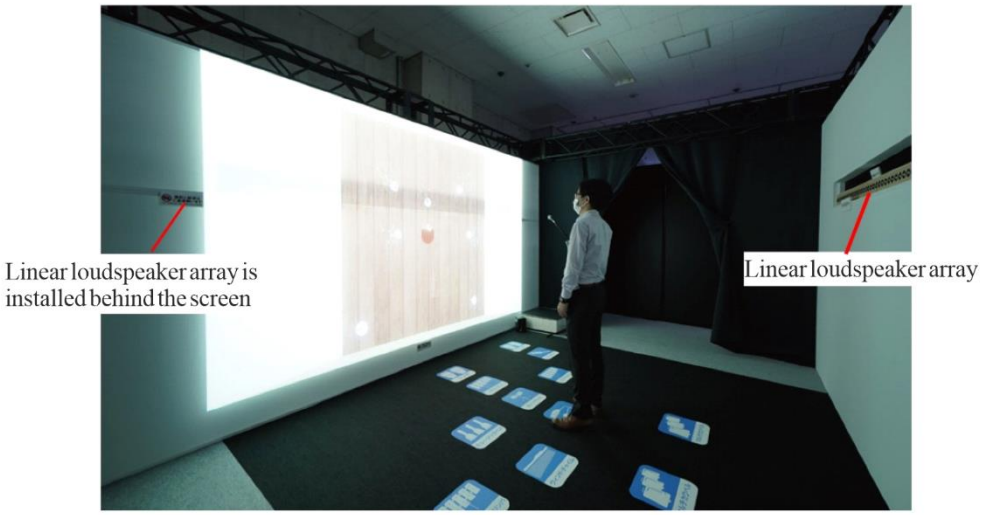
Audio content production and reproduction



Report BT.2420-46

FIGURE 73

Listening room equipped with linear loudspeaker arrays



Report BT.2420-47

4.5 Feedback of VR/AR trials

4.5.1 Feedback from producers and users

– Users experienced a high sense of presence (immersion) and overall positive reaction to content in the trials. This was associated with a reported positive experience from being able to freely choose where they directed their interaction with the content. However, some voiced a desire for higher image quality and image resolution. In addition, it is necessary to study and address the effects of VR/AR viewing on motion sickness as well as younger viewers.

– Production: In terms of lessening motion sickness and improving immersion, it is necessary to consider changes to the production process. This includes modification in camera placement and camera resolution. Since it may become necessary for HMDs to have an age limit, it will, in turn, become necessary to develop alternative immersive experiences and content that even children can enjoy. The distance between the cameras and the subject is a

critical variable that influences the production success. More study and control of the impact of this variable will help improve the experience and consistency of effective video production. Currently, super-wide-angle lenses are used during video capture. This has required a high-resolution content capture to facilitate angle changes during zooming.

– Receiver: Resources are insufficient for using current receivers to develop and display 360-degree VR video internally. Both CPUs and memory need to be considerably upgraded. The level of specification needed is content dependent. Improved receivers are needed that enable separation of the monitor and processing parts for continued expansion.

4.5.2 Applicability to programme genres

Applications of VR/AR are being considered that take advantage of the respective strengths of news, drama, sports, documentaries, and music. Each of these requires selection of an appropriate camera and recording system. Example relevant experiences benefitting from targeted selection include:

– Disaster reporting, where it is necessary to be able to communicate to the viewer the current surrounding situation.

– Multiple perspectives, such as to allow for multiple angles during live broadcasts of sports and entertainers and for participating in programmes virtually.

– Sports content when the distance from the camera is short. This close distance provides a sense of presence but means small changes in resolution make it harder to understand what is going on. This has limited usage to less detailed content and scenes.

– A form of a second screen for hybrid broadcasting and online distribution.

4.5.3 Is ‘live’ important?

– This depends on the service content. Nonetheless, content is being captured and developed that takes full advantage of the unique characteristics of ‘live’ to provide a high sense of presence.

4.5.4 How might VR technologies impact storytelling?

– During live VR experiences, within a production a user is able to select their individual viewpoint. They are also often able to magnify the content in a particular location. These experiences are in contrast to conventional viewing methods where content creators and producers could rely on having control over viewpoint changes. Historically these perspective modifications happened during identified optimal moments in the content. Allowing user flexibility requires notable changes to standard practices for content recording and production.

4.5.5 Future activities

– VR/AR are already gaining importance for events, and it can be expected that the speed of their popularization will increase as they catch the attention of large companies.

– Current AR/VR systems impose notable inconvenience and isolation to the user. This may inspire marked changes in the design of HMDs.

– Mainstream adoption requires improved development and cost savings in high-performance receivers.

5 Challenges

5.1 Possibilities of AISM

The new possibilities of advanced immersive media created new challenges. Solving these challenges is necessary to enable a fully immersive experience over a long period of time. Figure 74 (DVB Report Virtual reality – prospects for DVB delivery, Nov. 2016) depicts the processing stages of production, broadcast delivery, and consumption of AISM programme material. Each stage brings its own challenges.

FIGURE 74

Processing flow of AISM programme material in a broadcast scenario

Report BT.2420-48

5.2 Production challenges

5.2.1 Format for programme exchange of AISM programme material

Recommendation ITU-R BT.2123 provides video parameter values for 360-degree images in 3DoF applications such as projection mapping type, image resolution, frame frequency, and colorimetry by extending the parameter values for UHDTV and HDR-TV [23]. A standard format for volumetric content is yet to be developed.

Recommendation ITU-R BS.2051 provides sound system parameters for advanced sound systems to support channel-based, object-based, or scene-based input signals or their combination with metadata [13]. The structure of a metadata model that describes the format and content of audio files, called the Audio Definition Model (ADM), is specified in Recommendation ITU-R BS.2076 [12]. The Broadcast Wave 64Bit (BW64) audio file format specified in Recommendation ITU-R BS.2088 can carry multichannel files and metadata [24]. The advanced sound systems supported by ADM and BW64 are considered the most prominent immersive audio solutions in VR applications (e.g. [6]). However, there may be parameters specific to advanced immersive audio content that have not been addressed yet. It should be investigated how the existing specifications can be used or potentially be extended to become the programme production format for linear narrative (and maybe non-linear/interactive) advanced immersive audio content.

5.2.2 Evaluation of quality of AISM experience for broadcast applications

A recommendation on how to evaluate the quality of the AISM experience would be beneficial for the production and distribution of AISM content. For the evaluation of video coding quality of omnidirectional media, a few subjective methods have been proposed (e.g. [25], [26]). Because advanced immersive systems create multisensory experiences through the combination of audio, video, interactivity, and haptics, the QoE of AISM systems might require new QoE assessment methods. Such methods may be based on existing psychophysical methods and possibly also on psychophysiological assessment methods (e.g. [27]).

5.2.3 Production guidelines

VR production tends to be more complex compared to conventional programme production for a number of technical and aesthetic reasons: In a 360-degree video recording, there is nowhere to hide the equipment and production crew from the viewer. Creative solutions during production (e.g. hide

crew behind props) and/or postproduction (e.g. stitching of two 180-degree shots into one 360-degree shot) are necessary. Further, to prevent sensory sickness (see § 5.4.2), quick changes in the video, often used as a cinematic storytelling technique (fast camera movements, fast changes of vantage points, scene switches), can only be used with care. The desire to tell a story under these constraints makes it necessary to find alternative techniques to direct the viewer's attention. Some broadcasters have reported on their learning experience in producing advanced immersive media (e.g. [28], [29], [30]). It has been said that VR storytelling has more in common with stage plays than with cinematic storytelling. Sky VR has published their current production guidelines, provisionally specifying their VR format[2] [31]. The VR guidelines produced by VR Industry Forum, aiming at best practices for VR services, addresses production guidelines for VR audio and video content.

5.2.4 Projection mappings

There are various types of ways to squeeze a 360-degree video image into a format applicable for current video encoders. The most prominent projection map is the ERP map, but there are various other methods (e.g. pyramid mapping, cube mapping). It is unlikely that consumer devices (e.g. HMDs) can support every projection mapping concept which creates the risk that content has to be produced or transcoded for different distribution channels and consumer devices.

5.2.5 Production workflow and authoring tools

For professional content creation, an automated production workflow is necessary for a timely delivery of high-quality content within budget. A studio workflow for linear narrative VR content may consist of stages such as footage ingestion, conversion, stitching, asset creation, audio production, tracking, layout, rendering, reviewing, delivering, and archival. While this workflow may be similar to an HDR video production, the amount of data that is being processed is usually significantly larger. For a non-linear narrative VR content production (i.e. where the content unfolds based on some user interaction), the VR workflow will include concepts and tools know from game development. VR and AR content production also benefits from new authoring tools tailored towards AISM content; for instance, sound design authoring tools to pan and visualize sound sources while watching the video content via HMDs or in 360° video player. It is unclear to which extent the AISM content production workflow can be integrated into the broadcast production workflow and infrastructure. Technical standards for interoperability that simplifies content exchange of raw and mastered VR content could help.

5.2.6 Volumetric video capture

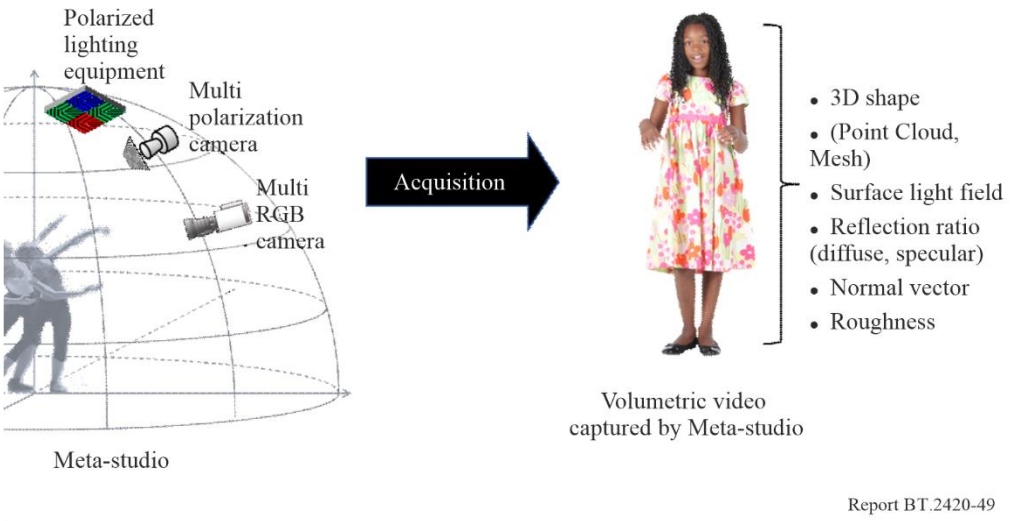
Content creation technologies that capture all available elements of three-dimensional subjects including shapes, surface light field and optical properties of surfaces (reflection ratios, normal vectors, and roughness) are required for content which provides spectators with highly immersive experiences. The concept of 'Meta-Studio' is shown in Fig. 75, which captures not only the shape and colours of the object but all subject information to generate a 3D volumetric object. The use of polarization cameras together with polarized lighting and multiple RGB cameras surrounding the subject makes it possible to acquire optical properties of surfaces such as the roughness and reflection ratio of the object surface. Surface light field information representing viewpoint-dependent textures is essential for creating a 3D object in photo-realistic and natural ways that can be viewed from any virtual viewpoint in high quality.

---

[2] Video: 2-4K resolution, ITU-T H.264|MPEG-4 AVC, 25-50 FPS, 20-60 Mbit/s bitrate; audio: stereo or scene-based audio/ambisonics.

FIGURE 75

Configuration of Meta-Studio and subject information to be acquired



5.3 Delivery challenges

5.3.1 High transmission rates

The delivery of AISM content is the most significant challenge for current distribution systems. The transmission rates for VR content far exceed the data rates of current high-quality broadcast programmes. This stresses the network throughput and the processing capabilities of many consumer devices. Data buffering is often necessary to guarantee a consistent immersive experience.

To deliver high-quality VR content using unicast streaming, various methods are under development. These methods are based on the dynamic switching of view-dependent video streams that are compressed so that the active field of view receives high-quality video, and the periphery receives reduced quality video (see, e.g., [32], [33]). For preventing the perception of low-quality video content during head rotation, the seamless switch from one viewport video stream to another is crucial. For broadcasting of AISM content, those viewport-based solutions may not be feasible, and the entire video feed is broadcasted at once. A proof of concept for the broadcasting of AISM content was demonstrated in 2016 at the IBC: using an ASTRA satellite connection as a distribution path the transmission of a 10K × 2K panoramic video signal to multiple devices was showcased (https://www.vrfocus.com/2016/09/ses-and-fraunhofer-hhi-to-bring-vr-via-satellite-to-ibc).

5.3.2 VR distribution format

There is a standardized distribution format to deliver high-quality immersive linear narrative AISM content. MPEG has standardized the first version of Omnidirectional Media Format (OMAF) as ISO/IEC 23090-2 which defines not only a media format of VR/360° content but also IP-based delivery methods of OMAF content using MMT and MPEG-DASH (see more in § 6). Recommendation ITU-R BT.2133 provides guidance on using ISO/IEC 23090-2 to transport AISM content in IP-based broadcasting systems [34].

5.4 Consumption challenges

5.4.1 General

Many of the technical problems during production and transmission are driven by the needs to deliver high-quality content on the consumer side that (in the best case) exceeds the requirements of current TV broadcasting. For instance, many experts suggest that a minimum of 60 frames per second and bi-ocular 4K video images are necessary to enable enjoyable high-quality VR experiences over

HMDs (see also Sky's production guidelines [31]). The acceptable motion-to-sound and motion-to-photon latencies are other important technical parameters that contribute to both the realism and prevention of sensory sickness. Pointer devices or hand gesture trackers are also important for interactivity within the virtual world. These must be frictionless, intuitive, and reliable human-computer interface devices. On top of these demanding technical requirements, consumer studies show that there can be aesthetic concern with the design of current HMDs [35]. As stated in [36] current VR/AR content is inconsistently rendered across HMD devices leading to different VR/AR experiences. New standards to guarantee a quality of experience may be needed.

5.4.2 Sensory sickness (kinetosis)

Currently created VR content has mostly a duration of less than 15 minutes. This can be linked to the desire to avoid triggering sensory sickness (i.e. kinetosis) that may occur when using HMDs for a longer duration, especially with high motion content. Despite years of research, preventing sensory sickness is a hard problem for VR consumption. It is known to occur as a function of several technical visual rendering parameters (motion-to-photon latency, frames per second, flicker of the displayed view, display width) and human aspects (duration of exposure, personal sensitivity, motion control, general health, genetic background, gender, age, mood, anxiety, postural stability). A detailed discussion on sensory sickness can be found in [2, section 10].

5.4.3 Content presentation

Compared to conventional television and cinema content, there is currently only a small amount of professional VR content available. Consequently, on-demand VR services aim to enhance their content offerings by adding 2D content that can be watched in VR. For instance, Netflix, Hulu and HBO are offering VR environments in which the user is placed in a virtual home cinema where the streamed 2D content is presented on a large virtual 2D-rectilinear. The content is adapted to the user's head motion. The audio content over headphones can be the original stereo mix or the binauralization of the original spatial sound mix.

On the flip side, content creators may want to repurpose VR content and make it available on traditional television. This scenario requires a workflow where a view of the VR video is rendered to traditional TV formats prior to broadcasting and presented with the correctly spatially aligned immersive audio. The German broadcaster rbb tested such solution for regular SmartTVs (https://www.fokus.fraunhofer.de/en/fame/news/FAME_BandCamp).

5.4.4 VR consumer platforms

A recent user experience report of 10 popular VR applications was published by VR testing company Fishbowl VR and concluded that many VR services are not capable to deliver high-quality VR experiences. Users were asked to test and rate VR applications over HMDs in the categories UI, content variety and quality, picture quality, virtual environment, and loading times. Additionally, they were asked to rate their likelihood of recommending the app to a friend. The testers found immersive video to be enjoyable in powerful and unique new ways; however, current VR delivery platforms still need improvements. Only for two tested apps, users consistently said they have the quality to eventually displace regular TV, laptop or mobile viewing. These two apps were offering both original 360-degree content but also traditional 2D programme material via a large 2D-rectilinear screen in the virtual world. Technical issues such as poor video quality and the lack of interactivity as well as social engagement degraded the overall VR experience. Users mostly desire a way to participate in VR video experiences together with friends and family. Content buffering, which occurs frequently in the majority of applications, led to very poor user experiences.

6 Work of ITU-T on virtual reality

6.1 ITU-T Study Group 16

In June 2016, ITU-T Study Group 16 established a new Question 8/16 regarding Immersive Live Experience. In January 2017, ITU-T SG 16 hosted its second workshop on Immersive Live Experience (ILE). The objective was for participants to exchange information related to immersive services and technologies between several organizations and to identify standardization gaps. The workshop featured various presentations on VR, AR and related technologies:
http://www.itu.int/en/ITU-T/studygroups/2017-2020/16/Pages/ws/201701_ILE.aspx

6.2 ITU-T Study Group 12

At the ITU-T Study Group 12 meeting in January 2017, new work items "QoE for Virtual Reality (G.QoE-VR)" and "Subjective test methodologies for 360 degree video on HMD (G.VR-360)" were created under Question 13. These work items are intended to lead to several recommendations regarding QoE factors, QoE/QoS requirements, subjective test methodology, and objective quality estimation models for virtual reality (VR) services [37].

According to the baseline text for G.QoE-VR, the scope of the new Recommendation is as follows:

Virtual Reality (VR) is a new type of media different from the traditional video and audio media. It generates realistic images, sounds and other sensations that replicate a real environment, and simulates a user's physical presence in this environment, by enabling the user to interact with this space and any objects depicted therein using specialized display screens or projectors and other devices. The multi-sensory experiences, which can include sight, touch, hearing, and, less commonly, smell, are well coordinated and synchronized through the user's interaction and feedback. A person using virtual reality equipment is typically able to "look around" the artificial world, move about in it and interact with features or items that are depicted on a screen or in goggles as in the real world.

In order to understand whether QoE or user-perceived performance of the VR service is good or not, benchmarking is critical. This allows measurement of user-perceived performance or QoE in that environment. Compared with traditional video and audio, the multi-sensory experience in VR imposes a new set of requirements to QoE assessment. The challenge is to characterize VR's real-life immersive video, spatial-audio, and interactivity. Before benchmarking the QoE, it is important to address the requirements and basic factors assessing the VR quality for different VR services.

This draft Recommendation identifies different VR services and their respective requirements for Quality of Experience (QoE). This document also summarizes the key factors affecting user-perceived experience of a VR service, which can help to identify the methodologies for assessing the VR quality.

The scope of this Recommendation includes:

– VR service categories.
– QoE requirements for VR services.
– Categorization of influence factors.

7 Activities of other SDOs and VR groups

7.1 Activities of other SDOs

Many SDOs and industry groups have already started working or exploring subareas of AISM systems. The following list gives a best-effort overview of ongoing activities:

MPEG: The Motion Picture Expert Group (MPEG) is developing video and audio codecs, and transport and systems aspects with regards to VR and AR applications.

An important activity for AISM content exchange and delivery is the Omnidirectional Media Format (OMAF). OMAF is the first standard from MPEG providing formats for the support of immersive media. It specifies the application format for coding, storage, delivery, and rendering of omnidirectional images and video and the associated audio. The Final Draft International Standard (FDIS) of OMAF was produced in October 2017 as ISO/IEC 23090-2 (Part 2 of MPEG-I).

MPEG-H 3D Audio (ISO/IEC 23008-3:2017) is a next generation audio codec featuring highly efficient compression of channel-based audio, object-based, and scene-based immersive audio. For AISM applications, it specifies a 3DoF decoder-side sensor interface and normative specification for diegetic and non-diegetic audio processing.

In Q3 2016, MPEG conducted an informal survey to better understand the needs for standardization in support for VR applications and services. They received 185 responses and summarized their results and conclusions [6]. Based on this survey, MPEG concluded to launch a new project on immersive media (MPEG-I). A first set of specifications that defines up to 3DoF 360° VR was finalized as OMAF to support market launches of products and services in 2018. Based on a common belief that major market launch of VR 360 services will happen in 2020, a next set of specifications including a richer feature set (such as 6DoF) may be ready in 2019.

The Joint Video Exploration Team of MPEG and ITU (JVET) is working on the future H.266 video codec for which they (among other things) study the effect on compression when different warping methods are applied to the input 360° video before compression. On this basis, JVET has defined common test conditions, test sequence formats, and evaluation criteria for such content (http://www.content-technology.com/standards/?p=740).

JPEG: The Joint Photographic Experts Group (JPEG) is developing JPEG XT, an image format which features coding of omnidirectional (360-degrees) images; JPEG XS, a low latency compression formats for VR videos; and JPEG PLENO, a video format for point cloud, light field, and holographic images.
https://jpeg.org

DVB: In 2016 DVB carried out a study mission to determine the likelihood VR video will be commercially successful, and to find out how DVB can be involved. The study mission produced a detailed report [2]. An executive summary of this report is freely available [5].

The DVB study mission concluded that at least in the near term for broadcast use cases, untethered devices supporting 3DoF (such as slide-on HMDs) are more likely to be commercially successful than tethered devices. The dominant success factors of these VR devices and services are quality of experience (QoE), lack of sensory sickness, comfort and ease of use, cost and availability of equipment, cost and availability of content, and content desirability.

Further, it was concluded that DVB should cooperate with standards bodies working in VR, as members will need to adopt common specifications for delivery of VR content. Requirements are needed for the minimum technical quality of VR video and audio. Requirements should be completed by mid-2018. A questionnaire on VR broadcasting services has been circulated to further inform the process of defining commercial requirements. DVB is informing various standardization groups about their findings and their planned activities (e.g. MPEG, ITU-R, ITU-T). The DVB study mission continues to address topics such as AR, MR, and 6DoF VR.

3GPP: 3GPP's subgroup SA4 conducted a feasibility study on virtual reality media services over 3GPP. The technical report can be accessed online [3]. Further, the subgroup SA1 specifies service requirements for the 5G system which includes aspects related to support various VR and AR use cases.

SMPTE: Specifications on production for VR content are in the scope of SMPTE. The 2016 SMPTE conference had presentations in this area (e.g. [37], [38]).
http://www.tvtechnology.com/events/0025/smpte-how-can-you-ensure-an-effective-vrar-experience/279732

W3C: The W3C community is working toward standardizing WebVR, a JavaScript-based API that provides access to VR devices, sensors, and head- mounted displays through web browsers. At the time of writing this report, it was unclear if and when the W3C standards track adopts WebVR.
https://w3c.github.io/webvr
https://www.w3.org/2016/06/vr-workshop

IEEE: Under the umbrella of the Virtual Reality and Augmented Reality Working Group (VRAR) the IEEE launched standardization activities in the area of taxonomy and definitions for VR, AR devices; quality metrics for immersive video; and VR file and streaming formats. The IEEE is also organizing the IEEE VR conference series, one of the oldest academic consortia dedicated to the study of virtual reality from a science and engineering perspective.
https://standards.ieee.org/develop/wg/VRAR.html

End-to-end interoperability

CTA: In November 2016, the Consumer Technology Association (CTA) released a survey report [38] entitled "Augmented Reality and Virtual Reality: Consumer Sentiments". The study aimed to understand consumer's awareness and opinions regarding AR and VR technologies and gathered consumer feedback about envisioned VR/AR use cases. A summary of this report was presented at the CES 2017 and provided three recommendations. First, consumers need more education about the difference of VR vs. AR; VR device choices; and content delivery options. Second, to expedite consumer excitement and to avoid stereotyping VR technology, more diverse VR demos need to be produced and presented to the mass market. Third, to correct miscues or misuse of devices, the industry should encourage customer feedback of VR solutions.

Others: The **IETF** (Internet Engineering Task Force) explores mechanisms to carry high bandwidth media such as VR with low latency over the Internet, **Qualinet** has a task force for VR and immersive experiences (IMEx), and **Cable Labs** studies the requirement for support of VR in cable networks. **ARIB**, the Association of Radio Industries and Businesses published anticipated quality requirements and transmission rates for VR application over 5G networks [39].

7.2 Activities of VR industry groups

Multiple industry groups have been formed to promote the development and growth of the VR/AR industry.

VRIF: The VR Industry Forum (VRIF) is a not-for-profit company with the purpose to further the widespread availability of high-quality audio-visual VR experiences, for the benefit of consumers. Founded on 5 January 2017, the goals of VRIF include:

– Advocating voluntary industry consensus around common technical standards for the end-to-end VR ecosystem, from creation to delivery and consumption.
– Advocating the creation and adoption of interoperable standards; promoting the use of common profiles across the industry and promoting and demonstrating interoperability.
– Developing voluntary guidelines that describe best practices, to ensure high-quality VR experiences.

Describing and promoting the use of VR services and applications.

VRIF has also established a lexicon of VR terminology to encourage common usage of term and avoid misuse of terms causing confusion.
http://www.vr-if.org

VR Society: The VR Society was officially launched on July 13th, 2016. Its purpose is to accelerate the transformation, innovation, and profitability of the Virtual Reality Content, Distribution and Technology Business. Further, the society promotes and fosters a marketplace of ideas based on thought leadership, sharing of best practices, marketing, consumer research, industry analytics and advocacy. Together with the Advanced Image Society, the VR Society bestows the Lumiere Awards (the 2017 edition includes expansive categories for VR content).
http://thevrsociety.com

VRARA: The VR/AR Association (VRARA) is a global industry association founded in 2015 that aims to offer a connected local and global community of members through its initiatives. The VRARA is focused on creating a member community of VR/AR solution providers, content creators, and customers. The objective is to accelerate networking and sharing of knowledge through case studies. There are different industry committees such as VR Stories & Audience, Mobile VR & 360 Video, Entertainment, and Advertisement that strive to establish best practices, guidelines, and call-to-actions (e.g. recommendations for standards) in their areas of expertise.
http://www.thevrara.com

Khronos: On December 6th, 2016, Khronos, an open consortium of hardware and software companies, started an initiative to define a cross-vendor, royalty-free, open standard for access to modern virtual reality (VR) devices. Key components of the new standard will include APIs for tracking of headsets, controllers and other objects, and for easily integrating devices into a VR runtime. This will enable applications to be portable across VR systems, enhancing the end-user experience, and driving more choice of content to spur further growth in the VR market.
https://www.khronos.org/vr

SVA: The Streaming Video Alliance (SVA) is a consortium of organizations spanning the streaming video value chain. In November 2016, it announced a study group for VR and 360-degree video. In this group, participating members and selected outside parties seek to understand and document the VR industry in an effort to identify opportunities for developing best practices. The group's objective is to:

– Understand the VR market and how it is impacting traditional video experiences.
– Capture the state of VR technologies, the players, and use-cases.
– Catalogue existing standards efforts.

Digital Senses Alliance: The Digital Senses Alliance is an Industry Connections programme about to be formed by the IEEE Digital Senses Initiative which aims for cross-industry and cross-disciplinary collaborations to identify gaps in technologies and standards in the area of VR, AR, and Human Augmentation (HA). Further, it plans to provide training facilities and learning resources for how to create VR and AR content.
http://digitalsenses.ieee.org

OSVR: The Open Source VR group (OSVR) promotes a universal open source VR ecosystem for technologies across different brands and companies to avoid hardware fragmentation of HMDs and controllers and other challenges in the industry.
https://osvr.github.io/whitepapers/introduction_to_osvr/

DASH-IF: The Industry Forum for dynamic adaptive streaming over HTTP (DASH-IF) hosted a workshop on "Streaming Virtual Reality with DASH" in May 2016.
http://dashif.org

AES: The Audio Engineering Society has just formed a new technical group to advance the science and application of Audio for New Realities (AR/VR/MR). The initial objectives will include collating the state of the art in audio for new realities across recording, composition, sound design, spatial audio, environmental analysis and auditory scene synthesis, in order to develop technical workflows

that are practical and relevant to the industry and creative practitioners in the field. The AES also organized one of the first conferences purely on Audio for VR/AR:
http://www.aes.org/conferences/2016/avar

Bibliography

[1] P. Lelyveld. (2015, July) *Virtual Reality Primer with an emphasis on camera-captured VR*. [Online]. Available at:
http://www.etcenter.org/wp-content/uploads/2015/07/ETC-VR-Primer-July-2015o.pdf

[2] DVB. (2016, October) *DVB Virtual Reality – prospects for DVB delivery*, report of the DVB CM study mission on virtual reality study.

[3] 3GPP SA4. (2018, January) *Virtual Reality (VR) media services over 3GPP*. [Online]. Available at:
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3053

[4] H. Nagata, D. Mikami, H. Miyashita, K. Wakayama, and H. Takada, "Virtual reality technologies in telecommunication services", Journal of Information Processing, vol. 25, pp. 142-152, 2017.

[5] DVB. (2016, November) *Executive summary DVB study mission on virtual reality*. [Online]. Available at:
https://dvb.org/wp-content/uploads/2020/01/dvb_vr_study_mission_report_summary.pdf

[6] MPEG. (2016, October) *N16542* – summary of survey on virtual reality. Chengdu, China. [Online]. Available at:
https://dms.mpeg.expert/
http://mpeg.chiariglione.org/sites/default/files/files/standards/parts/docs/W16542%20MPEG%20VR%20Questionnaire%20Results%20Summary.pdf

[7] H. Kawakita, K. Yoshino, D. Koide, K. Hisatomi, Y. Kawamura, and K. Imamura, "AR/VR for various viewing styles in the future broadcasting", IBC2019.

[8] F. Williams. (2017, February) *OxSight uses augmented reality to aid the visually impaired*. [Online]. Available at:
https://techcrunch.com/2017/02/16/oxsight-uses-augmented-reality-to-aide-the-visually-impaired

[9] Outlyer Technologies. (2017) *360 / VR test ad campaign*. [Online]. Available at:
https://advrtas.com/?s=VR

[10] H. Kawakita, T. Niida and K. Hisatomi, "Development of a 360 Video Player for Head-Mounted Display to Enable Comparison Between Before and After Incident", Forum on Information Technology 2020 (FIT2020), Part III, K-010, pp. 361-363. (2020).

[11] D. Koide, H. Kawakita, K. Yoshino, K. Ono, K. Hisatomi, "Development of High-Resolution Virtual Reality System by Projecting to Large Cylindrical Screen", IEEE International Conference on Consumer Electronics (ICCE), 1.16 AVS (4)-1, 2020.

[12] Recommendation ITU-R BS.2076-2 (10/2019) *Audio Definition Model*. [Online]. Available at:
https://www.itu.int/rec/R-REC-BS.2076/en

[13] Recommendation ITU-R BS. 2051-2 (07/2018) *Advanced sound system for programme production*. [Online]. Available at:
https://www.itu.int/rec/R-REC-BS.2051/en

[14] Adrian Freed and Andy Schmeder, "Features and Future of Open Sound Control version 1.1. for NIME", In Proc. NIME 2009, pp. 116-120.

[15] Y. Kawamura, T. Kusunoki, Y. Yamakami, H. Nagata and K. Imamura, "Toward tele-experience: Enhanced viewing experience by synchronized UHDTV and free-viewpoint AR", IBC2020.

[16] Wavefront Technologies, Appendix B1. Object Files (.obj), Advanced Visualizer Manual.

[17] Frank Galligan. (2017, Oct) *Draco Bitstream Specification*. [Online] Available at:
https://google.github.io/draco/spec/

[18] K. Yoshino, H. Kawakita, T. Handa and K. Hisatomi, "Viewing Style of Augmented Reality/Virtual Reality Broadcast Contents while Sharing a Virtual Experience", 26th ACM Symposium on Virtual Reality Software and Technology (VRST '20), Article 76, 2020.

[19] T. Handa, M. Azuma, T. Shimizu, S. Kondo, M. Fujiwara, Y. Makino and H. Shinoda: "Ball-type Haptic Interface to Present Impact Points with Vibrations for Televised Ball-based Sporting Event", IEEE World Haptics Conference (WHCs), TP1A.14, pp. 85-90, 2019.

[20] M. Azuma, T. Handa, T. Shimizu, and S. Kondo: "Development of Vibration Cube to Convey Information by Haptic Stimuli", Proceedings of the 24th International Display Workshops, Vol. 24, pp. 128-130, 2017.

[21] M. Takahashi, M. Azuma, T. Handa, T. Ishiwatari, M. Sano and Y. Yamanouchi: "Real-time Sports Video Analysis for Video Content Viewing with Haptic Information," Proceedings of the ACM SIGGRAPH 2021, Poster, DOI: 10.1145/3450618.3469135, 2021.

[22] M. Azuma, T. Handa, and K. Komine: "OTOGI BOX: A Cubic Haptic Interface for Presenting a Story through Tactile Communication", Proceedings of the 30th International Display Workshops, Vol. 30, pp. 719-722, 2023.

[23] Recommendation ITU-R BT.2123-0 (01/2019) *Video parameter values for advanced immersive audio-visual systems for production and international programme exchange in broadcasting*. [Online]. Available at: https://www.itu.int/rec/R-REC-BT.2123/en

[24] Recommendation ITU-R BS.2088-1 (10/2019) *Long-form file format for the international exchange of audio programme materials with metadata*. [Online]. Available at: https://www.itu.int/rec/R-REC-BS.2088/en

[25] E. Upenik, M. Rerabek, and T. Ebrahimi, "A testbed for subjective evaluation of omnidirectional visual content", in 32nd Picture Coding Symposium, no. EPFL-CONF-221560, 2016. [Online]. Available at: https://infoscience.epfl.ch/record/221560/files/Testbed_for_omnidirectional_content_camready.pdf

[26] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes", in Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on. IEEE, 2015, pp. 31-36. [Online]. Available at: https://pdfs.semanticscholar.org/b3c9/a67d5c1aadcabf6c6c25ce9729f077c9f302.pdf

[27] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, "Psychophysiology-based QoE assessment: A survey", IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pp. 6-21, Feb 2017.

[28] J. Bloch. (2016, Oct) *5 secrets to making a virtual reality film*. [Online]. Available at: http://www.cbc.ca/news/canada/british-columbia/highway-of-tears-vr-doc-1.3800490

[29] M. Burns. (2017, March) *An exploration of VR and AR for sports: part i – the challenge facing immersive sports*. [Online]. Available at: http://www.svgeurope.org/blog/headlines/an-exploration-of-vr-and-ar-for-sports-part-i-the-challenge-facing-immersive-sports

[30] Streamshark.io. (2016, April) Ca*se study: Live streaming of the Australian Open in 360°*. [Online]. Available at: https://streamshark.io/blog/aus-open-360

[31] Sky. (2016, December) *Launch technical guidelines for 360 video content*. [Online]. Available at: https://static.skyassets.com/contentstack/assets/bltdc2476c7b6b194dd/blt7aa6b8cd22755a07/5e4eb79506f84d0d618d9160/launch-technical-guidelines-360-video-content.pdf

[32] E. Kuzyakov. (2017, January) *End-to-end optimizations for dynamic streaming*. [Online]. Available at: https://code.facebook.com/posts/637561796428084/end-to-end-optimizations-for-dynamic-streaming

[33] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC- compliant tile-based streaming of panoramic video for virtual reality applications", in Proceedings of the 2016 ACM on Multimedia Conference, ser. MM '16. New York, NY, USA: ACM, 2016, pp. 601-605. [Online]. Available at: http://doi.acm.org/10.1145/2964284.2967292

[34] Recommendation ITU-R BT.2133-0 (10/2019) *Transport of advanced immersive audio visual content in IP-based broadcasting systems*. [Online]. Available at: https://www.itu.int/rec/R-REC-BT.2133/en

[35] Perkins Coie LLP and Upload. (2016, September) *2016 augmented and virtual reality survey report*. [Online]. Available at: https://dpntax5jbd3l.cloudfront.net/images/content/1/5/v2/158662/2016-VR-AR-Survey.pdf

[36] P. Routhier, "Virtually perfect: Factors affecting the quality of a VR experience and the need for a VR content quality standard", in SMPTE 2016 Annual Technical Conference and Exhibition, Oct 2016, pp. 1-20. [Online]. Available at: http://www.sportsvideo.org/2016/10/28/smpte-2016-qa-digital-troublemakers-pierre-routhiers-13-rules-for-quality-vr

[37] Recommendation ITU-T G.1035 (05/2020) *Influencing factors on quality of experience for virtual reality services*. [Online]. Available at: https://www.itu.int/rec/T-REC-G.1035/en

[38] M. L. Champel and S. Lasserre, "The special challenges of offering high quality experience for VR video", in SMPTE 2016 Annual Technical Conference and Exhibition, Oct 2016, pp. 1-10.

[39] Association of Radio Industries and Businesses. (2014, October) *Mobile communications systems for 2020 and beyond*. [Online]. Available at: http://www.arib.or.jp/english/20bah-wp-100.pdf