International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BT.2154-0

(12/2022)

High-level system architecture for immersive video for presentation on various types of display devices

BT Series

Broadcasting service (television)

ITU International Telecommunication Union

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 http://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 Recommendations

(Also available online at http://www.itu.int/publ/R-REC/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	Radiowave 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
SNG	Satellite news gathering
TF	Time signals and frequency standards emissions
V	Vocabulary and related subjects

***Note**: This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2022

RECOMMENDATION ITU-R BT.2154-0

High-level system architecture for immersive video for presentation on various types of display devices

(Questions ITU-R 140-1/6 and ITU-R 143-2/6)

(2022)

Scope

This Recommendation provides a high-level system architecture for immersive video to be presented on various types of display devices. The architecture consists of video objects, scene description, renderer, and player as a minimum set of components. This Recommendation also identifies information to be transferred between the renderer and player.

Keywords

Immersive video, 6DoF, scene description, volumetric video, device adaptation

The ITU Radiocommunication Assembly,

considering

a) that immersive video, which enables end users to move around in a video space and watch video omnidirectionally from free viewpoints, provides a new enhanced visual experience;

b) that immersive video is represented by arranging video objects such as volumetric video, omnidirectional video and two-dimensional video in a three-dimensional space;

c) that various types of display devices are available for end-users such as head-mounted displays, smartphones and tablets, needing to be considered for presenting immersive video;

d) that an increasing number of interactive delivery platforms are available for distributing content to audiences;

e) that servers on networks including cloud and edge with increased computational power will effectively be used for rendering immersive video adapting to different types of display devices with different computing and display capabilities;

f) that a common architecture for immersive video to be presented on various types of display devices will facilitate the development of immersive video systems and applications,

recognizing

a) the suite of standards ISO/IEC 23090 – Information technology – Coded representation of immersive media;

b) Recommendation ITU-R BT.2123 – Video parameter values for advanced immersive audio-visual systems for production and international programme exchange in broadcasting;

c) Report ITU-R BT.2420 – Collection of usage scenarios of advanced immersive sensory media systems,

recommends

that immersive video systems targeting various types of display devices should be designed according to the high-level system architecture described in the Annex.

Annex

High-level system architecture for immersive video for presentation on various types of display devices

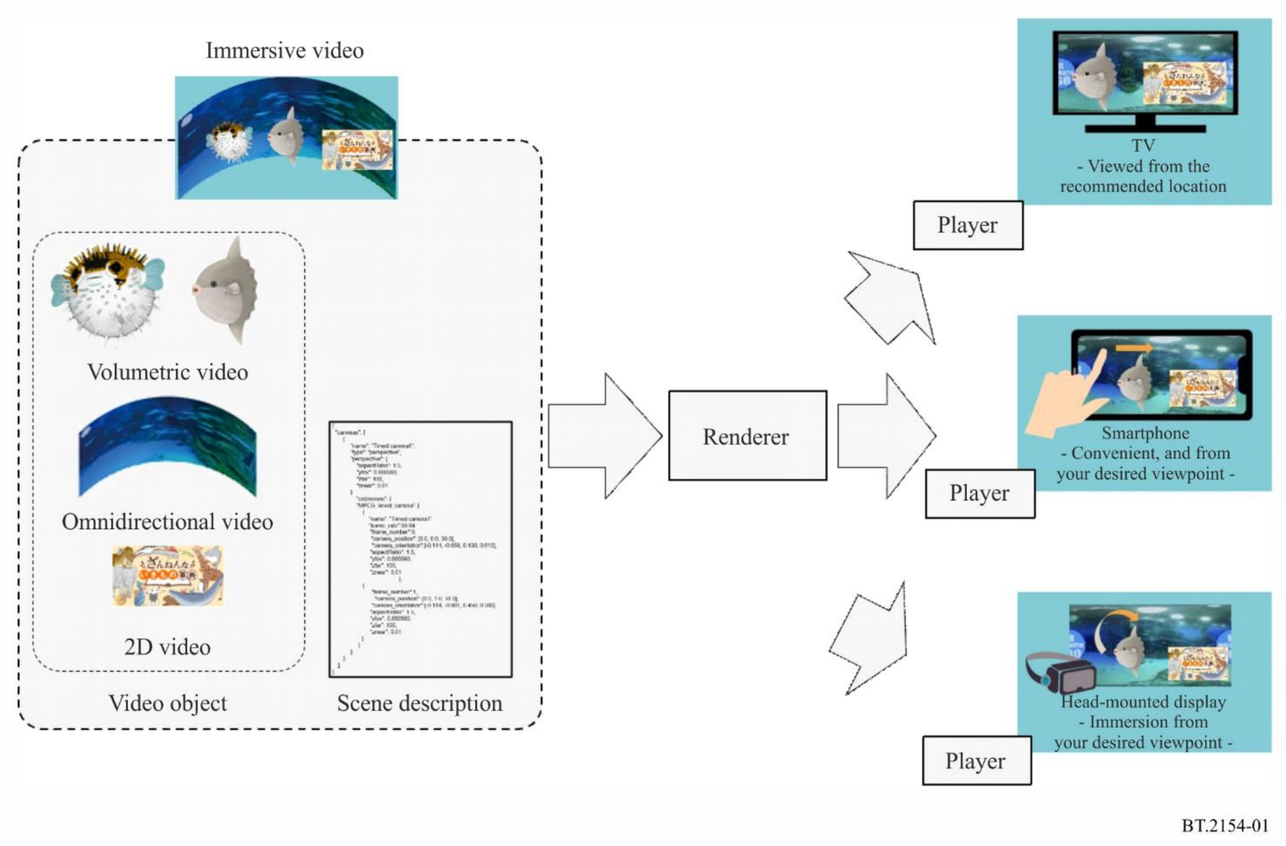
1 Overview

Figure 1 illustrates an overview of the system architecture for immersive video from composition to presentation on various types of display devices.

Immersive video is composed of scene description and multiple video objects referred from the scene description including volumetric video, which can represent the three-dimensional shape and texture of objects, omnidirectional video surrounding the objects, and two-dimensional video so that users can watch the video from any positions in any directions. Omnidirectional video and 2D video may have their depth information. Scene descriptions provide a time-series three-dimensional representation of immersive video such as position, orientation, and size of each object as well as their spatial and temporal arrangement in three-dimensional space.

FIGURE 1

Overview of immersive video from composition to presentation



A renderer constructs a three-dimensional space from various video objects indicated by the scene description and produces video to be viewed from the user's position and direction. The player of each display device presents the video from the rendered video in a way that best suites the device in accordance with user's viewing position and direction.

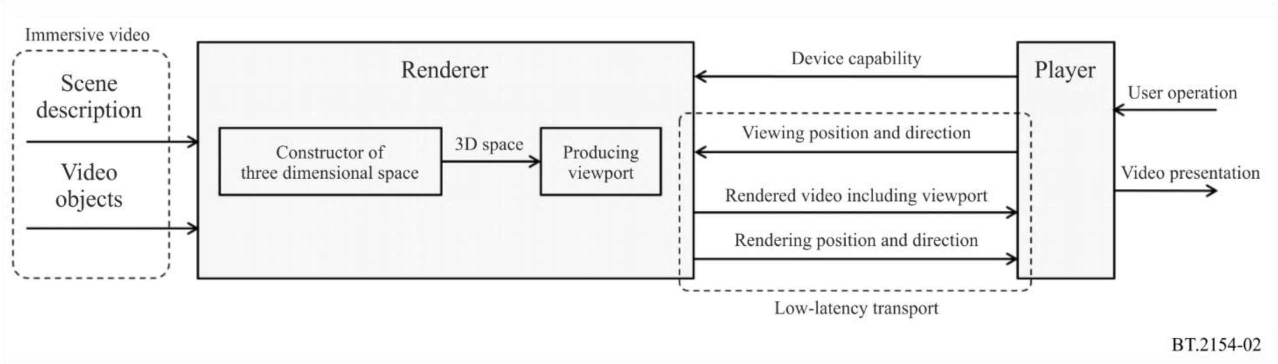
2 High-level system architecture

2.1 Definition

The high-level system architecture depicted in Fig. 2 is defined to present immersive video on various types of display devices.

FIGURE 2

High-level system architecture for immersive video



2.2 Immersive video

The immersive video represents a time-series three-dimensional space and is composed of video objects and scene description.

The video objects include volumetric video that represents the three-dimensional shape and texture of objects, omnidirectional video that surrounds the objects, and two-dimensional (2D) rectangular video. Omnidirectional video and 2D video may be associated with depth information.

The scene description defines a time-series three-dimensional space by referring to the multiple video objects and identifying position, orientation, and the size of each object as well as their spatial and temporal arrangement in three-dimensional space.

The scene description may also include information on the user's viewing position and viewing direction recommended by the content creator, i.e. recommended viewport, for the presentation on the display device.

2.3 Renderer and player

A renderer constructs a three-dimensional space from various video objects and the scene description. It also produces video to be viewed from the user's position and direction.

The player presents the video from the rendered video in a way that best suites the device in accordance with user's viewing position and direction.

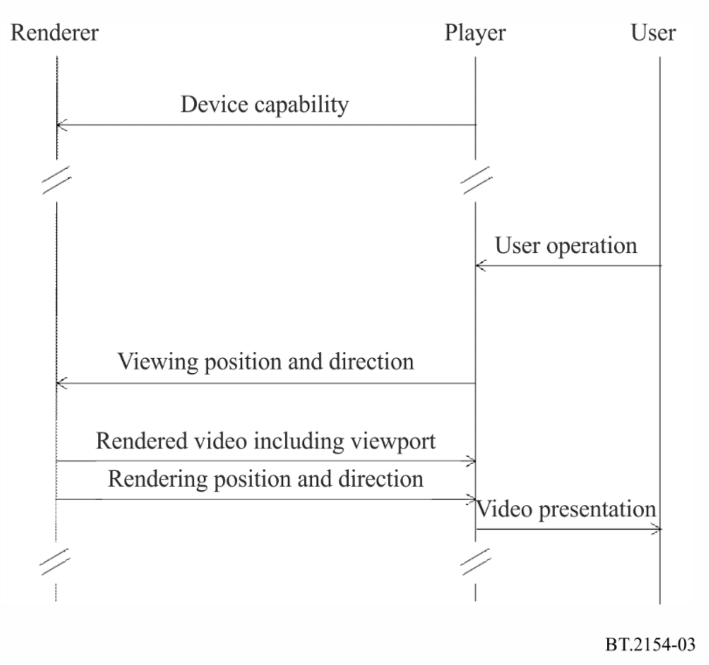
The functions of the renderer and the player should be separated so that the player does not need to process all the information in the three-dimensional space but only has to process a portion that needs to be presented on the device. The separation reduces the data size of the video objects to be processed by the player and processing load of the player, resulting in lighter processing players to be implemented. Even when additional types of video objects are to be introduced in the future, only the renderer needs to be updated to support them without updating the player.

2.4 Information to be transferred between renderer and player

Figure 3 shows the flow of information to be transferred between the renderer and the player.

FIGURE 3

Information to be transferred between renderer and player


Renderer
Player
User
Device capability
User operation
Viewing position and direction
Rendered video including viewport
Rendering position and direction
Video presentation
BT.2154-03

1 Before starting the rendering process, the render constructs a three-dimensional space on the basis of the scene description and the video objects, and the player notifies the renderer of the device capability including display resolution, field of view, and frame rate.

2 When a user starts viewing the content, the player notifies the renderer of user's viewing position and direction, which may change during watching the video, in accordance with the user's operation.

3 From the three-dimensional space, the renderer produces video including viewport to be presented according to the user's viewing position and direction notified. The renderer may produce video in a wider range of areas than the viewport to support the rapid movement of the viewing position and direction. In addition, the renderer may produce recommended-viewport video on the basis of the recommended viewport information in the scene description if included.

4 The rendered video is transferred to the player by indicating the rendering position and direction in the three-dimensional space used when the video was produced. A low-latency transport is to be used for transferring the video and the information on rendering position and direction.

5 The player presents the whole or part of the transferred video in accordance with the user's viewing position and direction.

Attachment
to the Annex
(informative)

Implementation example of high-level system architecture

1 Overview

This Attachment gives an example system in which the high-level system architecture for immersive video for presentation on various types of display devices defined in the Annex is implemented.

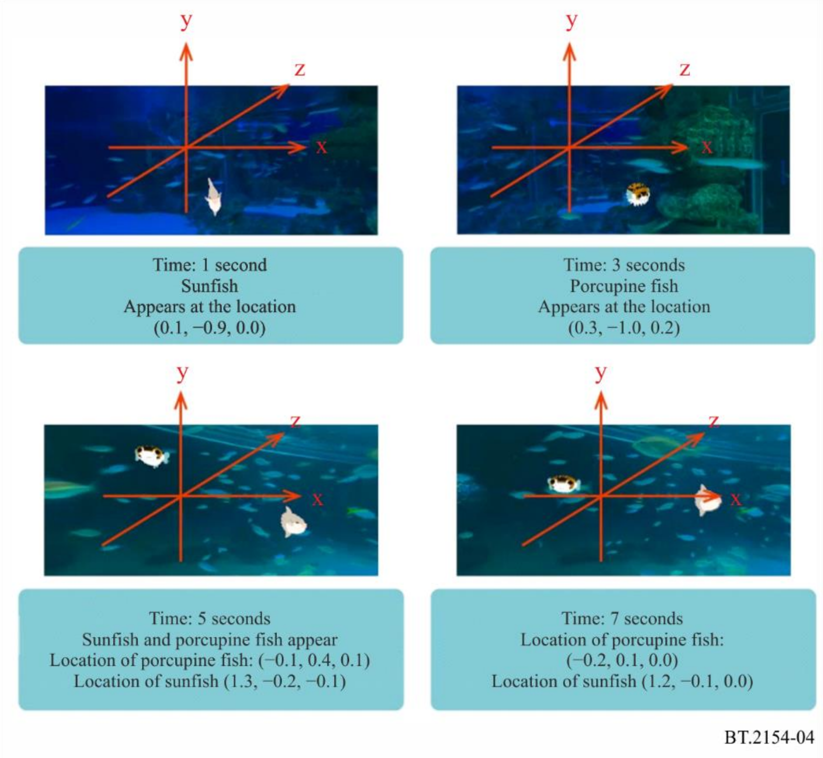
2 Immersive video

2.1 Scene description

The concept of scene description is shown in Fig. 4. As shown in the Figure, when the time instant is 1 second, a sunfish object appears at the location (0.1, −0.9, 0.0) in three-dimensional space. After 2 seconds, when the time instant is 3 seconds, a porcupine fish appears at the location (0.3, −1.0, 0.2). At this time instant, the sunfish object has disappeared. In this way, the scene description specifies the position, orientation and size of the objects in the three-dimensional space at every time instant.

FIGURE 4

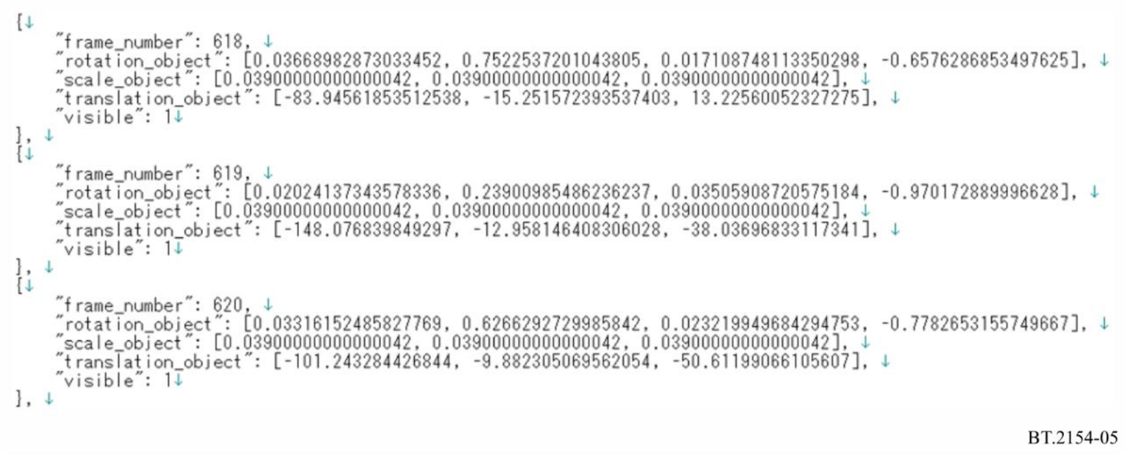
Arranging time-series objects in three-dimensional space by utilizing scene description



In this example, the extended format of GL Transmission Format (glTF2), which is specified at https://github.com/KhronosGroup/glTF/tree/master/specification/2.0, is used for the scene description. Figure 5 shows an example of the scene description.

FIGURE 5

Example scene description



```
[↓
    "frame_number": 618, ↓
    "rotation_object": [0.03668982873033452, 0.7522537201043805, 0.017108748113350298, -0.6576286853497625], ↓
    "scale_object": [0.0390000000000042, 0.0390000000000042, 0.0390000000000042], ↓
    "translation_object": [-83.94561853512538, -15.251572393537403, 13.22560052327275], ↓
    "visible": 1↓
], ↓
[↓
    "frame_number": 619, ↓
    "rotation_object": [0.02024137343578336, 0.23900985486236237, 0.03505908720575184, -0.970172889996628], ↓
    "scale_object": [0.0390000000000042, 0.0390000000000042, 0.0390000000000042], ↓
    "translation_object": [-148.076839849297, -12.958146408306028, -38.03696833117341], ↓
    "visible": 1↓
], ↓
[↓
    "frame_number": 620, ↓
    "rotation_object": [0.03316152485827769, 0.6266292729985842, 0.023219949684294753, -0.7782653155749667], ↓
    "scale_object": [0.0390000000000042, 0.0390000000000042, 0.0390000000000042], ↓
    "translation_object": [-101.243284426844, -9.882305069562054, -50.61199066105607], ↓
    "visible": 1↓
], ↓
```

BT.2154-05

2.2 Video object

As the video objects for volumetric video, point cloud streams obtained by compressing point-cloud-format volumetric video with ISO/IEC 23090-5 “Visual Volumetric Video-based Coding and Video-based Point Cloud Compression” are used.

As the omnidirectional video, the video obtained by Equirectangular Projection (ERP)-converted 360-degree video stored in ISO/IEC 23090-2 “Omnidirectional Media Format (OMAF)” is used.

In addition, two-dimensional rectangular video is used for overlay presentation.

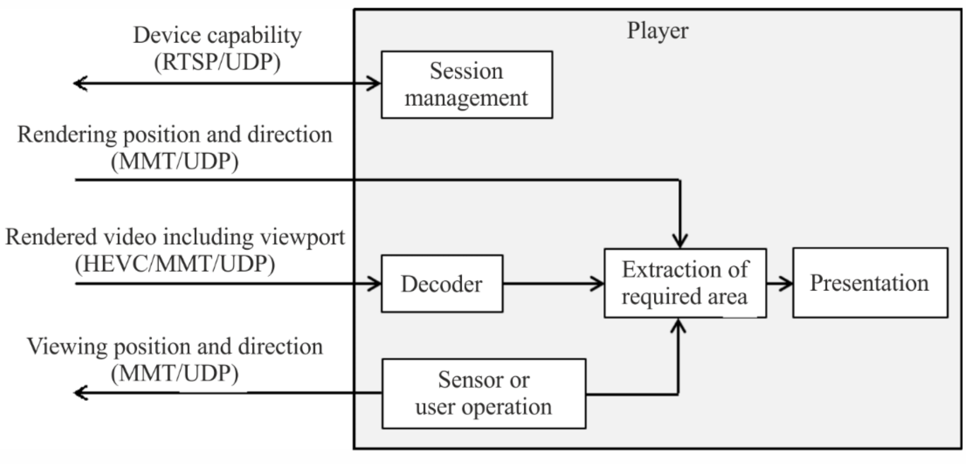
3 Implementation of renderer and player

3.1 Implementation of player

Players have been developed for a head-mounted display, a smartphone/tablet, and a conventional TV set, respectively. The player for conventional TV sets does not allow the user to change viewing position and direction. The functional blocks of these devices are depicted in Figs 6 and 7.

FIGURE 6

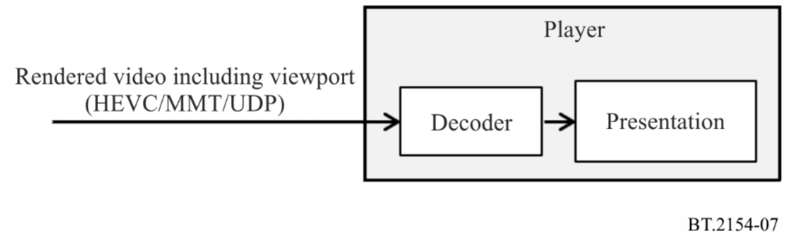
Functional blocks of the player for head-mounted display and smartphone/tablet



BT.2154-06

FIGURE 7

Functional blocks of the player for no user operation device assuming TV set



The player uses a Real-Time Streaming Protocol (RTSP, IETF RFC 7826) SETUP method to establish a session with the server and to inform the server of its capabilities including its display resolution, frame rate, field of view, and the available coding method used for compressing the video including the viewport.

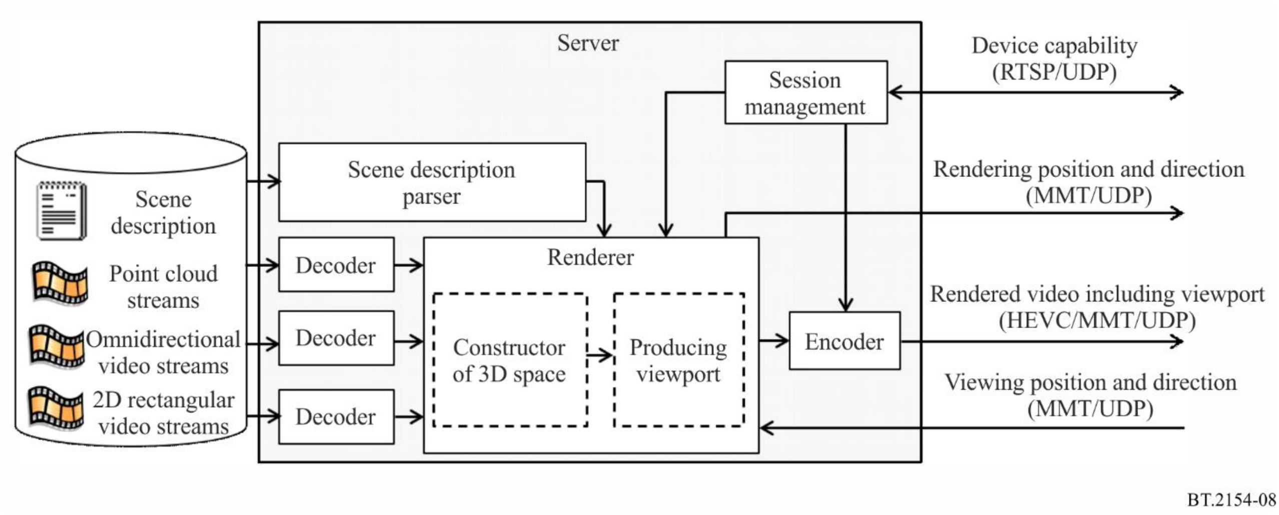
The user's viewing position and direction are notified to the server in the format of an MPEG Media Transport (MMT, ISO/IEC 23008-1) message. In the case of a head-mounted display, the viewing position and direction are determined in accordance with the user's own movement, and in the case of a smartphone/tablet, they are determined in accordance with the user's screen operation.

3.2 Implementation of renderer

A server having a renderer function is developed separately from the player. Different types of players are required in accordance with the device type, but the server is common regardless of the types of players. Figure 8 depicts the functional blocks of the server having a renderer function.

FIGURE 8

Functional blocks of the server having renderer function



The server parses the scene descriptions, decodes the required video objects in real time, and arranges them in the three-dimensional space in accordance with the scene descriptions. Then, from the three-dimensional space, the renderer produces video as a viewport that has a display resolution in accordance with the viewing position and direction notified by the player. Another viewport is produced on the basis of the recommended viewport information included in the scene descriptions for the device where the device capability notification is not performed and viewing position/direction are not changed.

The video including viewport produced by the renderer is compressed with High Efficiency Video Coding (HEVC, ISO/IEC 23008-2 | Rec. ITU-T H.265) as two-dimensional video and transported to the player in an MMT format. At the same time, the rendering parameters used to produce the viewport are transferred to the player in an MMT message format.

4 Presentation on three different types of display devices

4.1 Head-mounted display

As shown in Fig. 9, using a head-mounted display enables users to enjoy watching video in the desired direction from the desired location while freely moving around with a high feeling of immersion. This enables the user to view the objects from not only the front but also the back and side.

FIGURE 9

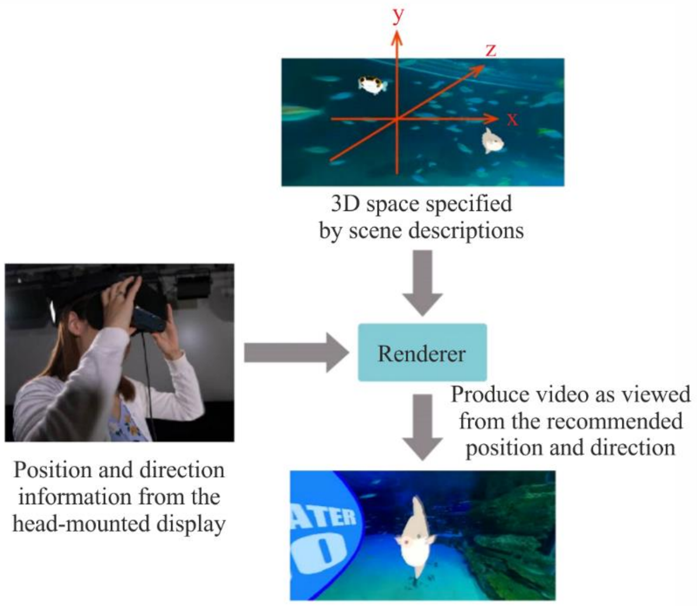
Viewing with a head-mounted display

BT.2154-09

In the system, the renderer produces the video in accordance with the viewing position and direction of the user as detected by sensors of the head-mounted display, and the player presents the produced video on the head-mounted display. The mechanism for presentation on a head-mounted display is shown in Fig. 10.

FIGURE 10

Mechanism for presentation on head-mounted display


y
z
x
3D space specified by scene descriptions
Renderer
Produce video as viewed from the recommended position and direction
Position and direction information from the head-mounted display

BT.2154-10

4.2 Smartphone

The viewing position and direction can be changed by operations on the smartphone screen, enabling users to view video in their desired direction from their desired location (see Fig. 11).

FIGURE 11

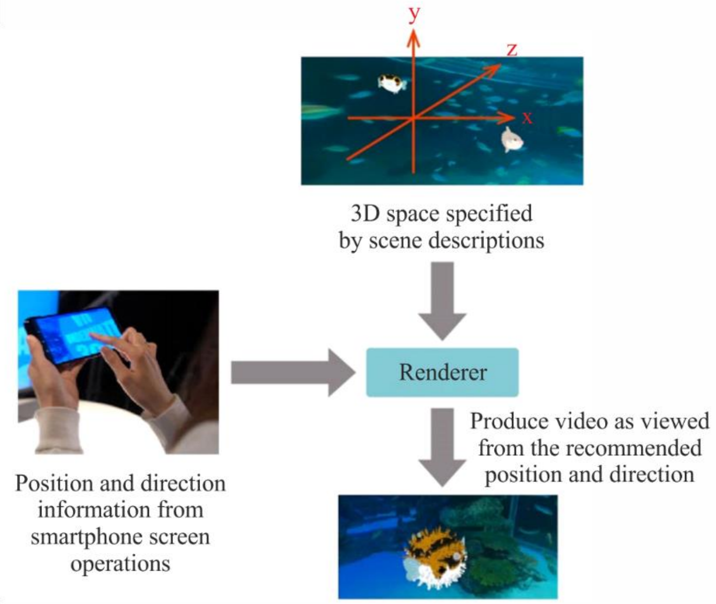
Viewing on smartphone

BT.2154-11

Like the case of presenting on head-mounted displays, the renderer produces video to be presented on smartphones on the basis of the scene descriptions. On smartphones, the player presents the video as viewed from the position and direction specified by the screen operations of the user (see Fig. 12).

FIGURE 12

Mechanism for presentation on smartphone


y
z
x
3D space specified by scene descriptions
Renderer
Produce video as viewed from the recommended position and direction
Position and direction information from smartphone screen operations

BT.2154-12

4.3 Television set

Although users cannot change the location and direction of a television set like on head-mounted displays and smartphones, they can still easily enjoy video from the viewing location and direction recommended by the content creator (see Fig. 13).

FIGURE 13

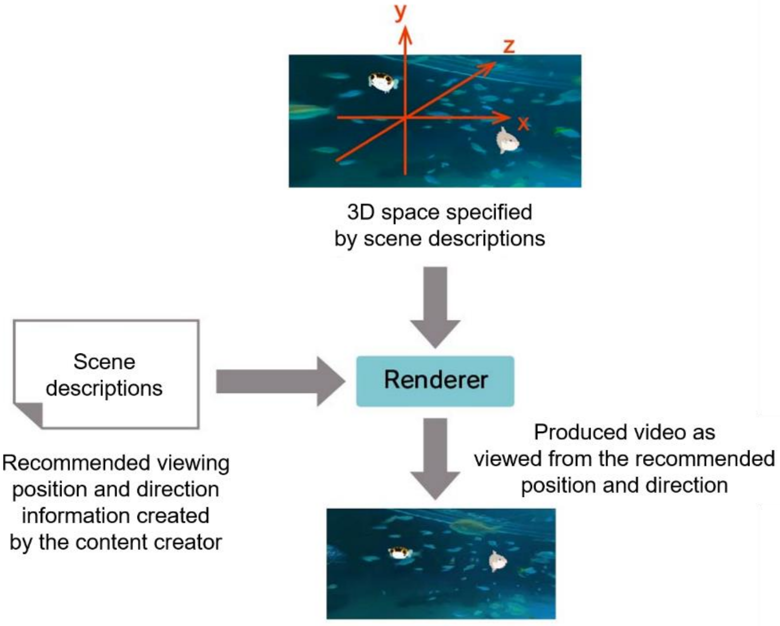
Viewing on television set

BT.2154-13

Even in this case, the renderer produces the video on the basis of the three-dimensional space information specified by the scene descriptions. However, since there are no user operations, information on the viewing position and direction is given by the scene descriptions as recommended viewport information. In accordance with this information, the renderer produces the video to be presented as illustrated in Fig. 14.

FIGURE 14

Mechanism for presentation on no user operation display (TV set)



5 References

Presentation on three-types of devices is available at the following URL:
https://www.nhk.or.jp/strl/english/open2021/tenji/3/index.html

Specifications used in the implementation are as follows:

Recommendation ITU-T H.265 | ISO/IEC 23008-2 (2020): Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High efficiency video coding

ISO/IEC 23008-1:2017: Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 1: MPEG media transport

ISO/IEC 23090-2:2021: Information technology – Coded representation of immersive media – Part 2: Omnidirectional media format

ISO/IEC 23090-5:2021: Information technology – Coded representation of immersive media – Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC)

IETF RFC 7826 (2016): Real-Time Streaming Protocol Version 2.0

glTF 2.0 Khronos Group, The GL Transmission Format (glTF) 2.0 Specification, available at https://github.com/KhronosGroup/glTF/tree/master/specification/2.0/

____