Unión Internacional de Telecomunicaciones

UIT-R

Sector de Radiocomunicaciones de la UIT

Recomendación UIT-R BT.2154-0
(12/2022)

Arquitectura de sistema de alto nivel para la presentación de vídeo inmersivo en diversos tipos de dispositivos de visualización

Serie BT

Servicio de radiodifusión (televisión)

Unión Internacional de Telecomunicaciones

Prólogo

El Sector de Radiocomunicaciones tiene como cometido garantizar la utilización racional, equitativa, eficaz y económica del espectro de frecuencias radioeléctricas por todos los servicios de radiocomunicaciones, incluidos los servicios por satélite, y realizar, sin limitación de gamas de frecuencias, estudios que sirvan de base para la adopción de las Recomendaciones UIT-R.

Las Conferencias Mundiales y Regionales de Radiocomunicaciones y las Asambleas de Radiocomunicaciones, con la colaboración de las Comisiones de Estudio, cumplen las funciones reglamentarias y políticas del Sector de Radiocomunicaciones.

Política sobre Derechos de Propiedad Intelectual (IPR)

La política del UIT-R sobre Derechos de Propiedad Intelectual se describe en la Política Común de Patentes UIT-T/UIT-R/ISO/CEI a la que se hace referencia en la Resolución UIT-R 1. Los formularios que deben utilizarse en la declaración sobre patentes y utilización de patentes por los titulares de las mismas figuran en la dirección web http://www.itu.int/ITU-R/go/patents/es, donde también aparecen las Directrices para la implementación de la Política Común de Patentes UIT-T/UIT-R/ISO/CEI y la base de datos sobre información de patentes del UIT-R sobre este asunto.

Series de las Recomendaciones UIT-R

(También disponible en línea en http://www.itu.int/publ/R-REC/es)

Series	Título
BO	Distribución por satélite
BR	Registro para producción, archivo y reproducción; películas en televisión
BS	Servicio de radiodifusión (sonora)
BT	**Servicio de radiodifusión (televisión)**
F	Servicio fijo
M	Servicios móviles, de radiodeterminación, de aficionados y otros servicios por satélite conexos
P	Propagación de las ondas radioeléctricas
RA	Radio astronomía
RS	Sistemas de detección a distancia
S	Servicio fijo por satélite
SA	Aplicaciones espaciales y meteorología
SF	Compartición de frecuencias y coordinación entre los sistemas del servicio fijo por satélite y del servicio fijo
SM	Gestión del espectro
SNG	Periodismo electrónico por satélite
TF	Emisiones de frecuencias patrón y señales horarias
V	Vocabulario y cuestiones afines

***Nota**: Esta Recomendación UIT-R fue aprobada en inglés conforme al procedimiento detallado en la Resolución UIT-R 1.*

Publicación electrónica
Ginebra, 2023

RECOMENDACIÓN UIT-R BT.2154-0

Arquitectura de sistema de alto nivel para la presentación de vídeo inmersivo en diversos tipos de dispositivos de visualización

(Cuestiones UIT-R 140-1/6 y UIT-R 143-2/6)

(2022)

Cometido

En esta Recomendación se define una arquitectura de sistema de alto nivel para la presentación de vídeo inmersivo en diversos tipos de dispositivos de visualización. La arquitectura está compuesta, como mínimo, de objetos de vídeo, descripciones de escenas, un procesador y un reproductor. En esta Recomendación se indica asimismo la información que se ha de transferir del procesador al reproductor.

Palabras clave

Vídeo inmersivo, 6DoF, descripción de escena, vídeo volumétrico, adaptación de dispositivo

La Asamblea de Radiocomunicaciones de la UIT,

considerando

a) que el vídeo inmersivo, que permite a los usuarios finales moverse en el espacio de vídeo y ver el vídeo omnidireccionalmente desde cualquier punto de vista, ofrece una nueva experiencia visual mejorada;

b) que el vídeo inmersivo se representa mediante la organización de objetos de vídeo, como el vídeo volumétrico, el vídeo omnidireccional y el vídeo bidimensional en un espacio tridimensional;

c) que hay diversos tipos de dispositivos de visualización a disposición de los usuarios finales, como los cascos, los teléfonos inteligentes y las tabletas, que han de considerarse para la presentación de vídeo inmersivo;

d) que hay cada vez más plataformas de entrega interactivas disponibles para la distribución de contenido al público;

e) que los servidores de red, incluidos en la nube y periféricos, con mayor potencia de cálculo se utilizarán efectivamente para el procesamiento de vídeo inmersivo en función de los distintos tipos de dispositivos de visualización con sus correspondientes capacidades de cálculo y reproducción;

f) que la existencia de una arquitectura común para la presentación de vídeo inmersivo en distintos tipos de dispositivos de visualización facilitará el desarrollo de sistemas y aplicaciones de vídeo inmersivo,

reconociendo

a) la serie de normas ISO/CEI 23090 – Information technology – Coded representation of immersive media;

b) la Recomendación UIT-R BT.2123 – Valores de parámetros de vídeo de los sistemas audiovisuales inmersivos avanzados para la producción y el intercambio internacional de programas en el ámbito de la radiodifusión;

c) el Informe UIT-R BT.2420 – Collection of usage scenarios of advanced immersive sensory media systems,

recomienda

que los sistemas de vídeo inmersivo destinados a diversos tipos de dispositivos de visualización se diseñen conforme a la arquitectura de sistema de alto nivel descrita en el Anexo.

Anexo

Arquitectura de sistema de alto nivel para la presentación de vídeo inmersivo en distintos tipos de dispositivos de visualización

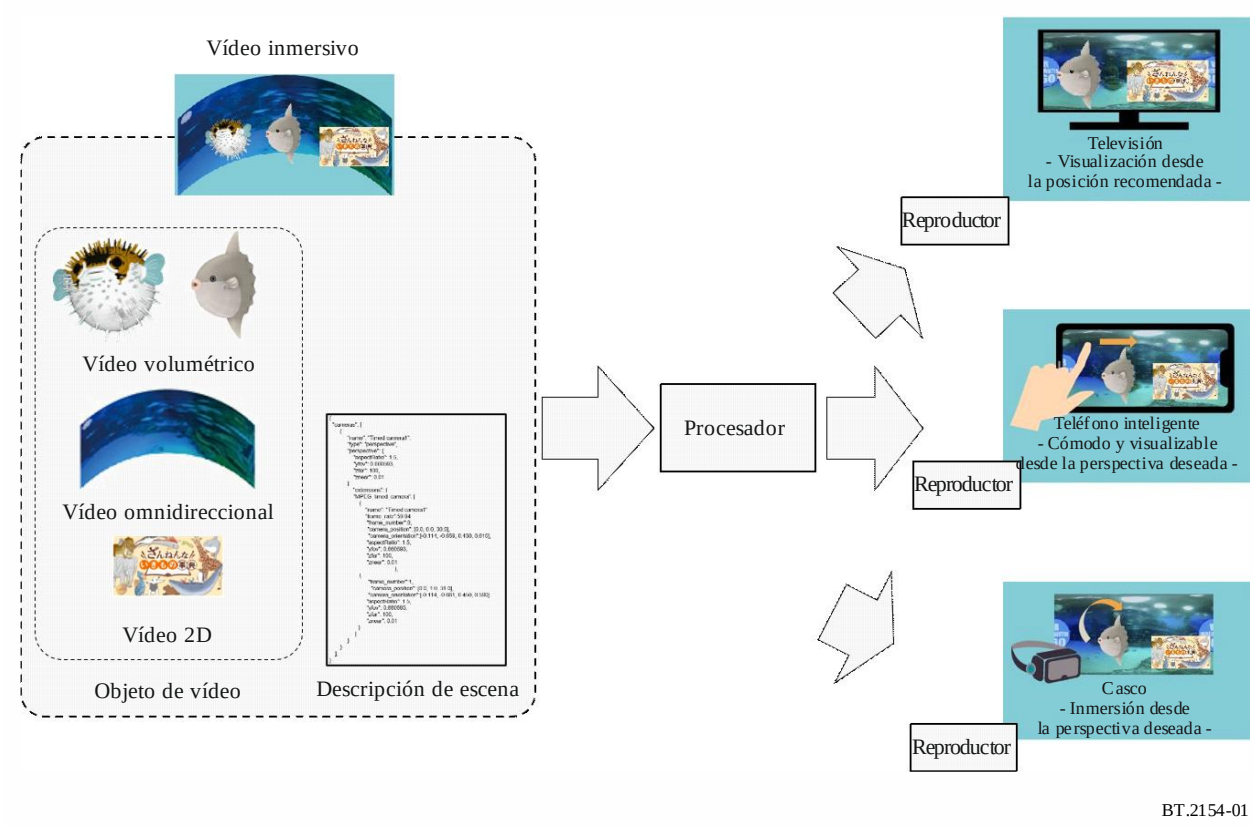
1 Introducción

En la Fig. 1 se muestra de manera resumida la arquitectura de sistema para el vídeo inmersivo desde la composición a la presentación en distintos tipos de dispositivos de visualización.

El vídeo inmersivo se compone de una descripción de escena y múltiples objetos de vídeo, indicados en la descripción de escena, incluido el vídeo volumétrico, que puede representar la forma tridimensional y la textura de los objetos, el vídeo omnidireccional que rodea a los objetos y el vídeo bidimensional para que los usuarios puedan ver el vídeo en cualquier posición y desde cualquier dirección. El vídeo omnidireccional y el vídeo 2D pueden poseer su propia información de profundidad. En las descripciones de escena se ofrece una representación tridimensional temporal del vídeo inmersivo, incluidas, entre otras cosas, la posición, la orientación y el tamaño de cada objeto, así como su disposición temporal y espacial en el espacio tridimensional.

FIGURA 1

Resumen del vídeo inmersivo desde la composición a la presentación



Un procesador construye un espacio tridimensional a partir de los distintos objetos de vídeo indicados en la descripción de escena y produce el vídeo que se verá desde la posición y perspectiva del usuario. El reproductor de cada dispositivo de visualización presenta el vídeo procesado de la manera que mejor se adapte al dispositivo, en función de la posición y perspectiva del usuario.

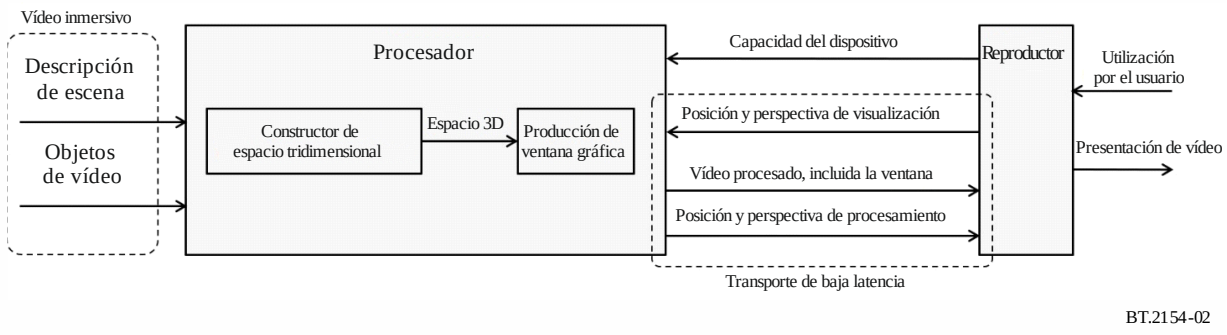
2 Arquitectura de sistema de alto nivel

2.1 Definición

La arquitectura de sistema de alto nivel ilustrada en la Fig. 2 está definida para presentar vídeo inmersivo en distintos tipos de dispositivos de visualización.

FIGURA 2

Arquitectura de sistema de alto nivel para vídeo inmersivo



2.2 Vídeo inmersivo

El vídeo inmersivo representa un espacio tridimensional temporal y se compone de objetos de vídeo y descripciones de escenas.

Los objetos de vídeo incluyen el vídeo volumétrico, que representa la forma tridimensional y la textura de los objetos, el vídeo omnidireccional que rodea a los objetos y el vídeo rectangular bidimensional (2D). El vídeo omnidireccional y el vídeo 2D pueden asociarse a información de profundidad.

En la descripción de escena se define un espacio tridimensional temporal haciendo referencia a múltiples objetos de vídeo e identificando la posición, la orientación y el tamaño de cada objeto, así como su disposición espacial y temporal en el espacio tridimensional.

La descripción de escena también puede incluir información sobre la posición y perspectiva de visionado del usuario recomendadas por el creador del contenido, es decir, la ventana gráfica recomendada para la presentación en el dispositivo de visualización.

2.3 Procesador y reproductor

Un procesador construye un espacio tridimensional a partir de diversos objetos de vídeo y de la descripción de la escena. También produce el vídeo que verá el usuario desde la posición y perspectiva que desee.

El reproductor presenta el vídeo procesado de la manera que mejor se adapte al dispositivo en función de la posición y perspectiva del usuario.

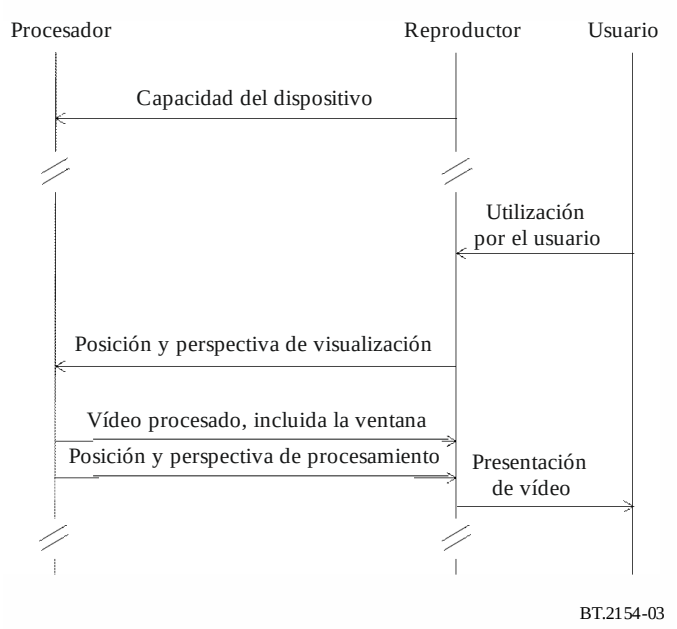
Las funciones del procesador y del reproductor deben diferenciarse a fin de que el reproductor no tenga que procesar toda la información en el espacio tridimensional, sino sólo la porción que se vaya a presentar en el dispositivo. Esta diferenciación reduce el tamaño de datos de los objetos de vídeo que debe procesar el reproductor y la carga de procesamiento del reproductor, gracias a lo cual pueden implementarse reproductores de procesamiento más ligeros. Así, cuando en el futuro se introduzcan tipos de objetos de vídeo adicionales, para soportarlos sólo será necesario actualizar el procesador y no el reproductor.

2.4 Información que se ha de transferir del procesador al reproductor

En la Fig. 3 se muestra el flujo de información que se ha de transferir del procesador al reproductor.

FIGURA 3

Información que se ha de transferir del procesador al reproductor



1) Antes de iniciar el procesamiento, el procesador construye un espacio tridimensional a partir de la descripción de escena y de los objetos de vídeo y el reproductor notifica al procesador la capacidad del dispositivo, incluida la resolución de imagen, el campo de visión y la velocidad de trama.

2) Cuando el usuario empieza a ver el contenido, el reproductor notifica al procesador la posición y perspectiva de visualización del usuario, que puede cambiar durante el visionado, en función de la utilización del usuario.

3) A partir del espacio tridimensional, el procesador produce el vídeo, incluida la ventana gráfica, que se presentará de acuerdo con la posición y perspectiva de visualización del usuario notificadas. El procesador puede producir vídeo en áreas más amplias que la ventana gráfica para soportar los cambios rápidos de posición y perspectiva de visualización. Además, el procesador puede producir el vídeo con la ventana gráfica recomendada, basándose en la información de ventana gráfica recomendada de la descripción de escena, si tal información está incluida.

4) El vídeo procesado se transfiere al reproductor indicando la posición y perspectiva de procesamiento en el espacio tridimensional utilizadas al producir el vídeo. Se ha de utilizar un transporte de baja latencia para transferir el vídeo y la información sobre la posición y perspectiva de procesamiento.

5) El reproductor presenta total o parcialmente el vídeo transferido en función de la posición y perspectiva de visualización del usuario.

Adjunto
al Anexo
(informativo)

Ejemplo de implementación de la arquitectura de sistema de alto nivel

1 Introducción

En este Adjunto se da un ejemplo de sistema en que se implementa la arquitectura de sistema de alto nivel para la presentación de vídeo inmersivo en distintos tipos de dispositivos de visualización definida en este Anexo.

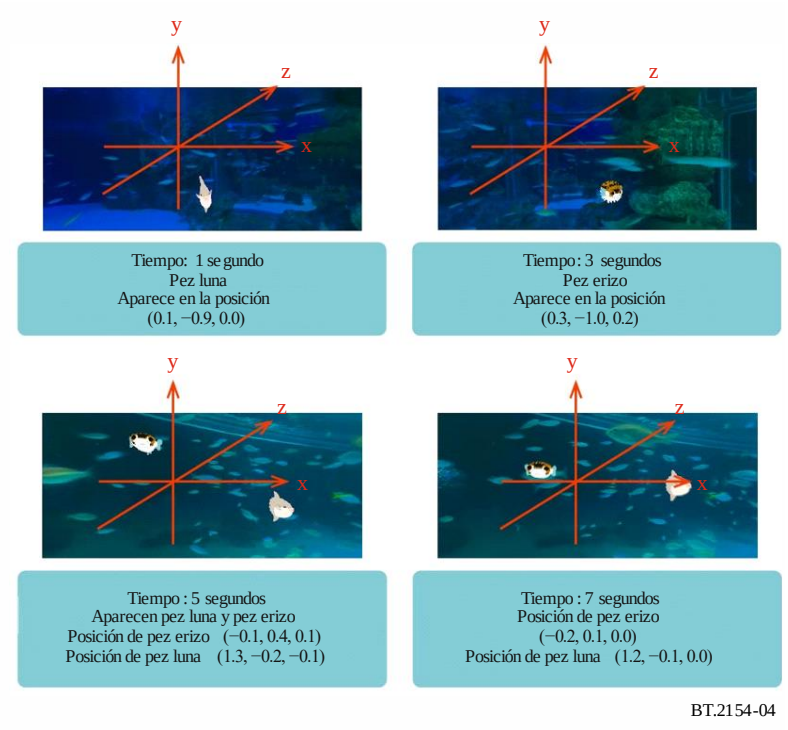
2 Vídeo inmersivo

2.1 Descripción de escena

El concepto de descripción de escena se ilustra en la Fig. 4. Como puede verse, cuando el instante es 1 segundo aparece el objeto pez luna en la posición (0.1, −0.9, 0.0) en el espacio tridimensional. Tras 2 segundos, cuando el instante es 3 segundos, aparece un pez erizo en la posición (0.3, −1.0, 0.2). En ese instante ha desaparecido el objeto pez luna. De esta manera, la descripción de escena especifica la posición, orientación y tamaño de los objetos en el espacio tridimensional en cada instante.

FIGURA 4

Disposición de objetos temporales en el espacio tridimensional utilizando la descripción de escena



En este ejemplo, se utiliza para la descripción de escena el formato ampliado del formato de transmisión GL (glTF2), especificado en https://github.com/KhronosGroup/glTF/tree/master/specification/2.0. En la Fig. 5 se muestra un ejemplo de descripción de escena.

FIGURA 5

Ejemplo de descripción de escena


```
[
    "frame_number": 618,
    "rotation_object": [0.03668982873033452, 0.7522537201043805, 0.017108748113350298, -0.6576286853497625],
    "scale_object": [0.03900000000000042, 0.03900000000000042, 0.03900000000000042],
    "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": [0.03900000000000042, 0.03900000000000042, 0.03900000000000042],
    "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": [0.03900000000000042, 0.03900000000000042, 0.03900000000000042],
    "translation_object": [-101.243284426844, -9.882305069562054, -50.61199066105607],
    "visible": 1
],
```


BT.2154-05

2.2 Objeto de vídeo

Como objetos de vídeo para el vídeo volumétrico se utilizan trenes de nubes de puntos obtenidas mediante la compresión de vídeo volumétrico en formato nube de puntos con ISO/CEI 23090-5 «Visual Volumetric Video-based Coding and Video-based Point Cloud Compression».

Para el vídeo omnidireccional se utiliza el vídeo obtenido mediante el vídeo a 360 grados con conversión de proyección equirrectangular (ERP) almacenado en formato ISO/CEI 23090-2 «Omnidirectional Media Format (OMAF)».

Además se utiliza el vídeo rectangular bidimensional para la presentación superpuesta.

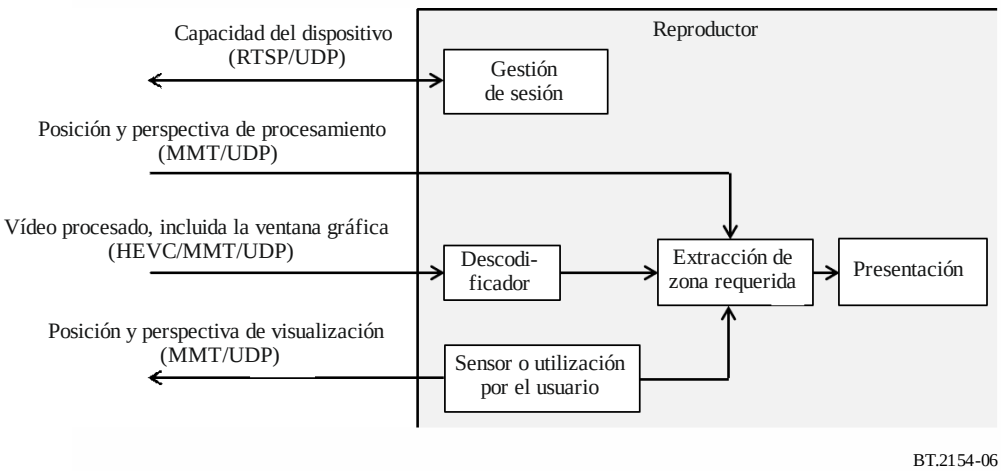
3 Implementación del procesador y el reproductor

3.1 Implementación del reproductor

Se han creado reproductores para cascos, teléfonos inteligentes/tabletas y televisores convencionales. El reproductor para televisor convencional no permite al usuario cambiar de postura o perspectiva de visualización. Los bloques funcionales de estos dispositivos de muestran en las Figs. 6 y 7.

FIGURA 6

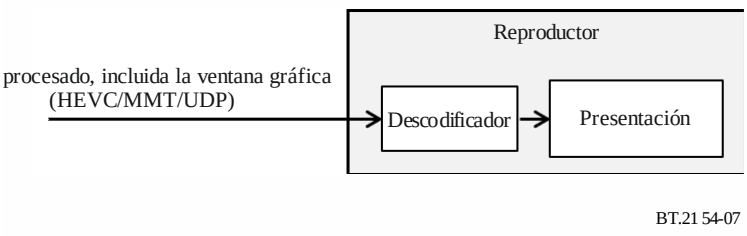
Bloques funcionales del reproductor para cascos y teléfonos móviles/tabletas



BT.2154-06

FIGURA 7

Bloques funcionales del reproductor para utilización independiente del usuario, por ejemplo, televisor



El reproductor utiliza el método SETUP del protocolo de reproducción directa en tiempo real (RTSP, IETF RFC 7826) para establecer una sesión con el servidor e informarle de sus capacidades, incluidas la resolución de imagen, la velocidad de tramas, el campo de visión y el método de codificación disponible utilizado para comprimir el vídeo, incluida la ventana gráfica.

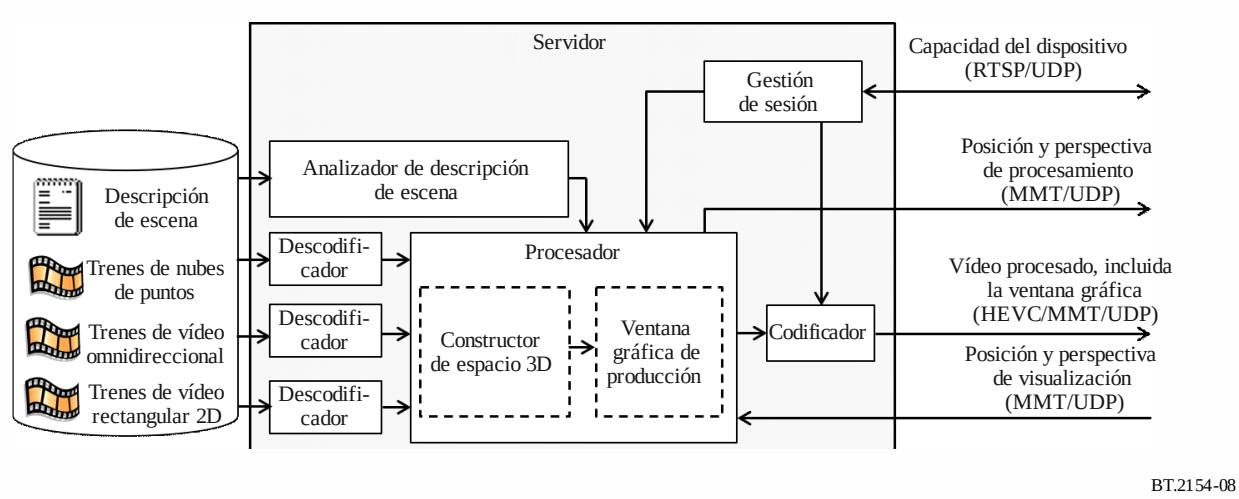
La posición y perspectiva de visualización del usuario se notifican al servidor en un mensaje de transporte de medios MPEG (MMT, ISO/CEI 23008-1). Si el dispositivo de visualización es un casco, la posición y la perspectiva de visualización se determinan en función del movimiento del propio usuario y, si se trata de un teléfono inteligente/tableta, se determinan según el funcionamiento de la pantalla del usuario.

3.2 Implementación del procesador

Un servidor con función de procesamiento es independiente del reproductor. Para cada tipo de dispositivo se necesitará un tipo distinto de reproductor, pero el servidor es común a todos los tipos de reproductor. En la Fig. 8 se muestran los bloques funcionales del servidor con función de procesamiento.

FIGURA 8

Bloques funcionales del servidor con función de procesamiento



El servidor analiza las descripciones de escenas, descodifica los objetos de vídeo necesarios en tiempo real y los dispone en el espacio tridimensional de acuerdo con las descripciones de escenas. A continuación, a partir del espacio tridimensional el procesador produce vídeo como una ventana gráfica con una resolución de imagen conforme con la posición y la perspectiva de visualización notificadas por el reproductor. Se produce otra ventana gráfica de acuerdo con la información de

ventana gráfica recomendada incluida en las descripciones de escenas para el dispositivo, cuando no se notifica la capacidad del dispositivo y no se modifican la posición/perspectiva de visualización.

El vídeo, incluida la ventana gráfica, producido por el procesador se comprime con codificación de vídeo de gran eficacia (HEVC, ISO/CEI 23008-2 | Rec. UIT-T H.265) como vídeo bidimensional y se transporta al reproductor en formato MMT. Al mismo tiempo, se transfieren al reproductor en formato de mensaje MMT los parámetros de procesamiento utilizados para producir la ventana gráfica.

4 Presentación en tres tipos de dispositivos de visualización distintos

4.1 Casco

Como se ve en la Fig. 9, la utilización de un casco permite al usuario ver el vídeo desde la perspectiva y posición deseadas, moviéndose libremente y disfrutando de una gran sensación de inmersión. El usuario puede ver los objetos no sólo de frente, sino también por detrás y por los lados.

FIGURA 9

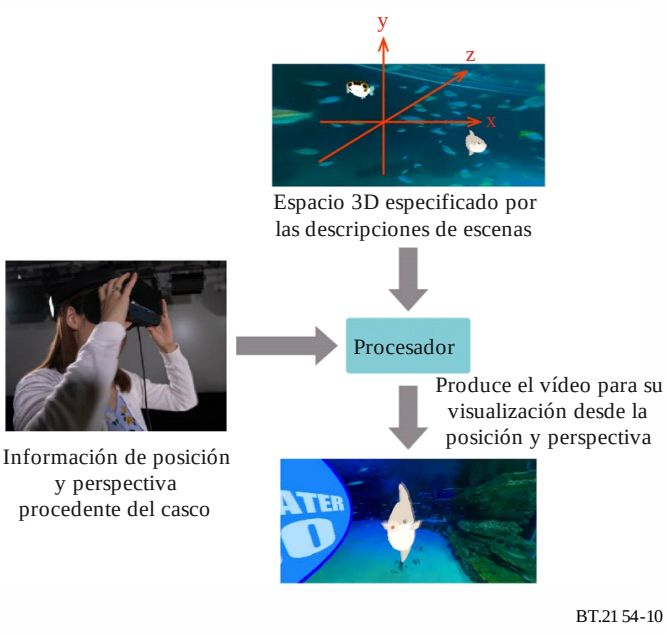
Visualización con un casco

BT.2154-09

En el sistema, el procesador produce el vídeo de acuerdo con la posición y perspectiva del usuario, detectadas mediante sensores en el casco, y el reproductor presenta el vídeo producido en el casco. El mecanismo de presentación en un casco se ilustra en la Fig. 10.

FIGURA 10

Mecanismo para la presentación en un casco

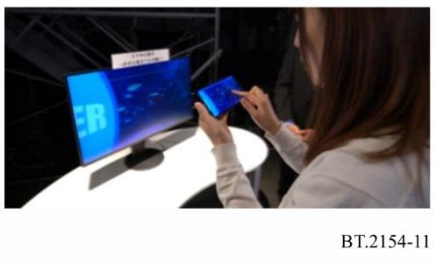


4.2 Teléfono inteligente

La posición y perspectiva de visualización pueden modificarse manipulando la pantalla del teléfono, lo que permite a los usuarios ver el vídeo desde la postura y perspectiva deseadas (véase la Fig. 11).

FIGURA 11

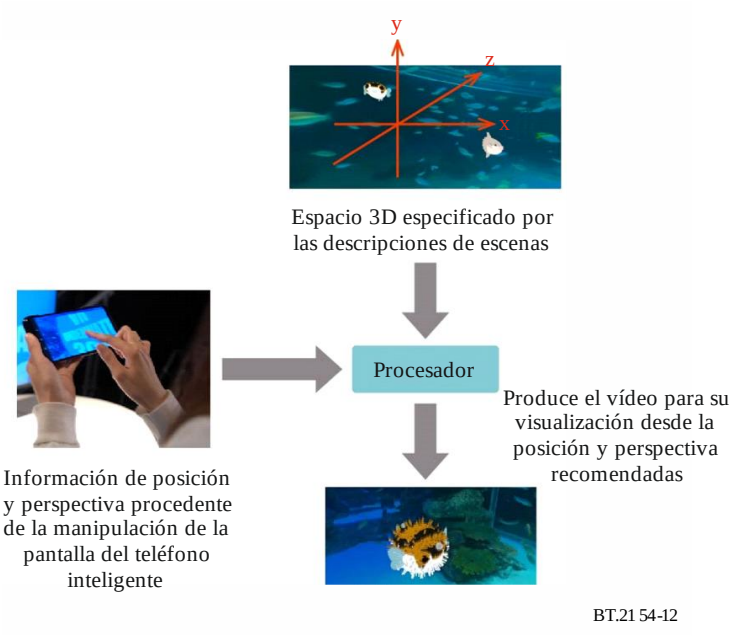
Visualización en teléfono inteligente



Como ocurre con la presentación en cascos, el procesador produce el vídeo que se presentará en un teléfono inteligente sobre la base de las descripciones de escenas. En los teléfonos inteligentes el reproductor presenta el vídeo en función de la posición y perspectiva especificadas por la manipulación de la pantalla por el usuario (véase la Fig. 12).

FIGURA 12

Mecanismo para la presentación en teléfono inteligente



4.3 Televisor

Aunque los usuarios no pueden modificar la posición y perspectiva de un televisor como con los casos o teléfonos inteligentes, sí pueden disfrutar del vídeo desde la posición y perspectiva de visualización recomendadas por el creador de contenido (véase la Fig. 13).

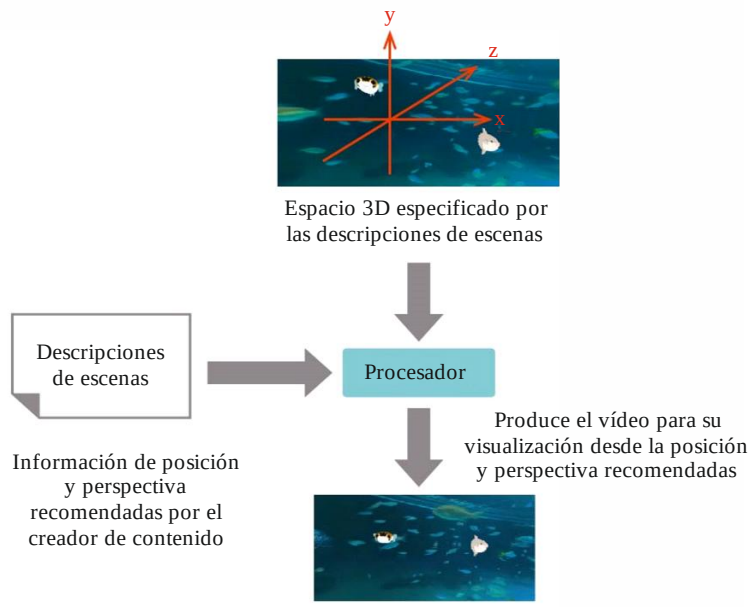
FIGURA 13

Visualización en televisor

También en este caso el procesador produce el vídeo basándose en la información tridimensional especificada en las descripciones de escenas. Sin embargo, al no haber intervención del usuario, la información de posición y perspectiva de visualización se obtiene de lo que en las descripciones de escenas se recomienda como ventana gráfica. De acuerdo con esa información, el procesador produce el vídeo que se ha de presentar, como se muestra en la Fig. 14.

FIGURA 14

Mecanismo para la presentación en dispositivos sin intervención del usuario (televisor)



BT.2154-14

5 Referencias

La presentación en tres tipos de dispositivos está disponible en la siguiente dirección: https://www.nhk.or.jp/strl/english/open2021/tenji/3/index.html

Para la implementación se utilizan las siguientes especificaciones:

Recomendación UIT-T H.265 | ISO/CEI 23008-2 (2020): Tecnología de la información – codificación de gran eficacia y entrega de medios en entornos heterogéneos – Parte 2: codificación de vídeo de gran eficacia

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

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

ISO/CEI 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, disponible en https://github.com/KhronosGroup/glTF/tree/master/specification/2.0/