UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE COMUNICACIONES

Evaluación de la QoE en un sistema de streaming adaptativo de vídeo 3D basado en DASH

TESIS DOCTORAL

Paola Fernanda Guzmán Castillo

Supervisores:

Dr. Juan Carlos Guerri Cebollada

Dr. Pau Arce Vila

Valencia, España

Junio - 2022

Abstract

The distribution of multimedia content, and in particular video streaming, currently dominates global Internet traffic and will become even more important in the future. Thousands of titles are added monthly to major service providers such as Netflix, YouTube and Amazon. In addition to the consumption of high-definition content becoming the main trend, an increase in the consumption of 3D content can be observed again. This fact has caused that issues related to content production, encoding, transmission, Quality of Service (QoS) and Quality of Experience (QoE) perceived by users of 3D video distribution systems became a research topic with numerous contributions in recent years.

This thesis addresses the problem of providing 3D video streaming services under variable bandwidth network conditions. In this sense, it presents the results of the evaluation of the QoE perceived by the users of 3D video systems, analyzing mainly the impact of the effects introduced in two of the elements of the 3D video processing chain: the encoding stage and the transmission process.

To analyze the effects of the encoding process on the quality of 3D video, the first stage deals with the objective and subjective evaluation of video quality, comparing the performance of different encoding standards and methods, in order to identify those that achieve the best ratio between quality, bit rate and encoding time. Also, in the context of transmission in a simulcast environment, the advantages of using asymmetric coding for 3D video transmission is evaluated as an alternative for bandwidth reduction while maintaining overall quality.

Secondly, for the study of the impact and performance of the transmission process, the work has been carried out on the basis of an adaptive dynamic over HTTP (DASH) transmission system in the context of both 2D and 3D video transmission, using different bandwidth variation scenarios. The aim has been to develop a framework for the evaluation of QoE in 3D adaptive video streaming scenarios, which allows analyzing the impact on the user's QoE against different bandwidth variation patterns, as well as the performance of the adaptation algorithm under these scenarios. The work focuses on identifying the impact on the user's Quality of Experience in aspects such as: frequency, type, range and temporal location of bandwidth variation events.

The proposed system allows to perform performance measurements in an automated and systematic way for the evaluation of DASH systems in the 2D and 3D video distribution service. The tool Puppeteer, the Node.js library developed by Google, has been used, which provides a high-level API to automate actions in the Chrome Devtools protocol, such as starting playback, causing bandwidth changes and saving the results of the quality change processes, timestamps, stops, etc. From this data, a further processing is performed that allows the reconstruction of the displayed video, as well as the extraction of quality metrics and the evaluation of the QoE of the users using the ITU-T P.1203 recommendation.

Resumen

La distribución de contenidos multimedia, y en particular el *streaming* de vídeo, domina actualmente el tráfico global de Internet y su importancia será incluso mayor en el futuro. Miles de títulos se agregan mensualmente a los principales proveedores de servicios, como Netflix, YouTube y Amazon. Y de la mano del consumo de contenidos de alta definición que se convierte en la principal tendencia, se puede observar nuevamente un incremento en el consumo de contenidos 3D. Esto ha hecho que las temáticas relacionadas con la producción de contenidos, codificación, transmisión, Calidad de Servicio (QoS) y Calidad de Experiencia (QoE) percibidas por los usuarios de los sistemas de distribución de vídeo 3D sean un tema de investigación con numerosas contribuciones en los últimos años.

Esta tesis aborda el problema de la prestación de servicios de transmisión de vídeo 3D bajo condiciones de red de ancho de banda variable. En este sentido, presenta los resultados de la evaluación de la QoE percibida por los usuarios de los sistemas de vídeo 3D, analizando principalmente el impacto de los efectos introducidos en dos de los elementos de la cadena de procesamiento de vídeo 3D: la etapa de codificación y el proceso de transmisión.

Para analizar los efectos de la codificación en la calidad del vídeo 3D, en la primera etapa se aborda la evaluación objetiva y subjetiva de la calidad del vídeo, comparando el rendimiento de diferentes estándares y métodos de codificación, con el fin de identificar aquellos que logran la mejor relación entre calidad, tasa de bits y tiempo de codificación. Así mismo, en el contexto de la transmisión en un entorno *simulcast*, se evalúa la eficacia de la utilización de las codificaciones asimétricas para la transmisión de vídeo 3D, como una alternativa para la reducción del ancho de banda manteniendo la calidad global.

En segundo lugar, para el estudio del impacto y el rendimiento del proceso de transmisión, se ha trabajado sobre la base de un sistema de transmisión dinámica adaptativa sobre HTTP (DASH) en el contexto de la transmisión de vídeo tanto 2D como 3D, utilizando diferentes escenarios de variación de ancho de banda. El objetivo ha sido el desarrollo de un marco de referencia para la evaluación de la QoE en escenarios de transmisión adaptativa de vídeo 3D, que permite analizar el impacto en la QoE del usuario frente a diferentes patrones de variación del ancho de banda, así como el rendimiento del algoritmo de adaptación frente a estos escenarios. El trabajo se enfoca en identificar el impacto en la Calidad de Experiencia del usuario que tienen aspectos como: la frecuencia, el tipo, el alcance y la ubicación temporal de los eventos de variación del ancho de banda.

El sistema propuesto permite realizar mediciones de rendimiento de forma automatizada y sistemática para la evaluación de los sistemas DASH en el servicio de distribución de vídeo 2D y 3D. Se ha utilizado Puppeteer, la librería Node.js desarrollada por Google, que proporciona una

API de alto nivel, para automatizar acciones en el protocolo Chrome Devtools, como iniciar la reproducción, provocar cambios de ancho de banda y guardar los resultados de los procesos de cambio de calidad, marcas de tiempo, paradas, etc. A partir de estos datos, se realiza un procesamiento que permite la reconstrucción del vídeo visualizado, así como la extracción de métricas de calidad y la evaluación de la QoE de los usuarios utilizando la recomendación ITU-T P.1203.

Resum

La distribució de continguts multimèdia, i en particular el *streaming* de vídeo, domina actualment el trànsit global d'Internet i la seua importància serà fins i tot mes gran en el futur. Milers de títols s'afegeixen mensualment als principals proveïdors de serveis, com ara Netflix, YouTube i Amazon. I de la mà del consum de continguts d'alta definició que es converteix en la tendència principal, es pot observar novament un increment en el consum de continguts 3D. Això ha fet que les temàtiques relacionades amb la producció de continguts, codificació, transmissió, Qualitat de Servei (QoS) i Qualitat d'Experiència (QoE) percebudes pels usuaris dels sistemes de distribució de vídeo 3D siguen un tema de recerca amb nombroses contribucions en els últims anys.

Aquesta tesi aborda el problema de la prestació de serveis de transmissió de vídeo 3D sota condicions de xarxa d'ample de banda variable. En aquest sentit, presenta els resultats de l'avaluació de la QoE percebuda pels usuaris dels sistemes de vídeo 3D, analitzant principalment l'impacte dels efectes introduïts en dos dels elements de la cadena de processament de vídeo 3D: l'etapa de codificació i el procés de transmissió.

Per analitzar els efectes de la codificació en la qualitat del vídeo 3D, a la primera etapa s'aborda l'avaluació objectiva i subjectiva de la qualitat del vídeo, comparant el rendiment de diferents estàndards i mètodes de codificació, per tal d'identificar aquells que aconsegueixen la millor relació entre qualitat, taxa de bits i temps de codificació. Així mateix, en el context de la transmissió en un entorn simulcast, s'avalua l'eficàcia de la utilització de les codificacions asimètriques per la transmissió de vídeo 3D, com una alternativa per la reducció de l'ampleada de banda mantenint la qualitat global.

En segon lloc, per a l'estudi de l'impacte i el rendiment del procés de transmissió, s'ha treballat sobre la base d'un sistema de transmissió dinàmica adaptativa sobre HTTP (DASH) en el context de la transmissió de vídeo tant 2D com 3D, utilitzant diferents escenaris de variació d'ample de banda. L'objectiu ha estat el desenvolupament d'un marc de referència per a l'avaluació de la QoE en escenaris de transmissió adaptativa de vídeo 3D, que permet analitzar l'impacte en la QoE de l'usuari davant de diferents patrons de variació de l'ample de banda; així com el rendiment de l'algorisme d'adaptació davant d'aquests escenaris. El treball s'enfoca a identificar l'impacte a la Qualitat d'Experiència de l'usuari que tenen aspectes com ara: la freqüència, el tipus, l'abast i la ubicació temporal dels esdeveniments de variació de l'ample de banda.

El sistema proposat permet realitzar mesuraments de rendiment de manera automatitzada i sistemàtica per a l'avaluació dels sistemes DASH en el servei de distribució de vídeo 2D i 3D. S'ha

utilitzat Puppeteer, la llibreria Node.js desenvolupada per Google, que proporciona una API d'alt nivell, per automatitzar accions al protocol Chrome Devtools, com iniciar la reproducció, provocar canvis d'ample de banda i desar els resultats dels processos de canvi de qualitat, marques de temps, parades, etc. A partir d'aquestes dades, es fa un processament que permet la reconstrucció del vídeo visualitzat, així com l'extracció de mètriques de qualitat i l'avaluació de la QoE dels usuaris fent servir la recomanació ITU-T P.1203.

Agradecimientos

Quiero expresar mis más sinceros agradecimientos a mis tutores. Al Dr. Juan Carlos Guerri, por su excelente orientación y apoyo constante a lo largo de este proceso de realización de la tesis doctoral. La culminación de esta tesis no habría sido posible sin su ayuda. Al Dr. Pau Arce por su paciencia, buen rollo y por estar siempre dispuesto a compartir su conocimiento.

Agradezco también a todos mis compañeros del grupo de Comunicaciones Multimedia, a los que siguen y a los que se han ido, han sido muchos años de experiencias compartidas, esto no habría sido lo mismo sin nuestros almuerzos (partidas de pit).

Mi agradecimiento especial es para Alberto, sin su amor y apoyo incondicional, no habría sido posible sacar adelante este proyecto. Y para Sara, mi princesa que con sus razonamientos y preguntas me ayuda a recordar que es lo que realmente importa.

Finalmente, agradezco inmensamente a mi familia. Especialmente a mi madre, mi padre y mis hermanos, por todo lo que me han dado; por haberme apoyado siempre en cada proyecto emprendido. A mi familia valenciana, por acogerme y hacer que me sienta como en casa desde el primer día.

No hay que apagar la luz del otro para lograr que brille la nuestra.

Gandhi

A mi familia

Tabla de contenido

Lista de Tablas

Lista de Figuras

Acrónimos

2D	Dos dimensiones
3D	Tres Dimensiones
3DTV	Three Dimensional Television
AV1	AOMedia Video 1
AVC	Advanced Video Coding
ACR	Absolute Category Rating
BD	Blu-Ray Disc
CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CDP	Chrome Devtools Protocol
CRF	Constant Rate Factor
CTU	Coding Tree Unit
CU	Coding Units
DASH	Dynamic Adaptive Streaming over HTTP
DCT	Discret Cosine Transform
DTV	Digital Television
DVB	Digital Video Broadcasting
FPS	Frames Per Second
GoP	Group of Pictures
HAS	HTTP Adaptive Streaming
HEVC	High Efficiency Video Coding
HLS	HTTP Live Streaming
HM	HEVC test Model
HTTP	Hypertext Transfer Protocol
HVS	Human Visual System
IP	Internet Protocol
ITU	International Telecommunication Union
JMVC	Joint Multiview Video Coding
JSON	JavaScript Object Notation
JVT	Joint Video Team
MPD	Media Presentation Description
MPEG	Moving Picture Experts Group
MOS	Mean Opinion Score
MVC	Multiview Video Coding
MVD	Multiview Video plus Depth
MVV	Multiview Video

OTT	Over the Top
PSNR	Peak Signal to Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
QP	Quantization Parameter
RD	Rate Distortion
S3D	Stereoscopic 3D
SI	Spatial Information
SRC	Source Sequence
SSIM	Structural Similarity Index
TCP	Transmission Control Protocol
TI	Temporal Information
UDP	User Datagram Protocol
VCL	Video Coding Layer
VIF	Visual Information Fidelity
VMAF	Video Multimethod Assessment Fusion
VoD	Video on Demand
VVC	Versatile Video Coding

Capítulo 1
Introducción

Como resultado del auge de las redes sociales y las plataformas de distribución de contenidos como TikTok, Instagram, YouTube, Netflix entre muchas otras, los servicios relacionados con la transmisión de contenidos multimedia en particular el *streaming* de vídeo, crece y evoluciona a una velocidad increíble. Según las últimas estadísticas de consumo de vídeo en línea, el 91,4% de los usuarios de Internet de todo el mundo ven vídeos digitales cada semana. Esto se refiere a cualquier tipo de vídeo, desde vídeos musicales y tutoriales hasta vídeos de juegos e influencers. Por su parte, las recientes mejoras en la tecnología de vídeo 3D, han reavivado un creciente interés hacia el consumo de este tipo de contenidos como una alternativa para expandir la experiencia del usuario. El vídeo 3D permite tanto percibir la profundidad de una escena en movimiento, como mostrar una escena desde múltiples puntos de vista y esto hace que tenga un amplio rango de aplicación que comprende entre otros: la producción de películas y videojuegos 3D, el desarrollo de sistemas inmersivos, vídeos 360 soportados en la realidad virtual estereoscópica, así como la ingeniería médica, donde es posible el modelado del cuerpo humano y sus órganos entre muchos otros. Sin embargo, debido al ancho de banda variable de las redes utilizadas para entregar el contenido multimedia, no siempre se puede garantizar una experiencia de reproducción fluida y de alta calidad. Lo anterior soporta la idea de que para proporcionar contenidos de vídeo en 3D de alta calidad es importante que los servicios se diseñen y gestionen con base en la calidad percibida por los usuarios en adelante QoE (*Quality of Experience*).

La transmisión de vídeo por canales de comunicación con pérdidas y de ancho de banda variable puede introducir artefactos en el contenido transmitido, y el efecto en el caso del vídeo 3D podría ser mucho más significativo en comparación con la transmisión del vídeo 2D convencional. Si bien los proveedores de servicio se enfocan en última instancia en la arquitectura general del sistema, buscando la infraestructura que mejor les permita entregar de forma confiable y rentable unos contenidos de calidad a clientes con altas expectativas. En términos generales, todas las tareas que intervienen en el proceso de distribución de contenidos de vídeo 3D (Figura 1.1), desde los problemas de calibración relacionados con la captura; las deficiencias introducidas por la codificación del vídeo; los errores o pérdidas durante la transmisión; hasta los efectos de visualización en relación con las pantallas o gafas 3D, pueden introducir efectos que influyan en la QoE de los usuarios.

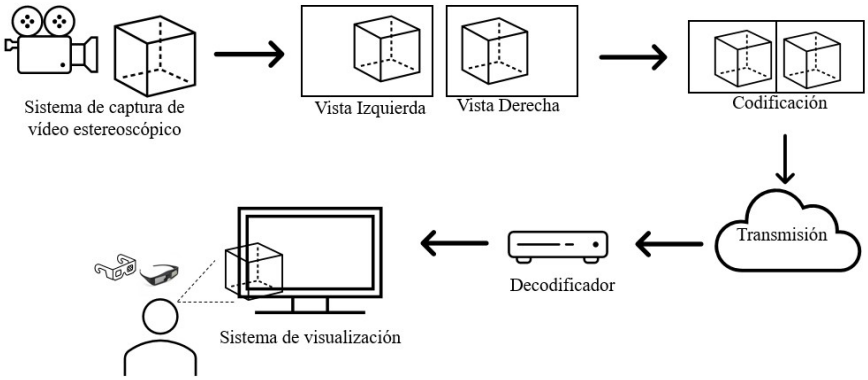


Figura 1.1. Sistema de distribución de vídeo 3D

En esta tesis se estudia cómo en un sistema de distribución de vídeo 3D, las degradaciones o distorsiones debidas a la codificación, así como las interrupciones y degradaciones durante la transmisión, afectan la percepción del usuario final. Para ello se han empleado métricas de evaluación tanto objetivas como subjetivas, y se propone una metodología que permita realizar estas valoraciones de forma sistemática. Las observaciones presentadas pueden servir de base para el diseño de futuros sistemas de distribución de vídeo 3D centrados en la QoE del usuario.

En este capítulo se introducen de forma general los principales conceptos relativos a la cadena de procesamiento de un sistema de distribución de vídeo 3D y su influencia en la evaluación de la QoE en vídeo 3D, estableciendo una base conceptual para el planteamiento de los objetivos de la tesis. Se describen también los objetivos y la estructura de este documento, así como las contribuciones más relevantes que abordan la evaluación automática de la calidad de experiencia del usuario para un sistema de transmisión adaptativa de vídeo 3D.

1.1. Conceptos generales

A pesar de que el vídeo 3D parecía alcanzar su punto máximo una década atrás con los efectos visuales épicos de películas para las salas de cine, como Avatar de James Cameron (2009), Hugo de Martín Scorsese (2011) o la Vida de Pi de Ang Lee (2012), la idea de cambiar la experiencia visual de los usuarios, proporcionando a los espectadores una escena de vídeo más inmersiva y natural, aumentó el interés por explorar la introducción de esta tecnología para el entretenimiento doméstico o las aplicaciones móviles [1][2]. Con esta expectativa, durante la última década se han realizado grandes esfuerzos para mejorar cada uno de los procesos que conforman el sistema de distribución de vídeo 3D como el que se muestra en la Figura 1.1 y que se describe a continuación:

- *Adquisición (Sistema de captura de vídeo estereoscópico):* nuestro cerebro genera una imagen en 3D en gran medida gracias a que tenemos dos ojos separados por una corta distancia (6,3 cm en promedio). Cada ojo capta una imagen ligeramente diferente de la escena que tiene delante, el cerebro fusiona estas dos imágenes y, dado que cada una de ellas está ligeramente desplazada con respecto a la otra, fenómeno conocido como

disparidad retiniana, se perciben las profundidades relativas (distancia percibida entre los objetos) y absolutas (distancia percibida del observador a los objetos) de los objetos en el espacio. La capacidad del cerebro para realizar estos cálculos se denomina estereopsis (o visión estereoscópica). Por tanto, el vídeo estereoscópico o vídeo 3D, es una señal de vídeo con componentes de señal izquierda y derecha cuya adquisición se simplifica mediante el uso de una cámara estereoscópica que facilita la captura sincronizada de esas dos señales. Sin embargo, estas componentes de señal izquierda y derecha también pueden ser generadas a partir de dos vistas marginalmente diferentes, producidas empleando un array de múltiples cámaras sincronizadas o sistema de vídeo multivista, que permite capturar una escena desde diferentes ángulos [3]; o bien mediante los sistemas de cámaras mejoradas con sensores de profundidad, que permiten la representación de una escena 3D a partir de una vista y su correspondiente mapa de profundidad o generados a partir de la conversión de una secuencia de vídeo 2D a la cual se le añade la información de profundidad.

- *Codificación (compresión y encapsulado):* el objetivo principal de la industria y los investigadores ha sido ofrecer una mayor resolución de vídeo y una experiencia multimedia envolvente en los hogares. Si bien se han producido grandes avances tecnológicos en lo que se refiere a los canales de transmisión, el ancho de banda sigue siendo un problema y como paso previo a su transmisión, los flujos de vídeo capturados se comprimen y son encapsulados en un formato de vídeo que sea compatible con la infraestructura disponible [4][5][6].

- *Transmisión:* los datos de vídeo 3D codificados pueden ser transmitidos a un solo receptor o a varios receptores simultáneamente. La prestación de servicios de vídeo 3D, requiere mecanismos de transporte eficientes, que permitan transmitir el volumen de datos de vídeo de las dos vistas con los enlaces disponibles, que generalmente son canales con pérdidas, con ancho de banda y retardos que varían con el tiempo, como las redes basadas en IP y sus extensiones inalámbricas, y los canales de transmisión terrestre y por satélite.

- *Visualización (Decodificación y renderizado):* los datos de vídeo 3D recibidos se descomprimen y se *renderizan* de acuerdo con los parámetros de visualización del usuario. La visualización de vídeo estereoscópico es un caso especial de visualización de vídeo 3D, en el que la profundidad de la escena se renderiza para cada usuario con la ayuda de un dispositivo de visualización especializado: gafas o una pantalla autoestereoscópica que se encargan de que cada ojo reciba la señal correcta.

El vídeo 3D ofrece una nueva experiencia visual a los usuarios mediante la percepción de la profundidad. Sin embargo, algunos usuarios se quejan de incomodidad visual y fatiga al ver vídeo 3D. Para ofrecer contenidos de vídeo 3D de alta calidad, además de los aspectos técnicos asociados a los distintos elementos de la cadena de distribución de vídeo, es importante diseñar y gestionar servicios basados en la QoE de los usuarios.

A diferencia de las imágenes y el vídeo en 2D, una de las principales cuestiones relativas a la evaluación de la QoE en 3D es su multidimensionalidad, ya que además de la calidad de la imagen, se debe evaluar la percepción de la profundidad y el confort visual, que probablemente ha sido un factor decisivo para la aceptación de las tecnologías de vídeo 3D. Así mismo, deben tenerse en cuenta tanto los factores propios del HVS (*Human Visual System*) [7] relacionados con la

percepción estereoscópica, así como aquellos relacionados con el entorno de los usuarios como: la localización, el costo del servicio o sus relaciones sociales, entre otros.

En conclusión, la QoE en vídeo 3D está sujeta a una serie de factores complejos y fuertemente interrelacionados, que han sido denominados factores de influencia [8] y que serán estudiados más adelante en el Capítulo 2 de esta tesis.

1.2. Planteamiento del problema y objetivos

Como ocurre con cualquier otro sistema, servicio o aplicación multimedia, saber si las expectativas de los usuarios finales están satisfechas, es crucial para el éxito y la implantación definitiva de las tecnologías de vídeo 3D en el mercado de consumo. Como se comentó en la sección anterior, cada uno de los procesos presentes en la cadena de distribución de vídeo 3D y los factores de influencia asociados a los mismos, pueden afectar la QoE de los usuarios. Por lo anterior, esta tesis aborda el problema de la prestación de servicios de transmisión de vídeo estereoscópico-3D sobre redes basadas en IP, analizando el impacto en la QoE de los usuarios, de los efectos introducidos en dos de los elementos de la cadena de procesamiento de vídeo 3D: la etapa de codificación y el proceso de transmisión tal como se muestra en la Figura 1.2.

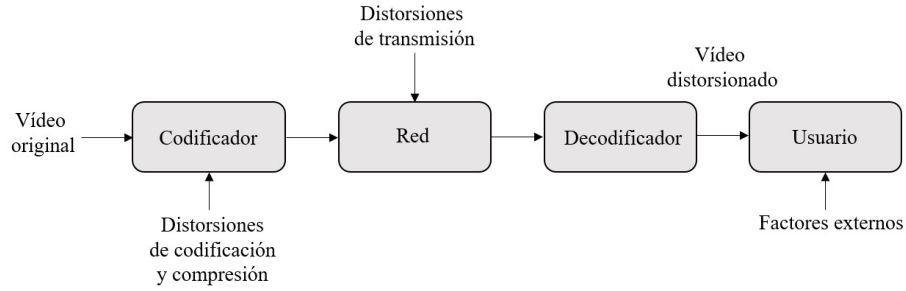


Figura 1.2. Fuentes de distorsión durante la transmisión de vídeo

En este contexto, el objetivo general de esta tesis es:

Evaluar el impacto que sobre la Calidad de Experiencia (QoE) percibida por los usuarios de los sistemas de vídeo 3D tienen: i) los parámetros configurables asociados con la compresión, segmentación y representación de los contenidos; ii) los parámetros de Calidad de Servicio (QoS) propios de la red de distribución basada en HTTP.

Para lograr el objetivo general, los objetivos específicos que se plantean son:

O1. Estudiar las técnicas de codificación de vídeo 3D disponibles en la actualidad y evaluar su impacto en la QoE del usuario en un escenario de distribución de vídeo sobre redes IP en ausencia de pérdidas.

O2. Validar el concepto de codificaciones asimétricas como alternativa para la reducción de la tasa de bits en sistemas de distribución de vídeo 3D.

O3. Evaluar la calidad del vídeo, en un escenario de transmisión adaptativa de vídeo 3D con diferentes condiciones de ancho de banda, retardo y porcentaje de pérdidas.

O4. Analizar la validez de las métricas objetivas de calidad de vídeo para automatizar el cálculo de la QoE en aplicaciones de distribución de vídeo 3D.

1.3. Contribuciones

Las contribuciones principales de esta tesis se resumen a continuación:

1. Desarrollo de un *testbed* para la comparación mediante métricas objetivas de calidad de vídeo, de los estándares de codificación H.264/AVC y H.265/HEVC, incluyendo sus extensiones para codificación de vídeo estereoscópico 3D (H264-MVC y H.265-HEVC 3D).
2. Propuesta de una metodología que permita, mediante los resultados de la evaluación de la calidad de cada una de las vistas que conforman la secuencia de vídeo 3D, seleccionar de forma automática las secuencias que deben estar disponibles en el servidor en un escenario de transmisión adaptativa de vídeo 3D.
3. Desarrollo de un *testbed* para la evaluación subjetiva por parte de los usuarios, de la calidad percibida frente a diferentes factores de compresión, incluyendo la codificación asimétrica del vídeo 3D en un escenario sin pérdidas, ni interrupciones durante la transmisión.
4. Desarrollo de un marco para la evaluación automatizada del rendimiento de los algoritmos de adaptación de los reproductores web multimedia, en el contexto de la transmisión adaptativa de vídeo sobre HTTP.

Además, esta tesis ha dado lugar a la publicación de artículos en revistas científicas y presentaciones en congresos. En el Apéndice A se presenta una lista detallada de las publicaciones derivadas de este trabajo.

1.4. Estructura de la tesis

Esta tesis está organizada en los siguientes capítulos:

- Capítulo 1 — Introducción

En este capítulo se presentan los conceptos básicos relativos a cada uno de los procesos que forman parte de la cadena de distribución de vídeo 3D (adquisición, codificación, transmisión y visualización), así como los principales factores por los que se puede ver afectada la QoE de un usuario de vídeo 3D. Se presentan, además, la motivación, el problema de investigación a resolver y los objetivos de la tesis y sus principales contribuciones.

- Capítulo 2 — Calidad de Experiencia en vídeo 3D

En este capítulo se presentan los conceptos fundamentales asociados a la evaluación de la QoE en escenario de transmisión de vídeo 3D. En primera instancia, se presentan los conceptos básicos asociados a la visión estereoscópica, la definición del término QoE y se exponen los principales factores presentes en la cadena de procesamiento del vídeo 3D, y que pueden afectar la QoE de los usuarios finales. A continuación, se presenta una descripción de las métricas más populares para la evaluación objetiva de la calidad del vídeo, que serán empleadas en el desarrollo de la tesis. Por último, se exponen las diferentes recomendaciones existentes, que definen las pautas

para la realización de pruebas de evaluación subjetiva en el contexto de los sistemas de transmisión de vídeo 3D.

- Capítulo 3 — Codificación de vídeo 3D

En este capítulo se realiza una comparación de la eficiencia en el proceso de compresión de vídeo 3D usando diferentes implementaciones de los estándares H.264-AVC y H.265-HEVC para la codificación por separado de cada una de las vistas del vídeo 3D, así como sus extensiones para la codificación estéreo y multivista (MVC). Así mismo, se valida el uso de las codificaciones 3D asimétricas, como una alternativa para minimizar la tasa de bits, verificando que se mantiene la QoE del usuario al degradar solo la calidad de una de las vistas. La comparación se realiza empleando tanto métricas de evaluación objetivas, como mediante la realización de pruebas subjetivas con usuarios. Finalmente, se propone una metodología basada en los resultados de la evaluación de la calidad, para la selección automática de las representaciones que deben estar disponibles en el servidor para su posterior transmisión.

- Capítulo 4 — Evaluación de la QoE de un sistema de transmisión adaptativa de vídeo sobre HTTP

En este capítulo se propone un marco para la evaluación automática y sistematizada de los algoritmos de adaptación de reproductores web multimedia, en un escenario de transmisión adaptativa de vídeo sobre HTTP. El sistema propuesto se soporta en el uso de herramientas de código libre como *Puppeteer*, la librería Node.js desarrollada por Google, que proporciona una API de alto nivel, para automatizar acciones en el protocolo *Chrome Devtools.*

Capítulo 2

Calidad de Experiencia en vídeo 3D

Los "pros y contras" de la tecnología 3D son en cierto modo obvios. Al introducir la percepción de la profundidad en el vídeo, el vídeo 3D ofrece una nueva experiencia inmersiva a los usuarios. Sin embargo, algunos usuarios se quejan de incomodidad visual y fatiga después de ver las películas en 3D.

En este sentido, la evaluación de la calidad del vídeo y mejorar la experiencia de visualización de los contenidos en 3D han sido y siguen siendo objeto de estudio por muchos autores como [9][10] [11], donde se plantean dos cuestiones básicas: 1) ¿qué factores afectan a la experiencia de visualización de los contenidos 3D? y 2) ¿cómo se puede medir la calidad de los contenidos visualizados y la experiencia del usuario?

Como paso previo a la comparación de los codificadores de vídeo 3D y el estudio de la calidad de experiencia de un sistema de transmisión adaptativa de vídeo sobre HTTP, que se presentarán en los capítulos 3 y 4 respectivamente, en este capítulo, en la sección 2.1 y 2.2, se presentan brevemente los conceptos relacionados con la percepción estereoscópica propia del sistema de visión humano y las tecnologías de visualización 3D respectivamente. En la sección 2.3 se introducen los principales aspectos y terminología relacionados con la evaluación de la Calidad de Experiencia (QoE) o la calidad percibida por el usuario del vídeo 3D. Se ofrece una definición del término QoE, ilustrando su multidimensionalidad, así como la descripción de los principales factores que influyen en ella. Finalmente, en la sección 2.4 se presentan los aspectos generales relativos a las principales métricas y métodos de evaluación de la calidad del vídeo 3D mediante pruebas de evaluación subjetiva y métricas objetivas.

2.1. Visión estereoscópica

La percepción de la profundidad nos permite percibir el mundo que nos rodea en tres dimensiones y calibrar la distancia de los objetos con respecto a nosotros mismos y a otros objetos. La percepción de la profundidad se basa en una serie de señales visuales de profundidad, que según se propone en la literatura [12], pueden clasificarse en dos categorías: señales monoculares y señales binoculares. Como su nombre indica, las señales monoculares requieren

la entrada de un solo ojo, mientras que las binoculares requieren las imágenes percibidas por los dos ojos.

2.1.1. Señales de profundidad monoculares

Dentro de las señales de profundidad monoculares pueden identificarse dos tipos: aquellas donde la información de la profundidad se puede extraer de una imagen estática y aquellas que proporcionan información de profundidad a partir del movimiento. La Figura 2.1 presenta algunas de estas señales.

Figura 2.1. Ejemplos de señales de profundidad monoculares.

- *Interposición:* cuando un objeto se superpone a otro, el objeto parcialmente oculto se percibe como más lejano
- *Luz y sombra:* se suele aplicar para las luces naturales, y para la mayoría de las luces artificiales, que la luz proviene de arriba. Los objetos en sombra pueden parecer más lejanos que los que están bien iluminados.
- *Tamaño relativo:* si dos objetos tienen aproximadamente el mismo tamaño, el objeto que parece más grande será juzgado como el más cercano al observador. Esto se aplica tanto a las escenas tridimensionales como a las imágenes bidimensionales. Dos objetos en un papel están a la misma distancia, pero la diferencia de tamaño puede hacer que el objeto más grande parezca más cercano y el más pequeño más lejano.
- *Altura en el campo visual:* se trata de una señal de profundidad basada en la posición vertical de un punto en el campo visual. Los objetos más lejanos suelen estar más altos en el campo visual.

- *Gradiente de textura:* la mayoría de las superficies, como los muros y las carreteras o un campo de flores, tienen una textura. A medida que la superficie se aleja del observador esa textura se vuelve más fina y parece más suave. Por tanto, la forma, el tamaño y la densidad de la textura también afectan a la percepción de la profundidad
- *Perspectiva lineal:* las líneas paralelas parecen encontrarse a medida que se adentran en la distancia. Cuantas más juntas estén las dos líneas, mayor parecerá la distancia.
- *Perspectiva aérea:* debido a los componentes atmosféricos (niebla, polvo o lluvia), los objetos más lejanos se perciben menos nítidos y con un ligero tinte azul.
- *Paralaje de movimiento:* es una señal de profundidad que existe en una escena dinámica. Cuando nos movemos o un objeto de la escena se mueve, los objetos que están más cerca de nosotros se mueven más rápido a través de nuestro campo de visión que los objetos en la distancia.

2.1.2. Señales de profundidad binoculares

La principal característica de las señales binoculares es que dependen de los dos ojos, así trabajan en las tres dimensiones.

- Disparidad binocular (estéreo): como los ojos humanos están separados horizontalmente (aproximadamente 6 cm), las imágenes retinales recibidas por los dos ojos son ligeramente diferentes. El cerebro es capaz de fusionar ambas imágenes en una sola imagen 3D, extrayendo la información de profundidad de la escena. Cuanto más lejos esté un objeto en 3D, más separadas están las dos imágenes.

Además de las señales visuales monoculares y binoculares, encontramos algunas otras asociadas al movimiento de los músculos de los ojos cuando se está observando un objeto. Éstas son la convergencia y la acomodación.

- Acomodación: ajustar el enfoque de un objeto en la escena mediante la tensión/relajación de los músculos oculares ciliares mejora la percepción de la profundidad. Eficaz para distancias inferiores a 2 metros.
- Convergencia: los dos ojos convergen cuando miran un mismo punto en un objeto 3D simultáneamente. Según el principio de triangulación, cuanto más cerca esté el objeto, más deben converger los ojos. Es eficaz para distancias inferiores a 10 metros. Al mirar un objeto a una distancia infinita, los ojos divergen hasta ser paralelos.

2.2. Sistemas de visualización 3D

Las pantallas 3D aprovechan los mecanismos del HVS (*Human Visual System*) en la percepción de la estereopsis. El cerebro humano puede fusionar las imágenes procedentes del ojo izquierdo y del ojo derecho, a partir de la disparidad retiniana y el HVS extrae la información de profundidad relativa, es decir, la distancia entre los puntos correspondientes en estas imágenes [13]. Así, la técnica básica de las pantallas 3D consiste en presentar las vistas izquierda y derecha por separado al ojo izquierdo y al ojo derecho. A continuación, las dos imágenes de la retina pueden combinarse en el cerebro para generar la percepción de la profundidad.

En función de si se requiere o no el uso de gafas, las pantallas 3D pueden clasificarse en dos tipos: estereoscópicas y autoestereoscópicas. Cuando se utiliza una pantalla estereoscópica, los

espectadores tienen que llevar un dispositivo óptico para dirigir las imágenes izquierda y derecha al ojo correspondiente, lo que se denomina visión asistida. En las pantallas autoestereoscópicas la tecnología de separación de ambas vistas está integrada en la pantalla, lo que se denomina visión libre.

Las pantallas estereoscópicas con visión asistida son muy utilizadas. Se clasifican en paralelas (*time-parallel*) o secuenciales (*time-sequential*). Con la tecnología paralela, las vistas izquierda y derecha se muestran simultáneamente en una o dos pantallas. En estos sistemas, los métodos más utilizados para dirigir cada vista al ojo apropiado, tal como se muestra en la Figura 2.2. son: 1) la multiplexación por color o anaglifo y 2) la multiplexación por polarización.

(a) (b)

Figura 2.2. Gafas pasivas sistemas de vídeo 3D. (a) Anaglifo. (b) Polarizadas

Empleando tecnología secuencial, las vistas izquierda y derecha de una secuencia estereoscópica se presentan en rápida alternancia. Los pares estereoscópicos se visualizan mediante gafas de enmascaramiento sincronizadas, más conocidas como gafas activas, que se abren y se cierran alternativamente para el ojo correspondiente. Este sistema aprovecha la característica del HVS que integra un par estereoscópico a través de un desfase temporal de hasta 50 ms [14].

A diferencia de las gafas activas, las gafas pasivas no requieren de una fuente de energía o baterías para poder ver el contenido en 3D, ya que no realizan ningún tipo de acción mecánica. Su principal ventaja respecto a los sistemas activos reside en las propias gafas, que son baratas y no necesitan sincronización con la pantalla. Entre las desventajas están la disminución de la resolución vertical a la mitad, y que la eficiencia de transmisión es menor, por lo que la imagen emitida es más oscura que la original.

A diferencia de las pantallas estereoscópicas, las pantallas autoestereoscópicas no necesitan gafas para presentar las dos vistas. Este tipo de pantallas envía las imágenes izquierda y derecha directamente al ojo correspondiente. Actualmente, las pantallas autoestereoscópicas pueden clasificarse en 1) multidireccionales; 2) volumétricas; y 3) holográficas. Para más detalles, se pueden consultar en [15].

2.3. Calidad de experiencia (QoE) en vídeo 3D

La calidad del vídeo 3D, muy asociada a la QoE de los usuarios, puede verse afectada por diversos niveles de degradación, propios de cada una de las etapas de un sistema de distribución de vídeo 3D. Hablamos de que el vídeo recibido por el usuario puede tener degradaciones debidas a la

producción, la compresión, la transmisión o la visualización. Por lo tanto, la investigación de la evaluación de la calidad del vídeo estereoscópico en cada una de estas etapas, y su impacto en la calidad percibida por el usuario desempeña un papel importante en el desarrollo de los sistemas de vídeo estereoscópico.

2.3.1. Definición de Calidad de Experiencia (QoE)

Dada la relación que existe entre ellos, los términos QoE (*Quality of Experience*) y QoS (*Quality of Service*) han sido usados durante mucho tiempo en la literatura de manera indistinta. Sin embargo, debe quedar claro que en la actualidad estos dos conceptos se encuentran separados. Se entiende que la QoS es un concepto puramente técnico, que se expresa y se define mejor en términos del rendimiento de la red/sistema y los elementos que los conforman, tal como se refleja en la definición dada por la ITU (*International Telecommunication Union*) [16], donde se define la QoS como "*La totalidad de las características de un servicio de telecomunicaciones que determinan su capacidad para satisfacer las necesidades explícitas e implícitas del usuario del servicio*". Esta definición, centrada explícitamente en el punto de vista del proveedor de servicios, no cubre muchos factores involucrados en los sistemas de transmisión de vídeo. Por ello, tras un exhaustivo trabajo para formular una mejor definición, incluyendo los efectos subjetivos (ej. expectativas del usuario y contexto) a los factores del sistema, se ha dado lugar a nuevas definiciones de este concepto. Se entiende como la definición más aceptada y precisa de la QoE la convenida por un consorcio internacional de instituciones de investigación en Qualinet 2012 [17] donde, tras desarrollar un completo marco teórico en relación con la evaluación de la calidad de los servicios multimedia, se dio lugar a la definición más precisa hasta el momento "*QoE es la medida del agrado o molestia del usuario de una aplicación o servicio. Es el resultado del cumplimiento de sus expectativas con respecto a la utilidad y/o el disfrute de la aplicación o el servicio a la luz de la personalidad y el estado actual del usuario*".

En el caso del servicio del vídeo 3D, la relación entre la QoE y la QoS asociada con los parámetros de codificación y las estadísticas de la red es más compleja, porque la calidad percibida por los usuarios es subjetiva y comprende una variedad de atributos, que van más allá de la calidad de la imagen, en general involucra aspectos como la naturalidad, la sensación de presencia, la profundidad, la inmersión, el confort visual, etc.

2.3.2. Calidad de Experiencia del vídeo 3D

En la actualidad, los usuarios pueden disfrutar de la visualización de contenidos 3D en terminales como televisores 3D, ordenadores personales y teléfonos inteligentes, gracias a los recientes avances realizados en estos dispositivos. Para ofrecer contenidos de vídeo 3D de alta calidad, además de los aspectos técnicos asociados a los distintos elementos de la cadena de distribución de vídeo, es importante diseñar y gestionar servicios basados en la QoE de los usuarios. Para la evaluación de la QoE, deben tenerse en cuenta tanto los factores propios del HVS relacionados con la percepción estereoscópica, así como aquellos relacionados con el entorno de los usuarios como: la ubicación, el costo del servicio o sus relaciones sociales, entre otros.

En conclusión, la QoE está sujeta a una serie de factores complejos y fuertemente interrelacionados, que han sido denominados factores de influencia (IF) [8] y que se dividen en

tres categorías [18]: factores de influencia del sistema, factores de influencia humanos y factores de influencia de contexto.

2.3.2.a. Factores de influencia humanos (FIH)

Los factores de influencia humanos son aquellas propiedades que pueden influir en la QoE, que son subjetivas e intrínsecas de los usuarios humanos, lo que las hace altamente complejas. En esencia puede ser cualquier propiedad o característica variante o invariable asociada al usuario. Dicha característica puede describir el contexto demográfico o socio-económico, la constitución física o mental o el estado emocional del usuario.

2.3.2.b. Factores de influencia del sistema (FIS)

Los factores de influencia del sistema son aquellas propiedades técnicas que influyen en la calidad de una aplicación o servicio y pueden estar relacionadas con el contenido, el hardware/software empleado para la adquisición, codificación/decodificación, transmisión o visualización.

Factores de influencia en el proceso de adquisición (Captura y edición): el contenido en sí mismo y características como: la tasa de bits, la cantidad de movimiento o el grado de profundidad, pueden influir notablemente en la experiencia visual de los usuarios finales, en esta etapa influyen también los aspectos asociados relativos a la captura, como el formato, la orientación y posición de la cámara, así como los parámetros de configuración de esta. Como se muestra en la Figura 2.3, dentro de las degradaciones típicas en esta etapa encontramos: el efecto teatro de marionetas (*Puppet Theater*), que hace que una imagen se vea antinaturalmente pequeña dentro de una escena; el efecto *Cardboard,* que hace que las imágenes 3D parezcan estratificadas, es decir, formadas por objetos planos sobre un fondo plano; y el efecto asíncrono, tanto espacial como temporal, entre las vistas izquierda y derecha.

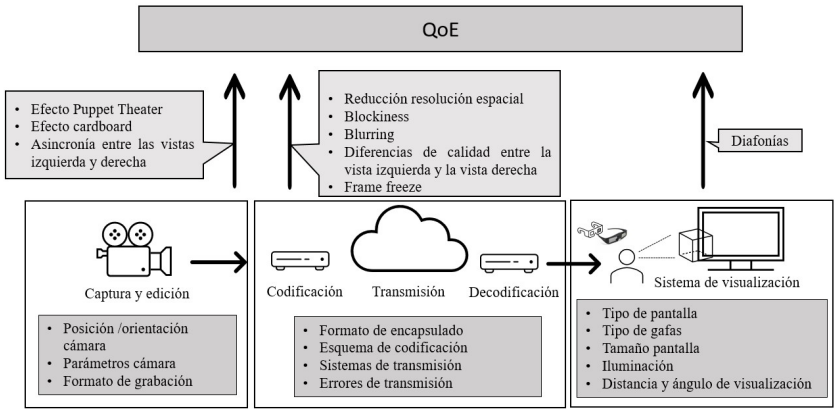


Figura 2.3. Factores de influencia de la QoE

Factores de influencia en el proceso de codificación: en segundo lugar, se encuentran los aspectos relativos al procesamiento del vídeo previo a la transmisión, aquí encontramos factores como el

formato de encapsulado, el esquema de codificación, la resolución o la tasa de fotogramas entre otros.

Factores de influencia en transmisión: por su parte, los factores de influencia asociados a la red incluyen aquellas propiedades inherentes al sistema de transmisión, como el ancho de banda, el retardo, el *jitter* o la tasa de pérdidas. Estas características relacionadas con la red pueden cambiar con el tiempo o a medida que un usuario cambia de ubicación, y están estrechamente relacionadas con la QoS de la red. De acuerdo con la Figura 2.3, dentro de las degradaciones típicas que se pueden introducir en esta etapa se encuentran: la reducción de la resolución espacial, la percepción de una imagen compuestas por bloques (*Blockiness*) o desenfocada (*Blurring*), diferencias de calidad entre la vista izquierda y la vista derecha, y el estancamiento o interrupción del flujo de vídeo, entre otros.

Factores de influencia en visualización: por último, tenemos los factores relacionados con los dispositivos finales del sistema o dispositivos de **visualización**. Cuando un usuario ve un contenido de vídeo en 3D, ve una imagen formada por dos imágenes de vídeo, vistas por separado por el ojo izquierdo y el derecho a través de unas gafas estereoscópicas. Las degradaciones típicas en este contexto son las diafonías producidas por el uso de gafas estereoscópicas. Como se comentó anteriormente, en un sistema de vídeo 3D la interfaz visual para el usuario es la pantalla autoestereoscópica o la combinación pantalla estereoscópica y gafas. La capacidad de estos dispositivos tendrá un tremendo impacto en la experiencia del usuario final. Por ejemplo, si una imagen de alta resolución (*pixels*) se muestra en una pantalla de baja resolución con pocos colores, gran parte de la información original de la imagen podría perderse. Sin embargo, si una imagen de baja resolución se muestra en una pantalla de alta resolución, lo más probable sería que el usuario perciba una imagen pixelada y borrosa, aunque el resultado final dependerá de las características de la pantalla y procedimiento final de escalado que realice sobre la imagen.

2.3.2.c. Factores de influencia del contexto (FIC)

Los factores de influencia del contexto son aquellos aspectos que describen el entorno de los usuarios, en términos de características físicas (la ubicación), temporales (la experiencia), sociales (las relaciones interpersonales durante el uso de la aplicación), económicas (los costes del servicio) y técnicas (la interacción entre el sistema y otros sistemas). Así, algunos de estos factores están fuertemente ligados a los factores de influencia humanos y del sistema, y son muy difíciles de cubrir en la evaluación de la QoE.

A diferencia de las imágenes y el vídeo en 2D, una de las principales cuestiones relativas a la evaluación de la QoE en 3D es su multidimensionalidad, ya que además de la calidad del vídeo, hay otros agentes que intervienen en la experiencia visual global en 3D. En esencia, la experiencia visual 3D viene determinada principalmente por la calidad de la imagen, la percepción de la profundidad y el confort visual. Además, hay que tener en cuenta algunos factores secundarios, como la naturalidad y la sensación de presencia, como se muestra en la Figura 2.4.

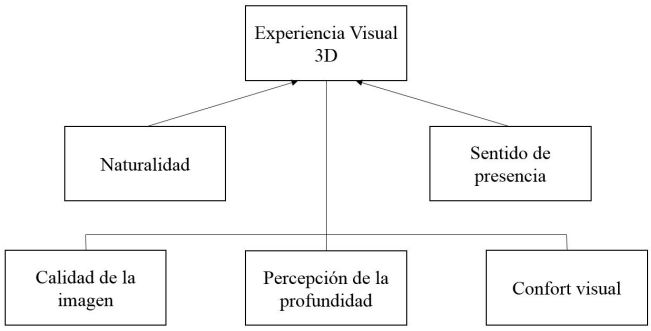


Figura 2.4. Modelo QoE para la evaluación del vídeo 3D

2.4. Métricas de evaluación de la calidad de vídeo

Aunque los primeros enfoques para evaluar la QoE de los usuarios de las tecnologías de vídeo 3D se han basado en los métodos y técnicas desarrollados para el vídeo 2D convencional, los nuevos factores que influyen en la QoE del vídeo 3D en comparación con el vídeo 2D (por ejemplo, la percepción de la profundidad, el confort visual, la naturalidad, la inmersión, las condiciones de visualización, etc.) hacen necesario revisar los métodos existentes y desarrollar nuevos enfoques para obtener evaluaciones fiables de la experiencia visual 3D. Así, en esta sección se presentan los avances en las alternativas de evaluación de la QoE 3D, tanto mediante pruebas de evaluación subjetiva (basadas en las opiniones directas de una serie de observadores sobre el sistema o la aplicación en estudio) como mediante métricas objetivas (que tratan de estimar automáticamente la calidad que percibirían los usuarios finales).

Como se comentó anteriormente, este trabajo se enfoca en la evaluación de la QoE en dos puntos concretos del proceso de distribución de vídeo 3D: la etapa de codificación y la transmisión. En la primera etapa, las métricas y métodos de evaluación de la calidad de vídeo están orientados a la comparación de los codificadores de vídeo 3D, mientras que, en la segunda, el objetivo se centra en evaluar la calidad del vídeo y, por tanto, la QoE en un escenario de transmisión adaptativa de vídeo sobre HTTP.

Al evaluar la eficiencia de un codificador se debe considerar que es importante comprimir la señal de vídeo tanto como sea posible manteniendo su calidad. Por lo general, encontramos dos categorías o métodos de medición de la calidad de vídeo: 1) la calidad objetiva que se mide utilizando modelos matemáticos que pueden ser automatizados; y 2) la calidad subjetiva que cuantifica la calidad del vídeo usando evaluadores humanos. La evaluación subjetiva de la calidad de vídeo resulta costosa en términos del tiempo requerido y la necesidad de evaluadores humanos y, además, puede resultar ineficiente en escenarios que involucren aplicaciones en tiempo real. Por esta razón, las métricas de calidad objetiva pueden ser una alternativa adecuada en un escenario de transporte adaptativo, donde es deseable tener una referencia de la calidad percibida por los usuarios a efectos de optimizar la red.

Si bien se ha trabajado mucho en la evaluación de la QoE para el streaming de vídeo 2D [19][20], la evaluación de la QoE para los servicios de transmisión de vídeo 3D sigue siendo un campo abierto de investigación, que ha ido avanzando en los últimos años [21][22][23]. De forma general, seguimos los aspectos típicos para evaluar la QoE: evaluación objetiva [24] y evaluación subjetiva [25].

2.4.1. Métricas objetivas de evaluación de la calidad del vídeo

Teniendo en cuenta que el vídeo estereoscópico 3D se puede procesar como dos secuencias o vistas independientes ($V_{Izquierda}$ y $V_{Derecha}$), la evaluación objetiva se basa en métricas bien conocidas que ya se han aplicado al vídeo 2D, como PSNR (*Peak Signal to Noise Ratio*) [26] , SSIM (*Structural Similarity Index*) [27] [28] y el más reciente sistema de control de calidad de vídeo VMAF (*Video Multimethod Assessment Fusion*), que se utiliza para controlar la calidad de la imagen de todos los vídeos codificados transmitidos por Netflix [29]. La implementación de las métricas de evaluación objetiva es simple, tiene bajo coste computacional y las medidas pueden reproducirse. Tanto el PSNR como el SSIM y el VMAF son métricas del tipo *full-reference*, es decir, se basan en la disponibilidad de la señal original, que es contrastada con la señal degradada, *frame* a *frame*.

PSNR representa la relación en decibelios entre la potencia máxima de una señal y la potencia del ruido no deseado que afecta a la fidelidad de su representación y se calcula de la siguiente manera:

$$PSNR = 10 Log_{10} \frac{255^2}{MSE}$$

$$MSE = \frac{1}{NM} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left| I_{org}(m,n) - I_{dec}(m,n) \right|^2 \quad (1)$$

Como se muestra en la ecuación (1), PSNR se mide con base en el error cuadrático medio (MSE) entre dos imágenes, de las cuales, una imagen es la imagen original y la otra es la imagen decodificada. M, N es el tamaño de la imagen. Los valores altos de PSNR indican que las imágenes de entrada y de salida son similares. Para los casos típicos, los valores de PSNR van desde 30 dB a 50 dB, valor a partir del cual las diferencias del vídeo codificado respecto al original se consideran imperceptibles. El PSNR típicamente solo se calcula para el canal de luminancia (Y-PSNR).

El índice SSIM es otra métrica de referencia completa, diseñada para medir la similitud entre dos imágenes. SSIM mide la calidad de la imagen decodificada mediante la comparación *frame* a *frame* de tres componentes (similitud de luminancia, similitud de contraste y similitud estructural) con base en una imagen sin comprimir o sin distorsión inicial como referencia. Los resultados se presentan en una escala de 0 a 1, siendo 1 la máxima calidad. SSIM está diseñado para mejorar los métodos tradicionales, como el PSNR y el MSE, que han demostrado ser menos consistentes con la percepción del ojo humano respecto a las variaciones en la luminosidad y contraste que pueden no afectar mucho a la calidad de la imagen, obteniendo valores más cercanos a la QoE del usuario.

La ecuación (2) corresponde al cálculo de SSIM

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_{2)}}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

$$SSIM(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma}$$

Y por último, VMAF predice la calidad subjetiva del vídeo basándose en una secuencia de vídeo de referencia y distorsionada. VMAF fue formulado específicamente por Netflix para correlacionar fuertemente con las puntuaciones subjetivas de MOS (*Mean Opinion Score*). Utilizando técnicas de aprendizaje automático, se utilizó una gran muestra de puntuaciones MOS para entrenar un modelo de estimación de la calidad. Se trata de una métrica *full-reference* que pretende aproximarse a la percepción humana de la calidad del vídeo. Esta métrica se centra en la degradación de la calidad debida a la compresión y al cambio de escala para la visualización. El VMAF estima la puntuación de la calidad percibida calculando las puntuaciones de múltiples algoritmos de evaluación de la calidad y fusionándolos mediante una máquina de soporte vectorial (SVM). Para predecir cuál es la percepción del usuario frente a un vídeo en concreto, VMAF combina 3 métricas de calidad: fidelidad visual, pérdida de detalle y efectos de movimiento.

- Fidelidad de la Información Visual (*Visual Información Fidelity* o VIF): métrica de calidad de imagen que se mide como una pérdida de fidelidad combinando cuatro escalas espaciales.
- Pérdida de detalle (*Detail Loss Metric* o DLM): métrica basada en la lógica de medir por separado la pérdida de datos que afecta a la visibilidad de contenido, y el deterioro redundante que distrae la atención del espectador.
- Efectos de Movimiento: medida de la diferencia temporal entre *frames* sobre la componente de la luminancia.

El VMAF se ha integrado en herramientas de análisis de vídeo como FFMPEG, Elecard StreamEye y MSU Video Quality Measurement Tool, poniéndose a la par de las métricas más populares como PSNR y SSIM. Además, este software de Netflix es de código abierto y se puede conseguir en Github por parte de otras empresas para que estas puedan de igual forma probar sus vídeos y solucionar los problemas que pueden generarse.

En VMAF los valores obtenidos para cada *frame* son promediados para ofrecer un único valor en una escala de 0-100, donde 100 indica que el vídeo evaluado es idéntico al original. El aprendizaje automático en el que se basa VMAF indica que puede ser cada vez más preciso con el tiempo. La Figura 2.5 muestra la escala establecida para cada una de las métricas comentadas, considerando que tienen escalas diferentes.

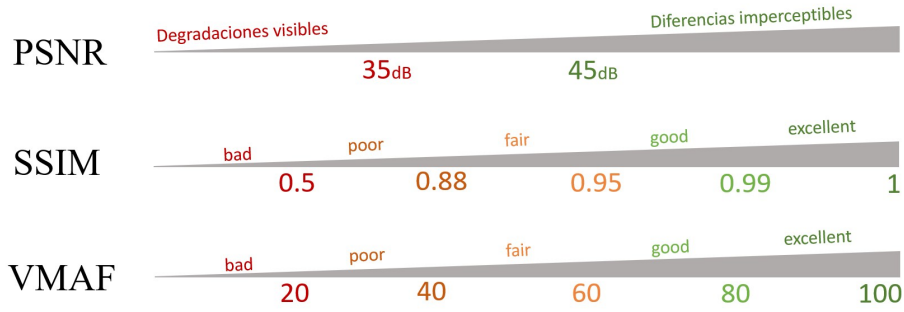


Figura 2.5. Escalas de métricas objetivas

Las métricas objetivas se pueden obtener de forma automática, sencilla y rápida. Sin embargo, su principal desventaja es que es necesario disponer de la imagen original y, en algunos casos, estas métricas tienen una muy baja correlación con la QoE del usuario. Ésta es la principal razón por la que se requieren las métricas de evaluación subjetiva.

2.4.2. Estándares para la evaluación subjetiva de la calidad de vídeo

La forma más sencilla de evaluar la calidad de un vídeo es mostrarlo a los usuarios y preguntarles su opinión (puntuación). Es por eso por lo que este tipo de evaluación de la calidad se denomina subjetiva, porque depende de la opinión del sujeto o usuario. Para eliminar la influencia de las desviaciones individuales de los usuarios en la evaluación, por ejemplo, lo que les gusta y lo que no les gusta, hay que preguntar a muchas personas y promediar sus opiniones (puntuaciones). Este tipo de métrica se ha definido como MOS (*Mean Opinion Score*). A lo largo de los años se han desarrollado muchos procedimientos diferentes de evaluación subjetiva de la calidad. Se diferencian en:

- Aspecto que se va a evaluar (por ejemplo, calidad, distorsión, fidelidad de la imagen).
- Las condiciones de la prueba (por ejemplo, con o sin referencia).
- El tratamiento de los datos (análisis estadístico).

Para obtener resultados fiables, todas las evaluaciones deben ser realizadas en condiciones definidas con precisión. Existen diferentes recomendaciones publicadas por la ITU-T que permiten unificar los criterios para la realización del proceso. Los temas abordados en estas recomendaciones incluyen: métodos de evaluación, escalas subjetivas, condiciones ambientales, distancia de visualización, tamaño de la pantalla y análisis de datos entre otros aspectos que se deben contemplar para la realización de los experimentos. Estos experimentos pueden responder a diferentes cuestiones, como la calidad del vídeo, la calidad de la profundidad, la naturalidad, la incomodidad visual, la calidad de la experiencia, la experiencia de visualización y la presencia.

Una de estas recomendaciones es la UIT-R BT.500-13 [30], cuyo ámbito de aplicación se limita a los contenidos de vídeo en 2D. Una extensión de este protocolo a sistemas estereoscópicos de 3DTV ha sido desarrollada por la UIT y está disponible como recomendación ITU-R BT.2021 [31], a la que se han añadido recientemente otras tres recomendaciones relacionadas con el vídeo 3D:

UIT-R P.914 Requisitos de visualización para la evaluación de la calidad de vídeo 3D [32], ITU-R P.915 Métodos de evaluación subjetiva de la calidad de vídeo 3D [33], y ITU-R P.916 Información y directrices para evaluar y minimizar la incomodidad visual y la fatiga visual del vídeo 3D [34]. En la evaluación subjetiva de la calidad de vídeo 2D, los observadores califican el vídeo en una única dimensión que cuantifica la calidad, pero en la evaluación subjetiva de la calidad del vídeo en 3D se utilizan otros indicadores de calidad específicos del 3D, como la calidad de la profundidad y el confort visual. Esto significa que para cada secuencia de vídeo 3D, los sujetos de prueba tienen que indicar tres calificaciones diferentes, a diferencia de una en el caso del vídeo 2D, lo que hace que el procedimiento de evaluación sea más engorroso y propenso a la variabilidad entre observadores.

La profundidad binocular añadida que introduce la tecnología 3D puede proporcionar a los espectadores no sólo una experiencia visual totalmente diferente y mejorada, sino también molestias y fatiga visuales. Este tipo de trastornos, contemplan una amplia gama de síntomas, por ejemplo, dolores de cabeza, cansancio, tensión ocular. Estas molestias son objeto de una atención cada vez mayor ya que, además de la disminución de la experiencia visual, están relacionadas con la salud y la seguridad de los espectadores.

2.4.3. Metodologías de evaluación subjetiva de la calidad de vídeo

Medir la calidad percibida por los usuarios mediante pruebas de evaluación subjetiva, requiere el uso de métodos, escalas de ponderación, equipamiento y condiciones específicas para su implementación. La elección del método apropiado dependerá de la aplicación u objetivos de la evaluación. A continuación, se presentará una breve descripción de las recomendaciones y los métodos de evaluación subjetiva más empleados en el análisis de la calidad de vídeo 3D.

2.4.3.a. Recomendación ITU-T P.915 - Métodos de evaluación subjetiva de la calidad de vídeo 3D

Esta recomendación describe la evaluación subjetiva del vídeo 3D. Incluye aspectos como: selección de contenidos, los métodos de evaluación, las escalas subjetivas, las condiciones del entorno, distancia de visualización, el tamaño de la pantalla y las indicaciones para hacer el análisis de los datos. Estos experimentos permiten obtener información sobre diversos aspectos como, la calidad del vídeo, la calidad de la profundidad, la naturalidad, la molestia visual, la sensación de presencia y la experiencia visual.

Selección de contenidos 3D: se especifica que las secuencias seleccionadas deben cubrir un amplio rango de vídeos 3D y se recomienda incluir vídeos con diferentes texturas, grados de profundidad y movimiento. Así mismo, se resalta la importancia de factores como la información espacial y temporal de la escena, considerando que estos parámetros juegan un papel muy importante en el proceso de compresión y, en consecuencia, en el grado de degradación sufrido por una secuencia cuando debe ser transmitida por un canal con una tasa de bits fija.

Duración del estímulo: según el objetivo de la prueba y del método seleccionado se recomienda el uso de secuencias en un rango entre 5 y 20 s. Con el objetivo de limitar la duración de las pruebas, se considera que 10 s es una medida ideal.

Distancia y ángulo de visión: la distancia de visualización debe ser de aproximadamente 3H (tres veces la altura de la imagen, H) para entornos de TV. Para monitores de PC, se recomienda de 1H

a 3H. Para aplicaciones multimedia (por ejemplo, dispositivos móviles), se recomienda de 6H a 10H. En el caso de las pantallas móviles, el sujeto ajustará la distancia de visualización según sus preferencias, el tamaño de la pantalla y la calidad del contenido. Por lo tanto, a efectos prácticos, los sujetos no están limitados cuando ven el contenido en sus dispositivos móviles, mientras que sí lo están cuando ven la televisión u otras pantallas fijas.

Número de usuarios: el número de usuarios para la realización de las pruebas es extremadamente importante. El número de pruebas necesarias variará en función de la metodología de la prueba subjetiva. Para los experimentos realizados en un entorno controlado, deben utilizarse 28 sujetos. Esto significa que, tras la selección de los sujetos, cada estímulo debe ser valorado por al menos 28 sujetos [35].

Duración del experimento: preferiblemente, un experimento debe diseñarse de forma que la participación de cada sujeto se limite a 1,5 horas, de las cuales no se emplee más de 1 hora en calificar los estímulos. Cuando se requieran experimentos más largos (por ejemplo, 3, horas de calificación de estímulos), se deben utilizar descansos frecuentes y una compensación adecuada para contrarrestar los impactos negativos de la fatiga y el aburrimiento.

Métodos de evaluación: las pruebas de evaluación subjetiva para vídeo 3D, permiten medir la opinión de los usuarios sobre diferentes escalas perceptivas relacionadas con: la experiencia visual, calidad de la imagen, confort visual y calidad de la profundidad.

La Figura 2.6 presenta un resumen de las principales características y escala de valoración de los tres métodos principales definidos para la evaluación subjetiva del vídeo 3D y que serán empleados más adelante en este trabajo. Para mayor información y detalle sobre cada uno de estos métodos y muchos otros se pueden consultar en la recomendación ITU-R BT.2021-1 (2015) [31].

<table>
<tr><th>ACR (Absolute Category Rating)</th><th>DCR (Degradation Category Rating) o DSIS (Double Stimulus Impairment Scale)</th><th>PC (Pair Comparison) – DSCS (Double Stimulus Comparison Scale)</th></tr>
<tr><td>• Método de estímulo único.
• Las secuencias se presentan y evalúan de una en una.
• Se recomienda para la evaluación de codificación y degradaciones espaciales.
• Se pueden obtener un gran número de evaluaciones en poco tiempo.</td><td>• Método de doble estímulo.
• Las secuencias se presentan por pares.
• El primer estímulo es la referencia y el segundo la secuencia degradada o que se quiere evaluar.
• El usuario debe emitir una ponderación sobre el grado de degradación percibido en la segunda secuencia respecto a la de referencia.</td><td>• Método de doble estímulo.
• Las secuencias se presentan por pares.
• Puede usarse para comparar el vídeo original con el vídeo degradado o dos vídeos con degradaciones diferentes.
• Las secuencias se presentan en orden aleatorio.
• Los usuarios deben evaluar el segundo estímulo con relación al primero.</td></tr>
<tr><td>
<table>
<tr><th colspan="2">Escala ACR</th></tr>
<tr><td>5</td><td>Excelente</td></tr>
<tr><td>4</td><td>Bueno</td></tr>
<tr><td>3</td><td>Regular</td></tr>
<tr><td>2</td><td>Pobre</td></tr>
<tr><td>1</td><td>Malo</td></tr>
</table>
</td><td>
<table>
<tr><th colspan="2">Escala de degradación DCR-DSIS</th></tr>
<tr><td>5</td><td>Imperceptible</td></tr>
<tr><td>4</td><td>Perceptible pero no molesto</td></tr>
<tr><td>3</td><td>Ligeramente molesto</td></tr>
<tr><td>2</td><td>Molesto</td></tr>
<tr><td>1</td><td>Muy molesto</td></tr>
</table>
</td><td>
<table>
<tr><th colspan="2">Escala de degradación PC-DSCS</th></tr>
<tr><td>-3</td><td>Mucho peor</td></tr>
<tr><td>-2</td><td>Peor</td></tr>
<tr><td>-1</td><td>Ligeramente peor</td></tr>
<tr><td>0</td><td>Igual</td></tr>
<tr><td>1</td><td>Ligeramente major</td></tr>
<tr><td>2</td><td>Mejor</td></tr>
<tr><td>3</td><td>Mucho mejor</td></tr>
</table>
</td></tr>
</table>

Figura 2.6. Métodos evaluación subjetiva vídeo 3D

Los métodos y técnicas descritos no pueden, por su propia naturaleza, responder a las necesidades de cada experimento subjetivo. Se da libertad para modificar el método de ensayo para adaptarlo a un experimento concreto. Tales modificaciones entran en el ámbito de la Recomendación. Algunas de las modificaciones más usadas se basan en la inclusión de una

"Referencia" oculta, es decir, incluir una secuencia sin degradar en el grupo de secuencias que deben ser valoradas por los usuarios.

La valoración dada por un usuario para cada secuencia examinada se denomina "*opinion score*". La media de estas puntuaciones, obtenida generalmente para cada sistema investigado, se denomina MOS. La inclusión de la "Referencia" permite calcular la "*difference opinion score*", que es la diferencia aritmética entre las puntuaciones dadas a las secuencias degradadas y "Referencia" de cada secuencia del estudio. La media de estas valoraciones se denomina DMOS (*Difference Mean Opinion Score).*

Puesto que las condiciones de visualización y las características de los dispositivos y tecnologías empleadas pueden tener un impacto directo en la valoración de los usuarios. La información relativa a los dispositivos y condiciones de visualización se abordan con más detalle en la recomendación ITU-R P.914 Requisitos de visualización para la evaluación de la calidad de vídeo 3D [32].

2.4.3.b. Recomendación ITU-T P.916 - Información y directrices para evaluar y minimizar la incomodidad y la fatiga visuales producidas por el vídeo 3D

Esta recomendación incluye información y directrices para evaluar y evitar las molestias y la fatiga visuales provocadas por los vídeos en 3D. Describe las posibles causas de la incomodidad visual y los síntomas de la fatiga visual 3D, incluidos los problemas causados por la visualización de la televisión estereoscópica y autoestereoscópica. Esta información está pensada principalmente para el diseño y la realización de evaluaciones subjetivas de la calidad de vídeo 3D.

Entre los aspectos abordados en este documento y como complemento a la información dada en la recomendación ITU-R P.914, se tratan temas relacionados con el entorno de visualización 3D, profundizando en la relación entre el espectador y el *display* 3D y dando pautas sobre las distancias de visualización y las zonas de confort.

Así mismo, se presentan los síntomas que pueden indicar fatiga visual 3D [36], y que deben ser tenidos en cuenta en el diseño de las pruebas. Algunos de ellos son:

- Dificultad para seguir el movimiento en la televisión.
- Fatiga ocular: visión borrosa, ojos secos, dolor de ojos, escozor, ojos pesados, nebulosos, ojos calientes, parpadeo y ojos llorosos.
- Dificultad para enfocar: visión doble, dificultad para ver de cerca, dificultad para ver de lejos y problemas para fusionar imágenes estereoscópicas.
- Síntomas de fatiga general: sensación de pesadez en la cabeza, dificultad de concentración, mareos, rigidez de hombros y de cuello.
- Dolor de cabeza: dolor en la sien y dolor en el centro de la frente.
- Náuseas: vómitos y vértigo.
- Oscurecimiento visual transitorio después de ver la televisión.

La recomendación indica que, si una persona siente alguno de estos síntomas, puede ser un indicio de fatiga visual. Debe dejar de ver la televisión en 3D hasta que desaparezca el efecto secundario.

A pesar de que las pruebas subjetivas proveen la mejor aproximación a lo que es la calidad percibida por el usuario, su implementación no siempre es viable debido a aspectos como: tiempo requerido para la aplicación de las pruebas, número de usuarios requerido, ambiente controlado y requerimientos de equipamiento especifico. Por esta razón la ITU-T se ha dado a la labor de definir la recomendación P.1203, que tiene como objetivo permitir la evaluación subjetiva de forma automática de la QoE, sin la participación de usuarios.

2.4.4. Implementación ITU-T P.1203

La ITU-T P.1203 [37] describe un modelo para la evaluación subjetiva de la calidad percibida por los usuarios de un servicio de transmisión de contenido audiovisual. El estándar ITU-T P.1203 ha sido desarrollado específicamente para los sistemas de *streaming* tipo HAS (*HTTP Adaptive Streaming*).

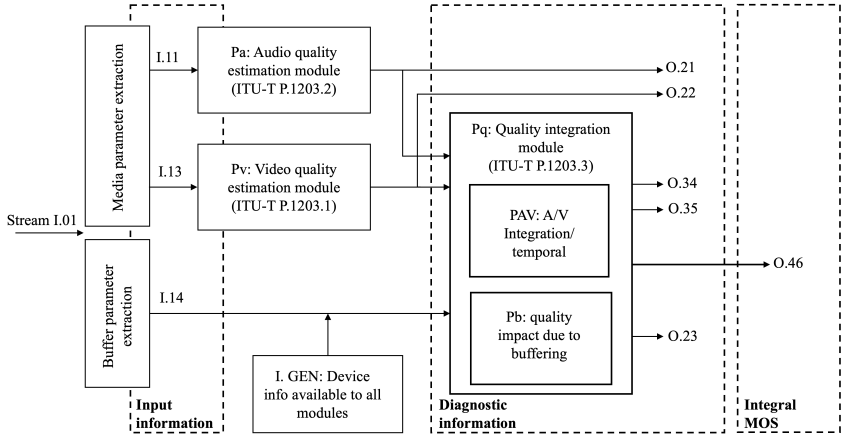


Figura 2.7. Diagrama de los diferentes módulos y componentes definidos en la ITU-P.1203.
Fuente: ITU-T P.1203 [37]

Como se puede observar en la Figura 2.7., el modelo P.1203 consta en realidad de tres módulos individuales: uno para la estimación de la calidad de vídeo (P.1203.1, Pv), otro para la estimación de la calidad de audio (P.1203.2, Pa) y uno más para la integración audiovisual (P.1203.3, Pq). La entrada I.01 denota el flujo de bits, del que se deriva la información de entrada específica para el audio (I.11), el vídeo (I.13) y el retardo de carga inicial y las paradas (I.14).

El modelo Pv (P.1203.1) puede funcionar en cuatro modos de operación según la información de entrada disponible, desde el modo 0 al modo 3. Estos modos se distinguen según la cantidad de información de entrada disponible I.13.

Los modos definidos son:

• Modo 0: utiliza información de los metadatos (MPD DASH) como el códec, la resolución y la tasa de bits de las diferentes representaciones, así como el retardo inicial, duración del segmento y la información sobre las paradas producidas durante la reproducción.

• Modo 1: hace uso de la información utilizada en el Modo 0 con la adición de información de los flujos multimedia obtenida de la inspección de la cabecera de los paquetes.

• Modo 2: hace uso de la información utilizada en el Modo 1 más un 2% de los bytes del total de la transmisión que es utilizada para un análisis parcial.

• Modo 3: hace uso de la información utilizada en el Modo 2 y de la información de los flujos multimedia basada en el análisis del total del *bitstream.*

Se puede encontrar una descripción detallada del algoritmo del modelo P.1203.1 en [38]. P.1203 ofrece diferentes informaciones de salida que pueden utilizarse para el diagnóstico del servicio y proporciona una puntuación en una escala de 1 a 5. Además, P.1203 tiene en cuenta la información de las paradas para la estimación de la calidad y depende de los *codecs* utilizados en la codificación del contenido. La implementación disponible en github actualmente (*ITU-T Rec. P.1203 Standalone Implementation*), puede utilizarse para vídeos codificados en formato H.264 con una resolución full HD (1920x1080) y duración de entre 30 s y 5 min.

Capítulo 3

Comparación de codificadores de vídeo 3D

Un factor clave en el almacenamiento y la transmisión de contenido 3D, dado el impacto que puede tener en la calidad de vídeo recibida y percibida por el usuario, es el formato de representación y la tecnología de compresión que se utiliza. En la selección de un formato de distribución deben considerarse varios factores. Estos factores incluyen la capacidad de almacenamiento o el ancho de banda disponibles, la compatibilidad con los sistemas ya existentes, las capacidades de los reproductores y receptores, la calidad mínima aceptable, y la capacidad de proveer servicios futuros.

Con el objetivo de analizar el impacto que la codificación del vídeo 3D tiene en la QoE del usuario, como parte de esta tesis, se ha realizado un estudio comparativo entre los principales estándares disponibles para la representación y codificación de contenido 3D, en el contexto de un escenario de distribución adaptativa de vídeo 3D a través de Internet. Se han considerado los estándares de codificación de vídeo H.264/MPEG-4 AVC (*Advanced Video Coding*), H.265/HEVCH.265/MPEG-HEVC (*High-Efficiency Video Coding*) y sus correspondientes extensiones para vídeo multivista o 3D estereoscópico (H.264/MVC y MV-HEVC). En el contexto de la transmisión en un entorno *simulcast*, se evalúa también la eficacia de la utilización de las codificaciones asimétricas para la transmisión de vídeo 3D, como una alternativa para la reducción del ancho de banda manteniendo la calidad global. La comparación se ha realizado empleando métricas bien conocidas de evaluación objetiva de calidad de vídeo, como el PSNR, el SSIM y el VMAF, y cuya definición fue presentada en el capítulo anterior. Así mismo, también se han realizado pruebas de evaluación subjetiva con usuarios, siguiendo las pautas indicadas en la recomendación ITU-T P.915[33], relativa a los métodos de evaluación subjetiva de la calidad del vídeo 3D descrita en la sección 2.4.3.a del capítulo anterior.

Se ha evaluado la eficiencia de codificación en términos de calidad del vídeo, la calidad de la profundidad y el malestar visual percibido por los usuarios, tomando como referencia los valores de MOS (*Mean Opinion Score*) y las curvas RD (*Rate Distortion*) que ofrecen una relación entre la calidad del vídeo y la tasa de bits. Así mismo, se han considerado aspectos como el tiempo de codificación y el consumo de recursos hardware, que dada la tendencia cada vez mayor del consumo de contenidos multimedia en dispositivos móviles resultan ser factores relevantes.

El capítulo se distribuye como sigue. La sección 3.1 se describen los formatos de representación de vídeo 3D. La sección 3.2 presenta brevemente los aspectos más relevantes de los principales estándares de codificación incluidos en este estudio. En la sección 3.3 se presenta la metodología y resultados obtenidos de la comparación de los estándares de codificación. Finalmente, se presenta una sección de conclusiones, que busca establecer una correlación entre los resultados de la evaluación subjetiva y las métricas de evaluación objetiva empleadas.

3.1. Representación del vídeo 3D

El vídeo 3D se basa en la estéreopsis [12] (o disparidad retiniana). Se trata de una señal binocular posible debido al desplazamiento horizontal de los dos ojos y que da lugar a dos imágenes ligeramente diferentes proyectadas en cada una de las retinas. Esta disparidad binocular permite al cerebro humano estimar la profundidad del objeto por triangulación y hace que el usuario tenga una experiencia más convincente de la profundidad dentro de una escena. Para recrear la experiencia de visión estereoscópica, en la que cada ojo ve una imagen diferente, los sistemas de visualización de vídeo 3D se valen de la multiplexación temporal. El vídeo presentado alterna entre un *frame* para ser visto por el ojo izquierdo y un *frame* ser visto por el ojo derecho, y así sucesivamente. Estos formatos suelen denominarse "*frame secuencial*" [4]. La tasa de *frames* de cada vista puede reducirse para que la cantidad de datos sea equivalente a la de una sola vista. Sin embargo, antes de todo este proceso, es el formato de la señal 3D el que determina como se transmite cada *frame* del vídeo desde la fuente (servidor de vídeo) al dispositivo de visualización del cliente. Por esta razón, como paso intermedio para su codificación y posterior transmisión, las secuencias capturadas deben ser convertidas de un formato de producción a un formato de transporte.

Como se muestra en la Figura 3.1, los formatos de producción de vídeo 3D se pueden dividir en dos clases principales, según la tecnología empleada para su adquisición: formatos de solo vídeo y formatos de vídeo más profundidad [39]. Los formatos de sólo vídeo más populares son: el vídeo estéreo clásico (CSV) que consiste en un par de secuencias, correspondientes a las vistas del ojo izquierdo y del ojo derecho, obtenidas por la grabación clásica mediante cámaras estereoscópicas y que no requiere procesamiento adicional, pero que provoca la mayor tasa de bits en la codificación de vídeo 3D. Los formatos de vídeo estéreo de resolución mixta (MRS) donde se submuestrea una vista con el objetivo de reducir la tasa de bits en el proceso de compresión. La percepción general es cercana a la vista nítida, ya que, según la teoría de supresión binocular, el cerebro humano puede enmascarar algunos artefactos borrosos en una vista con la otra. Y el vídeo multivista (MVV) con más de 2 vistas de la misma escena, obtenidas desde diferentes ángulos. Por otra parte, tenemos los formatos basados en un vídeo monoscópico y una secuencia de mapa de profundidad por píxel asociada, este tipo de formatos ofrece menores tasas de bits en la compresión, ya que los datos de profundidad pueden comprimirse con mayor eficacia que los datos de vídeo. La segunda vista se sintetiza en el lado de la pantalla mediante el *renderizado* basado en imágenes de profundidad. Los formatos con profundidad son: vídeo más profundidad (V+D) y vídeo multivista más profundidad (MVD).

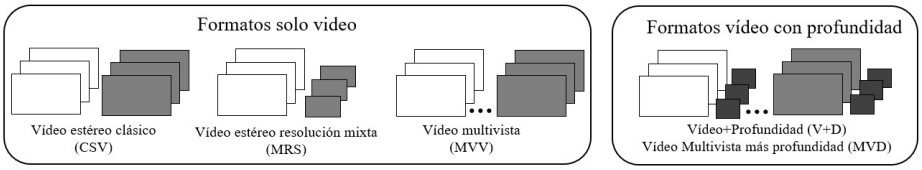


Figura 3.1. Formatos de video 3D

Si bien los formatos con profundidad tienen mayor adaptabilidad con independencia de la resolución de la pantalla, la ubicación del usuario o su distancia a la misma en comparación con los formatos de sólo vídeo, en lo que respecta al post-procesamiento requieren algoritmos complejos, como la rectificación y la estimación de la profundidad. Además, a pesar de las importantes mejoras en este campo, las configuraciones de cámaras mejoradas con sensores de profundidad siguen siendo poco utilizadas debido a la limitada resolución espacial y rango de profundidad de los sensores. Por este motivo, en lo que resta de este trabajo se tendrán en cuenta únicamente los formatos de solo vídeo, sin profundidad.

Los formatos de solo vídeo en 3D constan de al menos dos secuencias de vídeo. Por este motivo, al pensar en su transmisión y almacenamiento se platean dos cuestiones fundamentales: ¿cómo *representarlos*? Y ¿cuál es el mejor formato de *compresión*? Cualquier decisión debe tomarse teniendo en cuenta el ancho de banda disponible o la capacidad de almacenamiento, así como la calidad y las expectativas del usuario.

3.1.1. Formatos de representación de vídeo 3D

En sus inicios, la distribución de vídeo 3D estaba soportada por las redes de difusión de vídeo digital DVB (*Digital Video Broadcasting*), estando sujeta su reproducción a la capacidad de los terminales para decodificar el vídeo 3D. Por tal motivo, buscando ser compatibles con los sistemas de distribución de vídeo 2D ya existentes y evitando tener que realizar actualizaciones masivas de infraestructuras de transmisión por parte del proveedor y hardware de recepción para el usuario, los sistemas de distribución de vídeo 3D adoptaron mayoritariamente los formatos de vídeo estereoscópico-3D (S3D) compatibles con frame (FC, *frame-compatible*) y los de resolución completa (FR, *full- resolution frame- compatible*) [4][5][40].

En los formatos FC, la generación del vídeo se basa en el submuestreo y la multiplexación de las imágenes de las vistas izquierda y derecha en una sola imagen o secuencia de imágenes, con el objetivo de mantener la misma tasa y tamaño de frame que la distribución de vídeo 2D [6]. Con los formatos FC, la multiplexación puede realizarse espacial o temporalmente. Con la multiplexación espacial, las imágenes de la vista izquierda y derecha se submuestrean primero y luego se combinan en una sola imagen. Existen diferentes alternativas para realizar el submuestreo de cada vista, siendo los esquemas *side-by-side* y *top-and-botton* los más utilizados, ya que son compatibles con la gran mayoría de sistemas de visualización disponibles en el mercado. En el formato *side-by-side* se aplica submuestreo horizontal para las vistas izquierda y derecha, lo que reduce la resolución horizontal en un 50%. Los frames submuestreados son entonces puestos uno al lado del otro para formar una única secuencia de imágenes, como se observa en la Figura 3.2a.

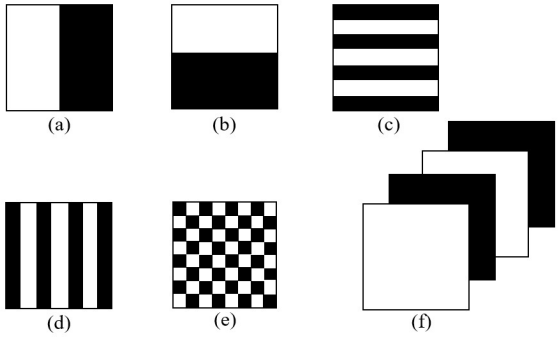


Figura 3.2. Formatos *Frame-compatible* de representación de vídeo3D. (a) *Side-by-side*. (b)*Top-and-botton*. (c) Entrelazado Horizontal. (d) Entrelazado Vertical. (e) *Checkerboard*. (f)Multiplexado Temporal

Así mismo, en la Figura 3.2b. muestra como en el formato *top-and-botton* se submuestrean verticalmente la vista izquierda y la derecha y se pone una debajo de la otra. En el formato de entrelazado o líneas intercaladas, las vistas izquierda y derecha están de nuevo submuestreadas, pero se ponen juntas de forma intercalada horizontalmente, Figura 3.2c, o verticalmente, Figura 3.2d. En el formato *checkerboard* que se muestra en la Figura 3.2e, las vistas izquierda y derecha son submuestreadas empleando un patrón de rejilla y multiplexadas en un único *frame* con el diseño de un tablero de ajedrez, y se requieren gafas activas para ver el 3D en este formato. Finalmente, la Figura 3.2f presenta un formato basado en la multiplexación temporal o *frame* sequential, donde los fotogramas de las dos vistas son presentados manteniendo su resolución espacial, pero de forma intercalada, y la velocidad de fotogramas de cada vista debe reducirse para que la cantidad de datos sea equivalente al de una sola vista. Como puede verse, el principal inconveniente de los formatos *frame-compatible*, es la reducción de la resolución espacial o temporal debido al uso de técnicas de diezmando y de interpolación.

Por su parte, en los formatos denominados de resolución completa (*full-resolution frame-compatible*), se toman los vídeos con resolución Full HD, normalmente 1920x1080 o 1280x720, para todas las vistas y no se aplica ningún tipo de reducción de resolución espacial o temporal, lo que significa que los datos brutos generados son directamente proporcionales al número de vistas, el doble en el caso del vídeo estereoscópico-3D (S3D). Es por eso por lo que, han aparecido los formatos multivista basados en las nuevas técnicas de codificación MVC (*Multiview Video Coding*) [41] y MVD (*Multiview Video Plus Depth*) [42], que, si bien tienen problemas de compatibilidad con los sistemas ya existentes, mejoran significativamente las tasas de compresión en función de la calidad y permiten el envío de múltiples vistas, aumentando el ángulo de visión de los usuarios en sistemas basados en pantallas auto-estereoscópicas, entre otros casos.

3.2. Estándares de codificación de vídeo 3D

Desde principios de los años 90, el desarrollo de normas de codificación de vídeo ha estado impulsado por dos espacios de aplicación paralelos: la comunicación de vídeo en tiempo real y la distribución o difusión de contenidos de vídeo. Dada la gran cantidad de modelos de negocio y aplicaciones disponibles en el mercado, existe una creciente demanda global de una mayor eficiencia en la transmisión de archivos multimedia, y en especial en hacer frente al problema de

que ningún códec se adapta a todos los casos de uso. El termino códec es la contracción de *enCOde/DECode,* y hace referencia a que los códecs funcionan codificando el vídeo para su distribución o almacenamiento y descomprimiéndolo para su reproducción o uso.

En cuanto al origen de los códec, por una parte, encontramos las especificaciones basadas en estándares que han sido publicadas por las organizaciones de normalización: la Unión Internacional de Telecomunicaciones (UIT) y el Grupo de Expertos en Imágenes en Movimiento (MPEG)[43]. Aquí tenemos los códecs H.264/AVC, H.265/HEVC y H.266/VVC que tienen dos nombres porque fueron creados conjuntamente por MPEG, que los llamó AVC (*Advanced Video Coding*), HEVC (*High Efficiency Video Coding*) y VVC (V*ersatile Video Coding*), y la UIT (H.264, H.265, H.266). Y, por otra parte, tenemos especificaciones con filosofía de código abierto como (VP8 y VP9) promovidas por empresas como Google o consorcios formados por proveedores de vídeo bajo demanda, productores de contenidos de vídeo, empresas de desarrollo de software y proveedores de navegadores web como es el caso de la AOMedia (*Alliance for Open Media*), responsable del desarrollo de nuevos formatos de codificación diseñados para la transmisión de vídeo sobre Internet principalmente, como AV1 (*AOMedia Video 1*).

Si bien la evolución de los codificadores de vídeo empieza en 1984 con la publicación de la primera norma de codificación de vídeo digital con el nombre de ITU-T H.120 [44], no es sino hasta 1990 con la publicación del estándar para vídeo conferencias H.261 [45] que se introducen importantes mejoras en el proceso de codificación con la inclusión del esquema de codificación híbrido. El esquema de codificación híbrido utiliza la predicción *intra-frame* junto con la compensación de movimiento para eliminar la información redundante en el plano temporal y la codificación de la señal de error resultante en el dominio de la transformada, mediante el uso comúnmente de la DCT (*Discrete Cosine Transform*).

Como evolución del estándar MPEG-1 (ISO/IEC 11172-2) publicado en 1993, aparece el estándar MPEG-2 que se publicó en 1994 como ISO/IEC 13818-2 [46]. MPEG-2 es el primer estándar que incluye una ampliación para permitir la codificación de vídeo multivista (estéreo) y fue adoptado por la UIT como H.262 [47], constituyéndose en la primera norma de codificación de vídeo conjunta de los dos organismos de normalización.

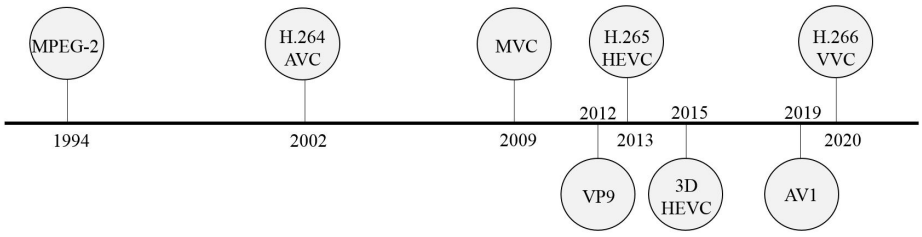


Figura 3.3. Línea de tiempo de las especificaciones de codificación de vídeo más populares

Como se puede observar en la Figura 3.3, debido al aumento de la demanda de contenidos multimedia por parte de los usuarios, en las últimas dos décadas se ha evidenciado un incremento en el desarrollo de nuevos codificadores de vídeo. Se resalta la aparición de los populares codificadores estándar: H.264/MPEG-4 (H.264/AVC) [48] y HEVC/H.265 [49] que, al igual que sus antecesores, pueden ser empleados para la codificación de vídeo 3D bajo un esquema simulcast,

donde las vistas izquierda y derecha se codifican por separado como si se tratase de dos vídeos 2D. La ventaja de este esquema de codificación simulcast es que permite utilizar la mayor parte de la infraestructura existente para vídeo 2D. Sin embargo, como las imágenes de las diferentes vistas están muy correlacionadas, mucha información entre las dos vistas es redundante. Por eso, en el caso particular del vídeo 3D, para una compresión eficaz, además de la predicción temporal, se debe aprovechar la redundancia inter-vista. Es por eso por lo que, buscando una mayor eficacia en la codificación, a la publicación de cada uno de estos estándares le ha seguido la publicación de extensiones o versiones específicas para la codificación de vídeo estéreo o vídeo multivista, conocidas como H.264/MVC [41] y MV-HEVC 3D [50][51], donde se introduce la predicción entre-vistas y que se describen con mayor detalle más adelante.

La compresión de vídeo es un área de competición abierta y muchos desarrolladores de codificadores están trabajando de forma paralela a los equipos de la ITU-T y ISO/IEC. Es por esto por lo que en la línea temporal de la Figura 3.3, podemos encontrar formatos de codificación de vídeo de código abierto y libres de derechos como VP8 [52] y VP9 [53] desarrollados por Google en 2010 y 2013 respectivamente como base del proyecto WebM [54]. VP9 proporciona una reducción de 50% del *bitrate* respecto a VP8 y con el objetivo de igualar o superar el rendimiento de H.265/HEVC. Incorpora un tamaño de macrobloque que llega hasta 64x64 con posibilidad de un granulado específico de bloques de 4x4 para el modelo temporal. Soporta 10 modos de predicción *intra-frame* y cuatro modos de predicción *inter-frame*. Utiliza la Transformada Discreta del Coseno (DCT), la Transformada Asimétrica Discreta del Seno (ADST) y la Transformada de Walsh-Hadamard para el modelo espacial. Se ha diseñado un filtro para eliminar los defectos de los bordes de los macrobloques. Soporta HDR (*High-Dinamic-Range*) y permite la codificación sin pérdidas. Al igual que H.265 permite el procesamiento en paralelo, además de escalabilidad temporal y espacial. Tres años más tarde Google lanzó VP10 alcanzando mejoras de hasta 40% en cuanto a compresión respecto a VP9 con el coste de incrementar el tiempo de codificación.

Continuando con la línea temporal encontramos el codificador AV1 [55] desarrollado como su nombre indica por la AOMedia como sucesor de VP9. Se terminó en 2018 y su principal propósito fue obtener una compresión mayor respecto a sus antecesores además de proporcionar escalabilidad con dispositivos modernos y diferentes enlaces de datos con una complejidad de decodificación muy baja. AV1 ofrece una reducción del 30% respeto al *bitrate* medio obtenido con VP9.

Finalmente, como resultado del trabajo del JVT (*Join Video Experts Team*), encontramos el estándar VVC [56] también conocido como H.266, ISO/IEC 23090-3, y MPEG-1 parte 3 y que fue finalizado en Julio de 2020. H.266/VVC ofrece un 50% más de compresión que H.265/HEVC. VVC permite incrementar el tamaño de bloque en el modelo temporal a 128x128 con bloques con particionados binarios o ternarios respecto al particionado cuaternario de HEVC. Permite particionados diferentes para los planos de luminancia y crominancia y habilita la aceleración hardware mediante el procesamiento en paralelo. En cuanto a la predicción *intra-frame,* se utilizan 67 modos de predicción en lugar de los 33 utilizados en HEVC, además de habilitar el uso de bloques rectangulares. En cuanto a la predicción *inter-frame,* permite la predicción a partir de dos imágenes de referencia, además de incrementar de 2 a 3 dimensiones los grados de libertad de los vectores de movimiento. Para la transformación se utilizan bloques no rectangulares realizando transformadas de diferentes tipos en función del modo de predicción. Se incrementa

el valor de QP o *quantizer* máximo de 51 a 63 para permitir tasas de bit menores. En la codificación entrópica se sigue utilizando CABAC (*Context Adaptative Binary Arithmetic Coding*).

Hay que tener en cuenta que para los nuevos y más eficientes codecs como H.266/VVC, H.265/HEVC, VP9 y AV1, los fabricantes de navegadores y dispositivos están divididos en su soporte. Por ejemplo: el navegador Safari de Apple es compatible con H.265/HEVC, pero no con VP9. Google y Firefox tienen soporte para VP9 y AV1, pero no para H.265/HEVC. Dado que al hablar de códecs la capacidad más relevante es la reproductibilidad y universalidad, para el desarrollo de esta tesis, hemos trabajado con los codificadores estándar H.264/AVC y H.265/HEVC, incluyendo sus versiones para vídeo estereoscópico-3D, tomando como criterio su amplio soporte de dispositivos.

3.2.1. H.264/AVC y H.264/MVC

H.264/AVC (*Advanced Video Coding*) [57], conocido formalmente como Recomendación UIT-T H.264 e ISO/IEC 14496-10 (MPEG-4 Parte 10), es una norma que define un códec de vídeo de alta compresión, desarrollada conjuntamente por el VCEG (*ITU-T Video Coding Experts Group*) y el MPEG (*ISO/IEC Moving Picture Experts Group*) entre diciembre de 2001 y mayo de 2003 en el marco de lo que se conoce como el JVT (*Joint Video Team*). La intención del proyecto H.264/AVC era crear un estándar capaz de proporcionar una buena calidad de imagen con tasas de bits notablemente inferiores a los estándares previos (MPEG-2, H.263 o MPEG-4 parte 2), sin incrementar la complejidad de su diseño.

H.264/AVC incluye funcionalidades como el uso de transformadas DCT, *quantizer* (QP), estimación y compensación de movimiento con predicciones inter e *intra-frame* y codificación entrópica. Algunas de las mejoras que aporta respecto a codificadores anteriores es la predicción de tramas de tipo Intra, DCT de enteros de tamaño 4x4, uso de múltiples *frames* de referencia, macrobloques de tamaño variable, precisión de un cuarto de píxel para la compensación de movimiento y el uso de un filtro de *deblocking* con el objetivo de suavizar los bordes de los macrobloques. Mantiene tanto la estructura de la transformada discreta del coseno como la compensación de movimiento de las versiones anteriores, y su eficiencia en la reducción de tasa de bits es significativa. Todas estas mejoras hacen que el codificador aporte una mejora del 50% en la reducción del *bitrate* para un vídeo de la misma calidad. La resolución soportada por H.264 llega a 4K (4096x2160) con 60 fps. Para más información sobre H.264/AVC, se sugiere consultar el estándar [58], cuya última versión a la fecha es del (06/2019), o referencias como [59] y [57] que presentan una descripción detallada del mismo. A continuación, se resumen algunas de las principales características de H.264/AVC, que luego serán usadas entre otros por la extensión del estándar para la codificación de vídeo multivista conocida como H.264/MVC (*Multiview Video Coding*) [41]:

- Como complemento a las imágenes I (*Intra*), P (*Predicted*), B (*Bi-directional predicted*), en el perfil extendido del estándar H.264/AVC se incluyen las imágenes SP (*Switching P*) y la SI (*Switching I*), que sirven para codificar la transición entre dos flujos de vídeo. Permiten, sin enviar imágenes Intra muy costosas en tiempos de procesamiento, pasar de un vídeo a otro utilizando predicción temporal o espacial, con la ventaja de que permiten la reconstrucción de valores específicos exactos de la muestra, aunque se

utilicen imágenes de referencia diferentes o un número diferente de imágenes de referencia en el proceso de predicción.

- H.264 funciona procesando los fotogramas de vídeo mediante un estándar de compresión de vídeo orientado a bloques y basado en la compensación del movimiento. Las unidades básicas de procesamiento se denominan macrobloques. Cada macrobloque, puede ser descompuesto en subbloques de forma cuadrada o rectangular en el rango 4x4 hasta 16x16, lo que permite una segmentación más precisa de las regiones en movimiento.
- Para la predicción de movimiento, se emplea el método de predicción de vectores de movimiento (MVP). Se incluye la capacidad de utilizar múltiples vectores de movimiento por macrobloque (uno o dos por partición) con un máximo de 32 en el caso de un macrobloque B construido con 16 particiones de 4×4.
- Precisión de un cuarto de píxel para la compensación de movimiento, lo que permite una descripción precisa de los desplazamientos de las zonas en movimiento.
- Cambios en la DCT: tamaño 4x4 píxeles (8x8 en los perfiles FRExt), coeficiente entero, precisión infinita y eficiencia en su implementación para la que se requieren solo sumas y desplazamientos binarios.
- Se aumenta el rango dinámico del parámetro de cuantización QP de 0 a 51.
- Para reducir los artefactos producidos por la codificación basada en bloques, H.264/AVC especifica un filtro antibloques (*Deblocking filter*) que opera dentro del bucle de compensación de movimiento entre imágenes.
- H.264/AVC soporta dos métodos de codificación entrópica: CAVLC (*Context Adaptative Variable Length Coding*) y CABAC (*Context Adaptative Binary Arithmetic Coding*) que están basados en VLC (*Variable Length Coding)* de forma adaptativa.
- La estructura de predicción incluye el uso de tramas B jerárquicas, que representa una mejora en la eficiencia de compresión comparada con la codificación típica IBBP usada por H.262/MPEG-2[47].
- H.264/AVC soporta diferentes opciones de muestreo de crominancia, incluido el monocromático (4:0:0, 4:2:0, 4.2:2, 4:4:4).
- Según sus capacidades, se definen varios perfiles destinados a clases específicas de aplicaciones. Los perfiles se declaran mediante un código de perfil que permiten al decodificador reconocer los requisitos para decodificar un determinado flujo de bits
- El diseño de H.264/AVC incluye una VCL (*Video Coding Layer*) y una NAL (*Network Abstraction Layer*). Mientras la VCL crea una representación codificada del contenido fuente, la NAL formatea esos datos y proporciona información de cabecera de una manera que permite una personalización sencilla y eficaz del uso de los datos de la VCL para una amplia variedad de sistemas. Así, el contenido codificado estará organizado en unidades NAL, algunas de ellas denominadas unidades VCL NAL que contienen la información de las imágenes codificadas, y otras denominadas unidades non-VCL NAL, que están destinadas a información relacionada con la definición de parámetros (*parameter set*) y mensajes con información de mejora suplementaria
- SEI (*Supplemental Enhancement Information*). La información de mejora suplementaria y VUI (*Video Usability Information*), que son información adicional que puede insertarse en el flujo de bits con diversos fines, como indicar el espacio de color utilizado o diversas restricciones que se aplican a la codificación. Por ejemplo: los formatos FC pueden

funcionar sin problemas en las infraestructuras existentes y con los decodificadores de vídeo ya instalados. En un esfuerzo por facilitar y fomentar su adopción, el estándar H.264/MPEG-4 AVC [58] introdujo un nuevo mensaje de información de mejora suplementaria (SEI), que permite señalar la disposición de empaquetamiento de tramas utilizada. Dentro de este mensaje SEI se puede señalar no solo el formato de empaquetado de fotogramas, sino también otra información como la relación de muestreo entre las dos vistas y el orden de vista, entre otros. Al detectar este mensaje SEI, un decodificador puede reconocer inmediatamente el formato y realizar el procesamiento adecuado, como el escalado, la eliminación de ruido o la conversión del formato de color, de acuerdo con el formato FC especificado. Además, esta información puede utilizarse para informar automáticamente a un dispositivo posterior, por ejemplo, una pantalla o un receptor, del formato FC utilizado, señalizando adecuadamente este formato a través de interfaces compatibles, como la interfaz multimedia de alta definición (HDMI).

De julio de 2006 a noviembre de 2009, el JVT trabajó en la codificación de vídeo multivista MVC [41], especificado en el anexo H de H.264/AVC. Se trata de una ampliación de H.264/AVC hacia la televisión en 3D y la televisión de punto de vista libre de alcance limitado, que permite la construcción de flujos de bits que representan más de una vista en una escena de vídeo. Ese trabajo incluyó el desarrollo de dos nuevos perfiles de la norma: el *Multiview High Profile,* que admite un número arbitrario de vistas, y el *Stereo High Profile,* que está diseñado específicamente para el vídeo estereoscópico de dos vistas. A continuación, se resumen algunas de las características clave del H.264/MVC, y puede consultarse más información en [60] .

H.264/MVC proporciona una representación compacta para múltiples vistas de una escena de vídeo, equivalente a flujos multicámara sincronizados. El vídeo estereoscópico o a dos vistas ($V_{Izquierda}$ y $V_{Derecha}$) para la visualización 3D es un caso especial de vídeo multivista. Para la compresión de vídeo estereoscópico H.264/MVC aprovecha el hecho de que las vistas izquierda y derecha muestran la misma escena desde una perspectiva ligeramente diferente correspondiente a la distancia de separación entre los ojos. Estas imágenes son en general muy similares, por esto, como complemento a la predicción temporal y espacial, el estándar H.264/MVC incluye la predicción entre-vistas, incrementando la eficiencia de la compresión. En la Figura 3.4 se muestra un ejemplo de la estructura de predicción, donde se observa cómo una de las imágenes (Vista Izquierda) puede comprimirse sin referencia a la otra. Y la segunda imagen (Vista Derecha) puede predecirse a partir de la imagen ya codificada, al igual que en las imágenes relacionadas temporalmente se puede aplicar la compensación de movimiento para la compresión de vídeo. La combinación de la predicción temporal y la predicción entre-vistas es el principio básico de la compresión eficiente del vídeo estereoscópico convencional.

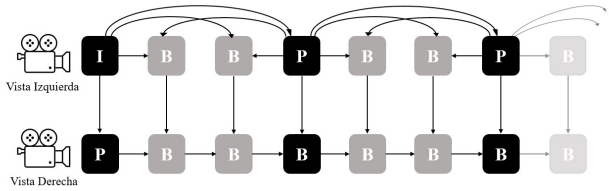


Figura 3.4. Predicción entre-vistas en MVC

Para mejorar su desempeño H.264/MVC aprovecha el uso de *frames* B jerárquicos para explotar las redundancias temporales y entre-vistas, al mismo tiempo que permite la compatibilidad con los sistemas de reproducción de vídeo ya existentes, mediante la inclusión de una "vista base" en la estructura del flujo MVC codificado. En reconocimiento a su capacidad de codificación de alta calidad y a su compatibilidad con versiones anteriores, la Asociación de *Discos Blu-Ray* seleccionó el perfil estéreo de la extensión H.264/MVC como formato de codificación para el vídeo 3D con resolución de alta definición.

3.2.2. H.265/HEVC y H.265/HEVC 3D

H.265/HEVC (*High Efficiency Video Coding*) [49] ha sido desarrollado por el JCT-VC (*Joint Collaborative Team on Video Coding*), que es un grupo de expertos en codificación de vídeo creado por el VCEG del ITU-T y el MPEG de la ISO/IEC en 2010, como respuesta a la demanda de mejorar el rendimiento de compresión en resoluciones de vídeo de alta definición (HD) y de ultra alta definición (UHD). En junio de 2013, H.265/HEVC fue formalmente publicada como Recomendación ITU-T H.265 [61] y, en noviembre de 2013, publicada por ISO/IEC como ISO/IEC 23008-2 (MPEG-H parte 2) [62], cuya última revisión (versión 7) fue aprobada en noviembre de 2019. Estudios previos como el realizado en [63] han demostrado que en un escenario de vídeo 2D, HEVC es capaz de lograr la misma calidad de vídeo subjetiva que el H.264/MPEG-4 AVC *High Profile* mientras que requiere en media sólo un 50% de la tasa de bits. Así mismo, según los resultados obtenidos en [64], H.265/HEVC tiene un 27,3% de ahorro de *bitrate* respecto a VP9 según las medidas objetivas realizadas en el estudio.

Al igual que su predecesor, H.265/HEVC es un esquema híbrido de compresión de vídeo en el que la primera imagen utiliza únicamente predicción *intra-frame* mientras que todas las demás utilizan predicción *intra* e *inter-frame*. HEVC ha sido diseñado para aprovechar resoluciones muy altas como 4K y 8K, que suelen contener áreas muy amplias de bloques similares. Por lo tanto, una de las primeras cosas que se cambiaron fue el tamaño de bloque con el que opera el códec. El concepto de macrobloque utilizado en H.264 se convierte en un CTU (*Coding Tree Unit*). Además, H.265 permite procesado en paralelo con lo que se acelera el proceso de codificación. H.265 soporta resoluciones hasta 8K UHDTV (8192x4320) de hasta 300 fps. A continuación, se mencionan algunas de las principales características a modo de comparación con su predecesor.

- Cada CTU puede ser de tamaño 16x16, 32x32 o 64x64. A su vez los CTU se dividen en un árbol cuaternario QT (*Quad-Tree*) que permite particiones más pequeñas llegando a CUs (*Coding Units*), mejorando el subparticionado de la imagen en estructuras de tamaño variable, esto permite aprovechar grandes áreas homogéneas aumentando la

eficiencia de codificación. Cada CU se puede predecir tanto *inter-frame* como *intra-frame*. La imagen residual se codifica utilizando transformaciones de bloque.

- La especificación H.265/HEVC distingue entre tres tipos de imágenes: imágenes que se utilizan como referencia a corto plazo (*short-term reference*), imágenes que son usadas como referencia a largo plazo (l*ong-term reference*) e imágenes que no se utilizan como referencia. Así, las imágenes que serán usadas para la predicción de una imagen actual o futura dentro de la secuencia de vídeo codificada se recogen en el conjunto de imágenes de referencia (RPS, *Reference Picture Set*).
- Un filtro antibloques (*Deblocking Filter*) similar al utilizado en H.264/AVC funciona dentro del bucle de predicción entre imágenes. Sin embargo, el diseño se ha simplificado en lo que respecta a los procesos de toma de decisiones y filtrado, y se ha hecho más fácil el procesamiento en paralelo. Se incluye además un offset de la muestra adaptativa (SAO, *Simple-Adaptative Offset*), dentro del bucle de compensación de movimiento.
- H.265/HEVC especifica 35 modos direccionales para la predicción *intra-frame* en comparación con los 8 modos especificados por H.264/MPEG-4 AVC [49].
- En cuanto a la codificación entrópica H.265/HEVC solo admite el algoritmo CABAC que es el más eficiente y fundamentalmente similar a CABAC en H.264/AVC [39].

La Tabla 3.1 presenta un resumen comparativo de los principales aspectos entre el estándar H.264/AVC y H.265/HEVC.

Tabla 3.1. Comparativa H.264/AVC –H.265/HEVC

Parámetro	*H.264/AVC*	*H.265/HEVC*
Codificación entrópica	CABAC/ CAVLC	CABAC
Unidad de procesamiento de datos	MB (*Macro Blocks*)	CTU (*Coding Tree Units*)
Tamaño máximo de bloque	16x16	64x64
Formatos soportados	MP4, MOV, F4V, 3GP, TS	TS, MP4, 3GP, MKV
Direcciones de predicción Intra	*9*	35
Predicción de Movimiento	MVP *(Motion Vector Prediction)*	AMVP *(Advanced Motion Vector Prediction)*

La necesidad de dar soporte a los servicios basados en múltiples vistas y 3D, motivó la creación en julio de 2012 del *Joint Collaborative Team on 3D Video Coding Extension Development* (JCT-3V), con el mandato de trabajar en extensiones de codificación de vídeo multivista y 3D para H.264/AVC y H.265/HEVC entre otros estándares de codificación de vídeo. El JCT-3V ha desarrollado dos extensiones para HEVC [50], *Multiview* HEVC (MV-HEVC) y 3D-HEVC [65][66], que están integradas en la última versión del estándar [61]. Aunque 3D-HEVC mantiene la eficiencia de codificación en función del ahorro en tasa de bits de H.265/HEVC, el principal inconveniente es la dificultad para la adquisición de contenidos basados en el formato textura más mapa de profundidad. Por su parte, MV-HEVC sigue el mismo principio de diseño que la codificación de vídeo multivista (MVC), de la extensión multivista de H.264/MPEG-4 AVC, al mismo tiempo que mantiene la compatibilidad con el vídeo monoscópico codificado con HEVC.

3.3. Metodología para la comparación de codificadores de vídeo 3D

Considerando que la codificación de vídeo corresponde a una compresión con pérdidas, en un sistema de distribución de vídeo, la secuencia de vídeo recibida por el usuario está degradada y sólo se aproxima a la secuencia de vídeo de entrada original. Es por ello que el reto de los sistemas de codificación es conseguir la máxima reducción de *bitrate*, minimizando la degradación de la imagen de entrada, y conocer el impacto de la codificación en la QoE de los usuarios. Dicho impacto en la QoE resulta clave para valorar el desempeño de los diferentes estándares de codificación.

En el caso particular del vídeo 3D o vídeo estereoscópico, podemos hablar de que existen dos métodos de codificación: la *codificación simulcast o simultánea*, donde cada vista se codifica por separado y puede aplicarse a todos los formatos, aunque requiere la mayor tasa de bits, y la *codificación estéreo conjunta* (*Joint Stereo Coding*), donde como su nombre indica se realiza la codificación conjunta de ambas vistas. Este esquema de codificación, si bien es más eficiente que la codificación *simulcast*, ya que también se aprovechan las dependencias entre las vistas, sólo puede aplicarse a los formatos de sólo vídeo sin profundidad.

En la literatura se pueden encontrar trabajos como los presentados en [63][64] y [67], donde la capacidad de compresión de varias generaciones de estándares de codificación, incluidos los estándares H.264/AVC y H.265/HEVC, se compara mediante la utilización de métricas objetivas como el PSNR y los resultados de pruebas subjetivas. Si bien estos estudios han sido realizados en el contexto del vídeo 2D, sus resultados sirven de referencia para entornos de vídeo 3D basados en codificación *simulcast*. Así mismo, cabe destacar los estudios de validación realizados durante el desarrollo de los propios estándares de codificación, como HEVC [68], MVC [41] y MV-HEVC [69], donde se validan las degradaciones propias de la codificación.

En relación al estudio de las degradaciones de codificación o la calidad de experiencia en el contexto del vídeo 3D, encontramos estudios como el presentado en [70], donde se ha analizado y comparado el rendimiento de los esquemas de codificación H.264/AVC y H.264/MVC, empleando para la evaluación subjetiva el método de clasificación por categorías absolutas (ACR) [71]; o el presentado en [23] donde se estudia la evaluación subjetiva de la calidad de vídeo 3D empleando el estándar H.264/MVC en un sistema de 3DTV. Finalmente, cabe mencionar algunas investigaciones donde se estudia el rendimiento subjetivo de la codificación asimétrica de vídeo 3D, utilizando diferentes parámetros de cuantificación o resoluciones (espaciales o temporales) como los presentados en [72][73].

La Figura 3.5 muestra el esquema general que se ha seguido en este capítulo para el estudio comparativo de la calidad de vídeo 3D. En total se generan 3 grupos de secuencias codificadas para cada uno de los estándares bajo estudio (H.264 y H.265).

- Codificación estéreo conjunta (CEC): corresponde a las secuencias codificadas empleando la versión multivista de los codificadores estándar H.264 y H.265 (H.264/MVC y MV-HEVC).
- Codificación *simulcast* simétrica (CSS): aquí tenemos las secuencias codificadas empleando el esquema de codificación *simulcast* simétrico. Bajo este esquema, cada una de las vistas del par estereoscópico se codifica de forma independiente empleando

las versiones estándar compatibles para vídeo 2D de los codificadores H.264 y H.265 y usando exactamente los mismos parámetros de configuración.

- Codificación *simulcast* asimétrica (CSA): este último grupo corresponde a las secuencias codificadas empleando un esquema de codificación *simulcast* asimétrico. Al igual que en el caso anterior, se emplean las versiones estándar de los codificadores H.264 y H.265, pero en este caso, una de las vistas (vista derecha V_1) se codifica empleando un mayor factor de compresión con relación a la vista izquierda V_0. Con estas secuencias se busca validar el fenómeno perceptivo de supresión binocular [74], que hace referencia a que la calidad global percibida por el usuario se aproxima a la de la vista de mejor calidad, permitiendo reducir el ancho de banda mientras se mantiene la calidad percibida.

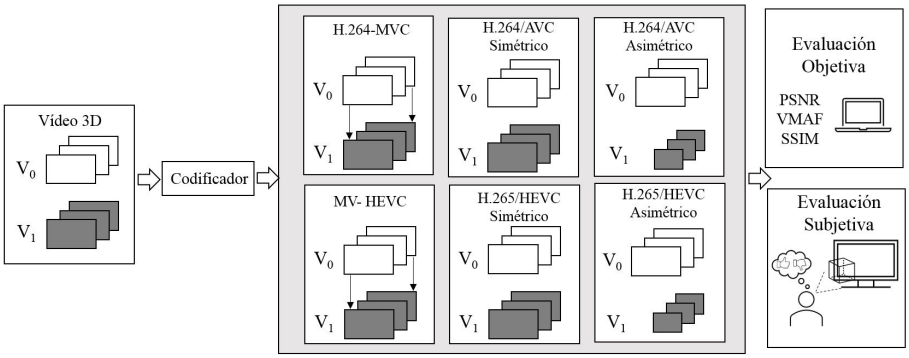


Figura 3.5. Esquema general comparación de codificadores

3.3.1. Selección de secuencias de prueba

Al igual que con distribución de contenido multimedia basado en vídeo 2D, la diversidad del contenido es una cuestión importante a la hora de evaluar el rendimiento del proceso de codificación. Las secuencias utilizadas en este trabajo han sido tomadas de las bases de datos de vídeos 3D HD estereoscópico: Nantes-Madrid-3D-Stereoscopic-V1 NAMA3DS [75] y RMIT3DV [76]. Dichas bases de datos están compuestas por secuencias estereoscópicas (vistas por separado) con resolución 1920x1080 a una tasa de 25 imágenes por segundo, diseñadas para representar una amplia gama de contenidos y condiciones visuales. Asimismo, se ha empleado la popular secuencia de animación generada por ordenador, Big Buck Bunny en 3D producida por Blender Foundation [77] con resolución 1920x1080 y una tasa de 30 imágenes por segundo por cada vista.

Con el fin de cuantificar y evaluar la variedad de los contenidos, tal y como lo describen las normas internacionales, como la recomendación ITU-T P.910 [71], se realizó una caracterización de las secuencias. En particular, se analizó la complejidad espacial y temporal mediante el cálculo del factor de información espacial (SI, *Spatial Information*) y el factor de información temporal (TI, *Temporal Information*) respectivamente. Siguiendo la recomendación ITU-T P.910, los factores SI y TI se calculan sobre la componente de luminancia de la vista izquierda. El factor SI mide la

actividad espacial de una escena, mediante el cálculo de la desviación estándar, sobre los píxeles de cada fotograma previamente procesado con filtros de Sobel. Por su parte, el factor TI se basa en la característica de diferencia de movimiento, que es la diferencia entre los valores de los píxeles en la misma ubicación en el espacio, pero en fotogramas adyacentes.

Se evaluaron inicialmente un total de 13 secuencias (SRC). La Figura 3.6. muestra los índices SI y TI calculados de cada una de las secuencias. Adicionalmente la Tabla 3.2 presenta la información general relativa a cada una (SRC01- SRC013), incluidos los índices SI y TI calculados en todos los casos para 10 s de vídeo.

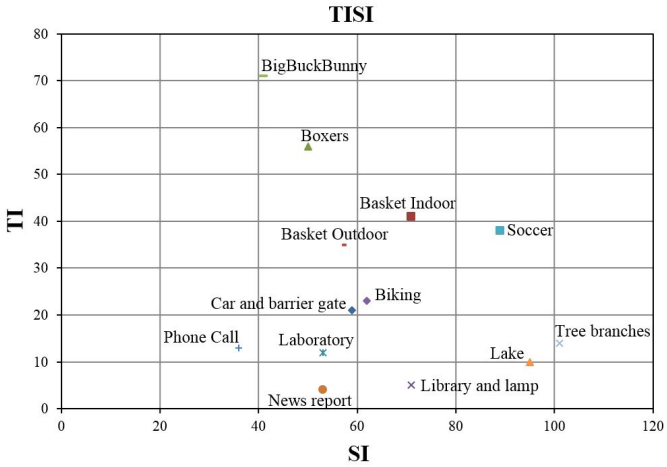


Figura 3.6. Gráfica índices TI-SI

Tabla 3.2. Resumen de características para la selección de las secuencias de vídeo

	Secuencia	*Resolución*	*Tasa de frames*	*Duración*	*SI*	*TI*
SRC01	*Car and barrier gate*	1920x1080	25 fps	16 s	59	21
SRC02	*Basket Indoor*	1920x1080	25 fps	16 s	71	41
SRC03	*Basket Outdoor*	1920x1080	25 fps	25 s	57	35
SRC04	*Library and lamp*	1920x1080	25 fps	10 s	71	5
SRC05	*Laboratory*	1920x1080	25 fps	16 s	53	12
SRC06	*News Report*	1920x1080	25 fps	16 s	53	4
SRC07	*Tree leaves*	1920x1080	25 fps	16 s	101	4
SRC08	*Boxers*	1920x1080	25 fps	16 s	50	56
SRC09	*Phone Call*	1920x1080	25 fps	16 s	36	13
SRC010	*Biking*	1920x1080	25 fps	77 s	62	23
SRC011	*Soccer*	1920x1080	25 fps	16 s	89	38
SRC012	*Lake*	1920x1080	25 fps	104 s	95	10
SRC013	*BigBuckBunny*	1920x1080	30 fps	634 s	41	71

A partir de sus valores SI-TI y teniendo en cuenta aspectos como el tipo de iluminación y las características del contenido, se han elegido un total de 4 secuencias. Una de tipo exterior (*Car and barrier gate*, SRC01), una de deportes (*Basket Indoor*, SRC02), una de interior *(Library and lamp*, SRC04) y otra de animación (*BigBuckBunny*, SRC013), tal como se muestra en la Figura 3.7. La Tabla 3.3 presenta un resumen de las principales características de cada una de las secuencias seleccionadas.

Tabla 3.3. Características principales de las secuencias de prueba seleccionadas

	Secuencia	*Iluminación*	*Movimiento*	*Características, textura*
SRC01	*Car and barrier gate*	Natural día	Movimiento Medio/Alto	Escena exterior. Cielo de fondo, coche en movimiento y barrera, árboles, construcciones, varios planos.
SRC02	*Basket Indoor*	Artificial/Interior	Movimiento Alto	Escena de deporte en interior. jugadores corriendo, pabellón deportivo, carteles publicitarios.
SRC04	*Library and lamp*	Artificial/Interior	Movimiento Bajo	Escena de interior. Primer plano lámpara movimiento pendular, estanterías y mesas, personas en segundo plano.
SRC013	*BigBuckBunny*	Digital/Generada por ordenador	Movimiento Medio	Escena de animación, personajes en movimiento con paisaje de fondo, arboles, cielo y césped, cambios de escena

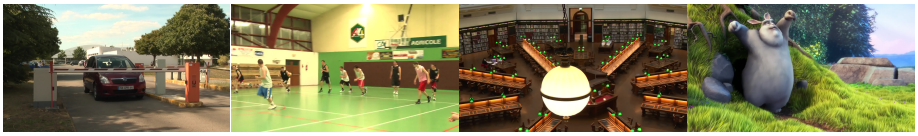


Figura 3.7. Secuencias seleccionadas. SRC01 (frame 116), SRC02 (frame 167), SRC04 (frame 117), SRC013 (frame 1623)

Dados los tiempos de codificación de algunas de las herramientas que serán empleadas, para esta primera fase se trabajara con 10 s de cada una de las cuatro secuencias seleccionadas. Como paso previo a la codificación, las secuencias son convertidas a formato YUV 420p, que es el formato sin compresión de entrada por defecto soportado por los codificadores de referencia que se van a emplear.

3.3.2. Selección de codificadores y parámetros de configuración

Como se muestra en la Figura 3.8, para la comparación de los estándares de codificación H.264/AVC y H.265/HEVC en el ámbito de la codificación de vídeo 3D, se han empleado los codificadores de referencia JMVC 8.5 (*Joint Multiview Video Coding*) [78] y el HM 16.7 (*HEVC test Model*) desarrollados por el JCTVC (*Joint Collaborative Team on Video Coding*) para las codificaciones empleando el método de codificación estéreo conjunta (CEC) o multivista. Por otra parte, se han empleado los codificadores JM 19.0 (*Joint Test Model*) H.264/AVC [79] y HM 16.18 (*HEVC Test Model*) H.265/HEVC [80], junto con las implementaciones de los estándares H.264 y H.265 en el codificador de código abierto FFMPEG (x264 y x265), para la codificación de las secuencias estereoscópicas, empleando el método *simulcast*. Recordemos que, procurando la compatibilidad con los sistemas de vídeo 2D, el concepto de codificación *simulcast* hace referencia a la codificación por separado de cada una de las vistas del par estereoscópico.

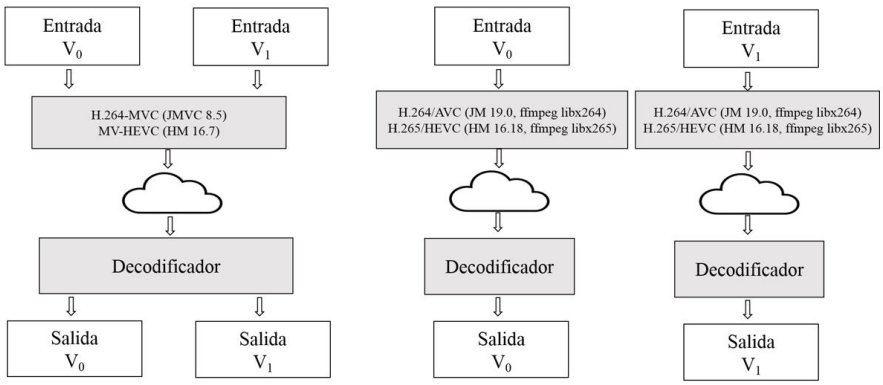


Figura 3.8. Implementaciones seleccionadas de los codificadores (MVC vs *Simulcast*)

Como se observa en la Figura 3.8, en la codificación estéreo conjunta (CEC), los procesos de codificación y decodificación se aplican a las vistas izquierda (V_0) y derecha (V_1) de forma simultánea, lo que conlleva un incremento del coste computacional para su ejecución. Como se mencionó anteriormente, una de las características de la CEC es que incluye la predicción inter-vistas para mejorar la capacidad de compresión. Esta predicción entre vistas puede estar limitada de forma general a los denominados *frames* de anclaje (*frames* tipo I o P), o extenderse a todos los *frames* de las secuencias de vídeo. En un trabajo previo a la tesis y dentro de la línea de investigación, se pudo comprobar la mejora en la compresión que se obtiene al incluir un mayor número de *frames* en el proceso de predicción inter-vistas.

En cuanto a la configuración de los codificadores, en este capítulo se normalizará en función de la calidad y, como consecuencia, se ajustará el tamaño del archivo. Por tal motivo, para generar las variaciones de calidad, se empleó el parámetro de cuantificación QP con los valores 24, 28, 32, 36 y 40. Así mismo, al tratarse de una comparación, se procuró que los parámetros de configuración fuesen equivalentes en todos los codificadores, eligiendo los mismos valores GoP (*Group of Pictures*) y separación entre *frames* Intra. En cuanto al perfil, se ha elegido el perfil *Main* que ofrece un mejor compromiso entre la complejidad de la decodificación, los requisitos exigidos en

los dispositivos finales y la calidad de vídeo obtenida, y soporta *frames* I, P, B. La Tabla 3.4 resume los principales parámetros de configuración empleados.

Tabla 3.4. Información general de configuración de los codificadores

	H.264/AVC	*H.265/HEVC*	*H.264/AVC - MVC*	*H.265/HEVC- MVC*
Implementación	JM 19.0	HM 16.18	JMVC 8.5	HM 16.7
QP	24, 28, 32, 36, 40	24, 28, 32, 36, 40	24, 28, 32, 36, 40	24, 28, 32, 36, 40
Tamaño GoP	8	8	8	8
Perfil	Main	Main	Main	Main
IntraPeriod	24	24	24	24
InputChromaFormat	4:2:0	4:2:0	4:2:0	4:2:0
SymbolMode	CAVLC	CABAC	CAVLC	CABAC

3.3.2.a. Presets FFMPEG

Como se comentó anteriormente, podemos encontrar diferentes implementaciones de los estándares de codificación H.264/AVC y H.265/HEVC. Unas de las más populares por su condición de código abierto son las disponibles en el software FFMPEG, definidas como libx264 y libx265 respectivamente. Existen algunos parámetros definidos por los estándares que son comunes para todos los codificadores, como es el caso de los algoritmos de codificación entrópica (CAVLC o CABAC) o los perfiles usados para equilibrar la complejidad de la decodificación (*Baseline, Main o High)*. Sin embargo, en el caso particular del software FFMPEG, encontramos otros ajustes que pueden influir directamente tanto en la calidad como en el tiempo de codificación, y son conocidos como *presets.*

Es importante señalar que no se pretende hacer un estudio exhaustivo de los *presets* y todos los parámetros implícitos en cada uno de ellos. Por lo tanto, teniendo en cuenta que, la comparación de los codificadores propuesta en este trabajo se basa en la codificación de calidad constante mediante la definición del parámetro QP y con el objetivo de conocer el comportamiento de los diferentes *presets* disponibles en FFMPEG (*Ultrafast, Superfast, Veryfast, Faster, Fast, Medium, Slow, Slower, Veryslow*), para la configuración de los codificadores x264 y x265, se han realizado pruebas empleando las 4 secuencias seleccionadas en la sección anterior y usando los valores de QP (24, 28, 32, 36 y 40).

De acuerdo con la documentación ofrecida por FFMPEG, un *preset* o preajuste ofrece un conjunto de opciones, que proporcionan una determinada relación entre la velocidad de codificación y la calidad de la compresión. Por tanto, si tenemos como objetivo un determinado tamaño de archivo o una tasa de bits constante, de forma general se presupone que se logrará una mejor calidad con un *preset* más lento. Así mismo, en un entorno de codificación de calidad constante como el nuestro, se entiende que la menor tasa de bits se obtendrá también eligiendo el *preset* más lento.

Sin embargo, aparecen algunas discrepancias según las pruebas realizadas, y tal como se observa en las Tablas 3.5 y 3.6, que presentan la calidad promedio obtenida empleando la métrica VMAF, para las implementaciones x264 y x265 respectivamente. Cada una de las filas de las Tablas 3.5 y 3.6 presenta la calidad obtenida empleando los 9 *presets* disponibles, desde el más rápido (*Ultrafast)* hasta el más lento (*Veryslow*). Los valores corresponden al promedio para cada valor de QP de los cuatro vídeos codificados. La escala de colores representa para cada fila la transición entre el *preset* de mejor calidad representado en color verde hasta el de peor calidad representado en color rojo.

Podemos ver que, si bien en la implementación FFMPEG x265 (Tabla 3.6) los resultados obtenidos se corresponden con el hecho de que los *presets* más lentos sean los que ofrece la mejor calidad, en el caso de la implementación FFMPEG x264 (Tabla 3.5) el comportamiento de los *presets* varía en función del QP empleado. En particular, llaman la atención los resultados obtenidos al emplear un valor de QP igual a 24, donde la variación entre la calidad ofrecida por cada uno de los *presets* es muy baja, dándose el caso de que el *preset* que ofrece la mejor calidad es a su vez el más rápido.

Tabla 3.5. Calidad promedio FFMPEG x264 (VMAF)

QP\Preset	*Ultrafast*	*Superfast*	*Veryfast*	*Faster*	*Fast*	*Medium*	*Slow*	*Slower*	*Veryslow*	*Delta*
QP24	94,94	92,27	92,08	93,67	94,00	94,17	94,29	94,35	94,18	3,10%
QP28	91,75	87,23	89,66	91,42	91,89	92,06	92,30	92,39	92,19	5,91%
QP32	85,56	78,92	83,07	85,89	86,52	86,70	87,16	87,35	87,11	10,69%
QP36	76,04	67,06	73,46	77,26	78,05	78,14	78,66	79,10	78,56	17,95%
QP40	62,95	52,00	60,41	65,01	65,79	65,76	66,30	66,97	66,40	28,80%
Promedio	82,25	75,49	79,74	82,65	83,25	83,36	83,74	84,03	83,69	11,31%
% Respecto al mejor	2,12%	10,16%	5,11%	1,65%	0,93%	0,80%	0,34%	0,00%	0,41%	

Tabla 3.6. Calidad promedio FFMPEG x265 (VMAF)

QP/Preset	*Ultrafast*	*Superfast*	*Veryfast*	*Faster*	*Fast*	*Medium*	*Slow*	*Slower*	*VerySlow*	*Delta*
QP24	94,89	95,45	95,22	95,21	95,31	95,39	96,26	96,36	96,37	1,56%
QP28	91,96	92,73	92,51	92,51	92,69	92,73	93,94	94,08	94,09	2,32%
QP32	87,21	88,52	88,17	88,16	88,52	88,49	90,14	90,38	90,40	3,66%
QP36	79,93	82,00	81,50	81,54	82,24	82,04	84,16	84,60	84,63	5,88%
QP40	69,60	72,06	71,66	71,70	72,90	72,35	74,77	75,51	75,54	8,52%
Promedio	84,72	86,15	85,81	85,82	86,33	86,20	87,85	88,19	88,21	4,12%
% Respecto al mejor	3,95%	2,33%	2,72%	2,70%	2,12%	2,27%	0,40%	0,02%	0,00%	

Así mismo, la última columna muestra el delta entre la mejor y la peor calidad por cada valor de QP. De aquí observamos que cuando el valor del QP es bajo, es decir, la calidad es buena, el impacto del *preset* es menor. Por ejemplo, según la Tabla 3.5, para un valor de QP igual a 24, encontramos que el delta entre la calidad más alta (*preset Ultrafast*) y la más baja (*preset Veryfast)*, es de 3.10%, frente al delta de 28,8% obtenido al emplear un QP igual a 40 (en este caso el delta representa la distancia entre la calidad ofrecida por los *preset Slower* y *Superfast* que

ofrecen la mejor y peor calidad respectivamente). Este comportamiento se observa también en el caso de FFMPEG x265 (Tabla 3.6), aunque los rangos de variación son menores tal y como se presenta de manera gráfica en las Figuras 3.9a. y 3.9b. Las Figuras 3.9a. y 3.9b representan el valor máximo, el valor mínimo y el valor promedio de la calidad (VMAF) en función del QP para las implementaciones libx264 y libx265 respectivamente. Se ve claramente como el rango de variación de la calidad ofrecida por los diferentes *presets* se incrementa a medida que se reduce la calidad del vídeo, o lo que es equivalente, en la medida que el valor del parámetro de codificación QP aumenta.

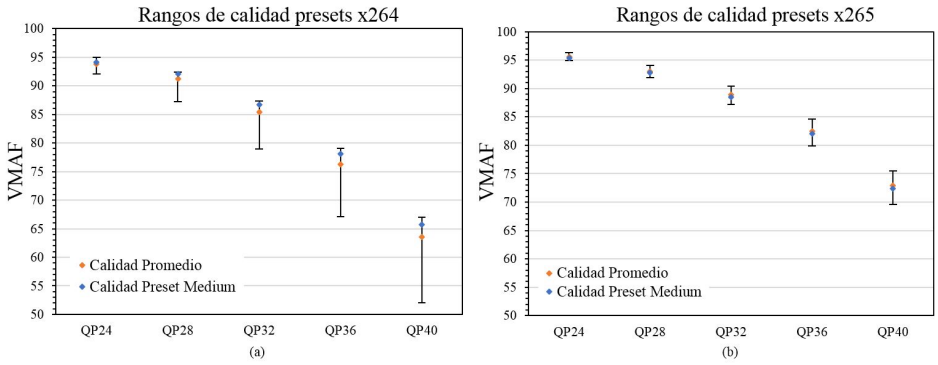


Figura 3.9. Rangos de variación de la calidad entre *presets*, para las implementaciones FFMPEGx264 (a) y x265 (b) en función del parámetro QP.

Sin embargo, la calidad no es el único aspecto por considerar al evaluar diferencias entre los distintos *presets*. Para tener una visión más general que permita identificar qué *presets* tienen un mejor desempeño, las Figuras 3.10a, 3.10b. y 3.10c. representan su respuesta en términos de la calidad (VMAF), tiempo de codificación y tasa de bits respectivamente, tanto para x264 como para x265. Es importante tener en cuenta que el tiempo de codificación dependerá del dispositivo empleado para realizar las codificaciones. En nuestro caso, las pruebas se realizaron empleando una estación Intel Core I7 9700 CPU 3 GHz.

Si observamos el comportamiento reflejado en estas gráficas y retomando la información contenida en las Tablas 3.5 y 3.6, al hacer una elección entre el *preset* de mejor calidad y el *preset Medium*, para la implementación x265 vemos como *Veryslow* (*preset* mejor calidad) incrementa por un factor superior a 15 el tiempo de codificación y ofrece solo un 2,27% de mejora en la calidad. Mientras, para la implementación x264, *Slower* (*preset* de mejor calidad) incrementa en un factor cercano al doble el tiempo de codificación, mientras ofrece una mejora de 0,8%. Viendo estas cifras hay que preguntarse si estas mejoras de calidad valen la pena teniendo en cuenta la reducción del rendimiento que conllevan. Por ejemplo, para el caso de x264, ¿vale la pena reducir el rendimiento en un 50% para aumenta la calidad en menos de 1 punto? Y en nuestro caso la decisión ha sido que no y se ha elegido el *preset* Medium para las configuraciones basadas en FFMPEG, considerando que ofrece un buen equilibrio entre todos los factores valorados.

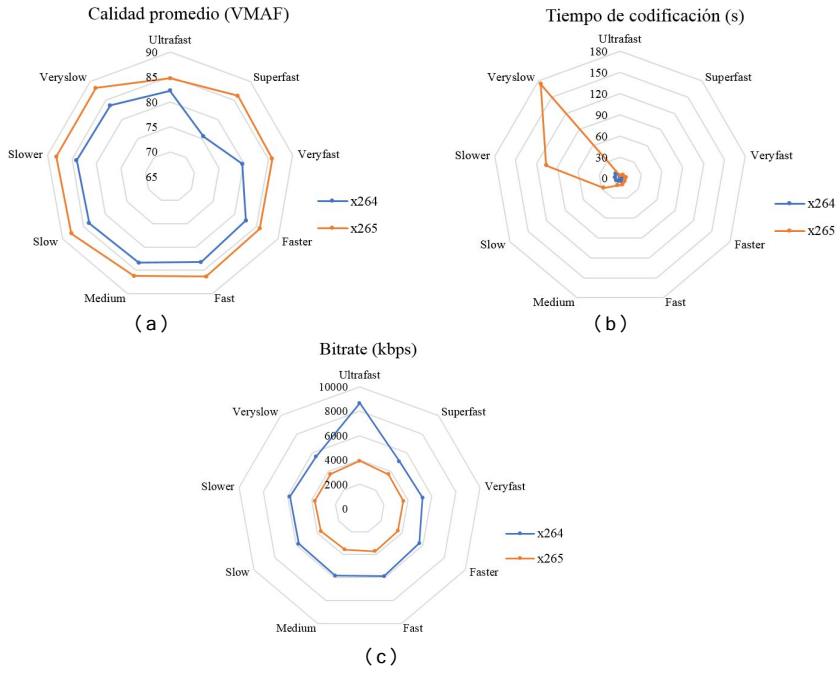


Figura 3.10. Presets en x264 y x265 en función de: (a) Calidad VMAF, (b) Tiempo de codificación, (c) *Bitrate*

3.3.3. Comparación de codificadores mediante métricas objetivas

Con base en los conceptos presentados en el Capítulo 2 respecto a la evaluación de la QoE de vídeo 3D, en esta sección se presenta la metodología y los resultados obtenidos en la evaluación de la QoE tanto mediante métricas objetivas (que tratan de estimar automáticamente la calidad que percibirían los usuarios finales) como mediante pruebas de evaluación subjetiva (basadas en la valoración emitida por los usuarios del sistema). En esta etapa, se han seleccionado las métricas PSNR y SSIM, que corresponden a la categoría FR (*Full-Reference*), es decir, que requieren el vídeo original para su cálculo, así como la métrica VMAF, que, según sus promotores mejora su capacidad de predecir la calidad a partir de una mejor aproximación al modelado del sistema de visión humano y la simulación de circuitos neuronales de bajo nivel que permiten reunir información sobre cómo el cerebro humano percibe la calidad.

3.3.3.a. Codificación simétrica

En esta primera parte, se muestran los resultados de la comparación de las diferentes implementaciones de los codificadores H.264 y H.265 bajo un esquema de codificación simétrica, es decir, codificando las dos vistas con el mismo factor de compresión, mediante implementaciones multivista o *simulcast* tanto de los codificadores de referencia como con la implementación FFMPEG. La Figuras 3.11., 3.12 y 3.13 muestran las curvas RD (*Rate Distortion*) en función del PSNR, SSIM y VMAF respectivamente, para las secuencias bajo estudio.

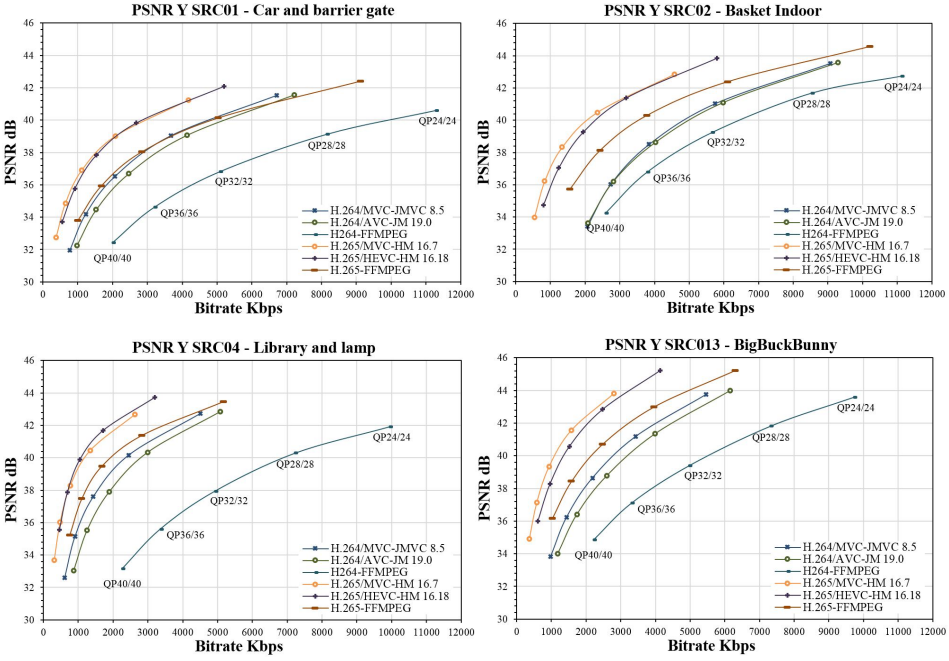


Figura 3.11. Curva RD (PSNR) comparación de codificadores

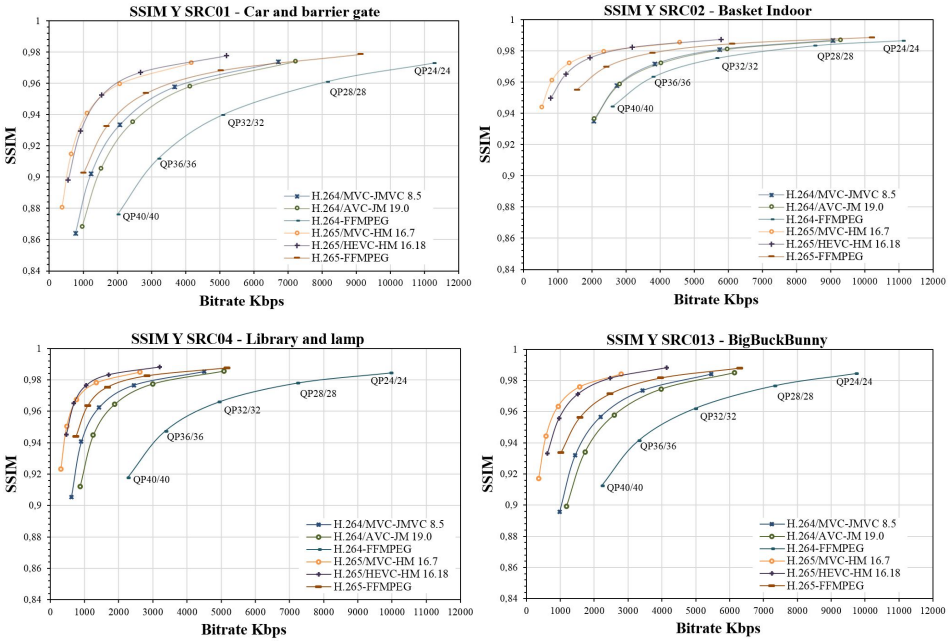


Figura 3.12. Curva RD (SSIM) comparación de codificadores

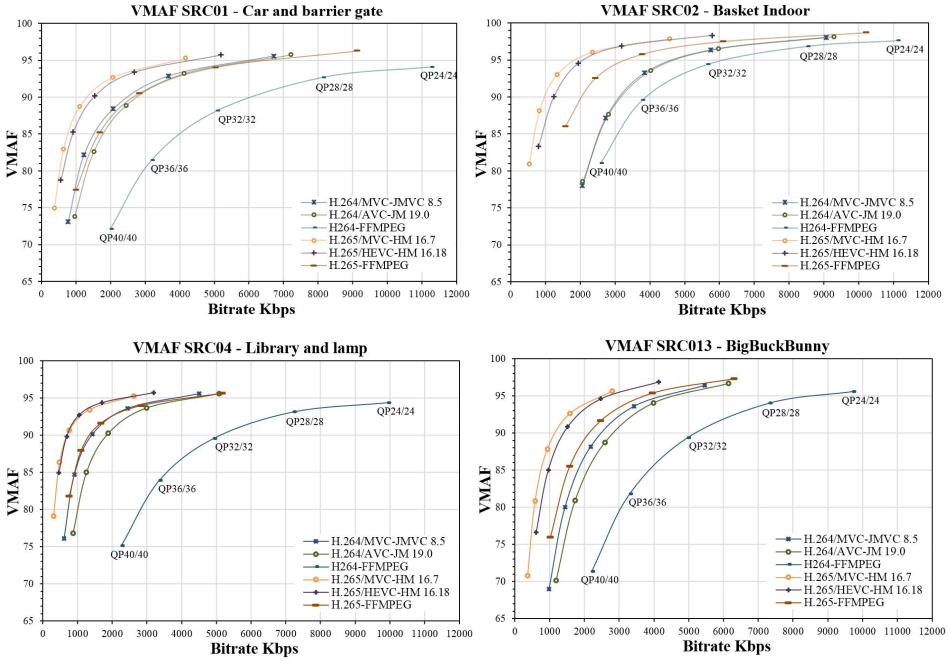


Figura 3.13. Curva RD (VMAF) comparación de codificadores

De acuerdo con los aportes de trabajos como [63][64][81], y según los resultados presentados en las curvas RD-PSNR (Figura 3.11), RD-SSIM (Figura 3.12) y RD-VMAF (Figura 3.13), los codificadores multivista tienen un mejor rendimiento en relación con los esquemas *simulcast* o vistas por separado. Esto se debe a que además de considerar las similitudes entre *frames* dentro de cada vista, explotan también la similitud intervista, lo que permite eliminar los *frames* de tipo Intra al codificar una de las vistas en función de la otra. Así mismo, se demuestra que, para una misma calidad de vídeo, los codificadores basados en el estándar H.265 proporcionan un ahorro significativo en el *bitrate* con respecto a los basados en el estándar H.264.

En base a las 4 secuencias bajo estudio, tal como se muestra en la Tabla 3.7., las implementaciones basadas en H.265/HEVC ofrecen en promedio una ganancia superior al 50% que el de sus equivalentes basados en H.264. En concreto, para un valor de VMAF=90, la implementación multivista del codificador H.265/MVC-HM 16.7 ofrece una ganancia de *bitrate* de 54,95% respecto a la implementación multivista del codificador H.264/MVC-JMVC 8.5, al mismo tiempo que las implementaciones H.265/HEVC-HM 16.18 y H.265 FFMPEG ofrecen una ganancia del 54,84% y el 59,25% respecto a sus equivalentes H.264/AVC-JM 19.0 y H.264 FFMPEG, respectivamente.

Tabla 3.7. Ganancia promedio de *bitrate* para una misma calidad VMAF=90

	Ganancia de *bitrate* respecto a				
Codificador	H.264 FFMPEG	H.264/AVC-JM 19.0	H.264/MVC-JMVC 8.5	H.265 FFMPEG	H.265/HEVC-HM 16.18
H.265/MVC-HM 16.7	79,36%	60,78%	54,95%	49,06%	12,97%
H.265/HEVC-HM 16.18	75,73%	54,84%	48,45%	41,14%	-
H.265 FFMPEG	59,25%	23,09%	11,80%	-	-
H.264/MVC-JMVC 8.5	50,91%	11,98%	-	-	-
H.264/AVC-JM 19.0	45,59%	-	-	-	-

Sin embargo, enfocados en nuestro siguiente objetivo que se centra en la transmisión de contenidos de vídeo 3D en un entorno adaptativo, además de la eficiencia de codificación en términos de las métricas objetivas, se ha valorado la opción de poder generar codificaciones asimétricas y los tiempos de codificación. La Tabla 3.8. muestra el incremento en los tiempos de codificación de los codificadores basados en H.265 respecto a los de H.264, así como la amplia diferencia porcentual entre los tiempos de codificación usando codificadores de referencia. Por tal motivo, en adelante se usará como herramienta para la codificación, el software FFMPEG con las librerías de codificación libx264 y libx265.

Tabla 3.8. Tiempos de codificación – Comparación de codificadores respecto a H.264 FFMPEG

Codificador	Tiempo respecto H.264 FFMPEG
H.265/MVC-HM 16.7	330,74%
H.265/HEVC-HM 16.18	293,01%
H.264/MVC-JMVC 8.5	221,95%
H.264/AVC-JM 19.0	215,84%
H.265 FFMPEG	25,15%

Tabla 3.9. Reducción de la tasa de bits de las codificaciones H.265 FFMPEG respecto a H.264 FFMPEG en función del QP y el tipo de secuencia

Secuencia	QP24	QP28	QP32	QP36	QP40
SRC01	19,04%	38,15%	44,13%	47,32%	49,76%
SRC02	8,06%	28,25%	33,12%	35,74%	39,43%
SRC04	47,92%	60,72%	65,89%	66,15%	66,58%
SRC013	35,08%	45,90%	50,25%	52,09%	53,59%

Por otra parte, la Tabla 3.9 presenta los resultados de reducción del *bitrate* en función del QP, de H.265 FFMPEG respecto a H.264 FFMPEG, para cada una de las 4 secuencias bajo estudio. En todos los casos, podemos observar como la eficiencia en términos de la reducción del *bitrate*, de las codificaciones hechas con H.265 FFMPEG respecto a las realizadas con H.264 FFMPEG se incrementa, a medida que la calidad del vídeo se reduce al aumentar el QP. Así mismo, vemos como para un mismo valor de QP, la ganancia de *bitrate* de las codificaciones hechas con H.265 FFMPEG respecto a las hechas con H.264 FFMPEG es menor en las secuencias con un nivel de movimiento alto/medio (SRC01 - *Car and barrier gate*, SRC02 - *Basket indoor*) y se incrementa en las secuencias con nivel de movimiento bajo (SRC04 - *Library and lamp*), confirmándose de esta

manera que tanto el grado de movimiento de la secuencia de vídeo, como el parámetro de cuantificación inciden directamente en la eficiencia de codificación de un estándar respecto al otro.

3.3.3.b. Codificación asimétrica

Ante la necesidad de lograr una mayor reducción del *bitrate* para la codificación del vídeo estéreo 3D, además, de los métodos de codificación multivista que, si bien se ha comprobado ofrecen una significativa reducción del *bitrate*, resultan poco eficientes en cuanto al tiempo de codificación y el consumo de recursos, en esta sección se estudiará el comportamiento de la codificación asimétrica, que aprovecha la capacidad del HVS para percibir las altas frecuencias, incluso si esta información está presente en sólo una de las vistas.

La codificación asimétrica pretende explotar la supresión binocular del HVS consiguiendo una compresión de vídeo más eficiente al representar una de las dos vistas con una calidad inferior sin que esto afecte la percepción global del usuario [82]. Esto es similar a lo que se ha hecho para el vídeo monocular en color, donde los canales de crominancia se codifican con menos bits que los de luminancia, porque el HVS es menos perceptivo a los cambios de color. En el caso del vídeo estéreo 3D, la asimetría puede lograrse tanto mediante la reducción de la resolución espacial [22] como mediante la reducción de la calidad de una de las vistas. En este trabajo, considerando que como parámetro para la configuración y comparación de los codificadores H.264 y H.265 se ha seleccionado el parámetro de cuantificación, la asimetría será analizada en función de la diferencia de calidad.

Para el análisis de la codificación asimétrica se emplearán las implementaciones FFMPEG x264 y FFMPEG x265 en un contexto de codificación *simulcast*, y las vistas izquierda (V0) y derecha (V1) serán codificadas mediante la combinación de los parámetros de codificación QP (24, 28, 32, 36 y 40). De esta forma, la nomenclatura QP24/28 corresponde a una representación del vídeo estéreo, en la que la vista izquierda V0 ha sido codificada con un QP=24, mientras que la vista derecha V1 ha sido codificada empleando un valor de QP=28. Así mismo, atendiendo al hecho de que, entre todas las métricas objetivas disponibles, VMAF es la que mejor se aproxima a lo que podría ser la representación perceptual de la calidad del vídeo, en adelante VMAF será la métrica empleada para realizar la comparación objetiva de los resultados de codificación.

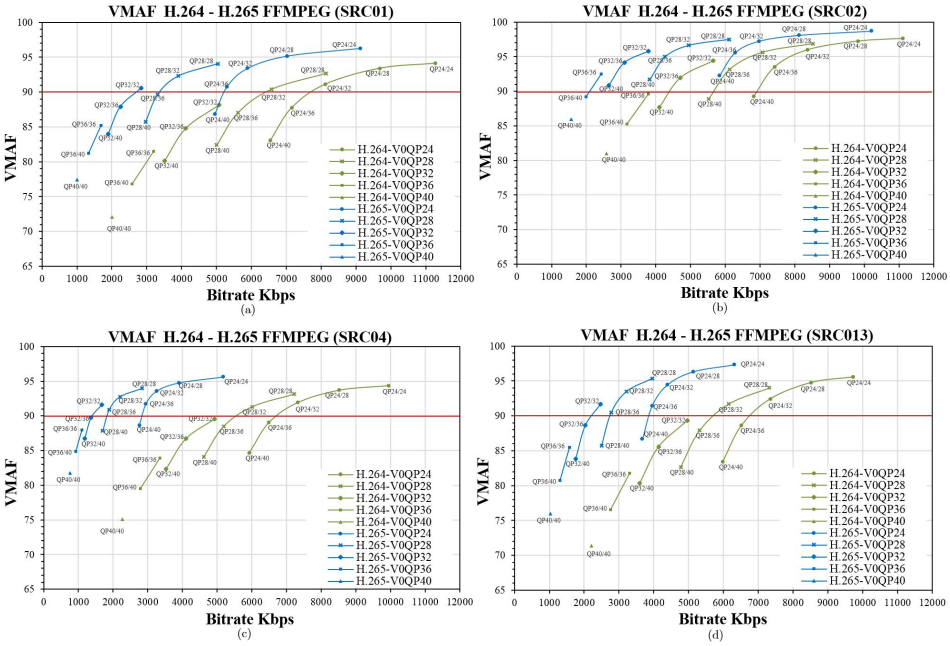


Figura 3.14. Curvas RD (VMAF) evaluación de calidad codificaciones asimétricas FFMPEG H.264 y FFMPEG H.265. (a) SRC01-Car and barrier gate. (b) SRC02-Basket Indoor. (c) SRC04-Library and lamp. (d) SRC013-BigBuckBunny

En las curvas RD de la Figura 3.14, las líneas trazadas en verde y azul corresponden a las codificaciones simétricas y asimétricas obtenidas empleando las implementaciones H.264 FFMPEG y H.265 FFMPEG, respectivamente. Podemos observar como para un determinado nivel de calidad (ej. VMAF=90), la generación de codificaciones asimétricas mediante la combinación de las codificaciones de las vistas por separado, permite obtener una mayor granularidad en los saltos de reducción del *bitrate*, suavizando de esta forma la sensación de degradación de la calidad.

La codificación de las secuencias se realizó bajo la premisa de dominancia del ojo izquierdo: manteniendo fija la vista izquierda V0 en los valores QP 24, 28, 32, 36 y 40 y degradando la vista derecha V1, obteniéndose en cada caso un total de 15 combinaciones, de las cuales 5 corresponden a las codificaciones simétricas QP24/24, QP28/28, QP32/32, QP36/36 y QP40/40 que equivalen al punto de mayor *bitrate*, pero también mayor calidad en cada línea. Las 10 restantes pertenecen a las codificaciones asimétricas. El número de combinaciones se reduce en función del QP de la vista izquierda V0, teniendo así en nuestro caso: 4 codificaciones asimétricas cuando V0=24 y se degrada V1 (QP24/28, QP24/32, QP24/36 y QP24/40), 3 si V0=28 (QP28/32, QP28/36, QP28/40), 2 para V0=32 (QP32/36, QP32/40) y 1 con V0=36 (QP36/40).

Tomando como ejemplo, la secuencia SRC013- *BigBuckBunny* Figura 3.14d, y estableciendo un umbral VMAF=90, se observa que, si disponemos de las codificaciones asimétricas, tomando siempre los puntos de mayor calidad de cada una de las curvas, podemos tener un total de 4 saltos de reducción de *bitrate* con su respectiva reducción de calidad (QP24/24, QP24/28, QP28/28, QP28/32). Por su parte, si se dispone únicamente de las codificaciones simétricas se tendrían solo

dos opciones de reducción de *bitrate* (QP24/24, QP28/28), implicando una reducción más abrupta de la calidad al pasar de una representación a la otra cuando las condiciones de variación de ancho de banda así lo demanden.

3.3.4. Comparación de codificadores mediante pruebas de evaluación subjetiva

Si bien es cierto, que gracias a la aparición de nuevas métricas de evaluación objetiva como VMAF, se ha conseguido mejorar significativamente la evaluación mediante métricas objetivas de la calidad del vídeo, en el caso particular del vídeo estereoscópico 3D, para aproximarnos a la percepción de calidad del vídeo por parte de un usuario, sigue sin ser suficiente obtener el promedio de la evaluación mediante métricas objetivas de cada una de las vistas por separado. Por lo tanto, para el estudio comparativo de los codificadores de vídeo 3D, hemos empleado pruebas subjetivas basadas en las ponderaciones hechas por usuarios siguiendo las metodologías DSIS y PC-DSCS descritas en el Capítulo 2.

En primer lugar, se hará una descripción de los aspectos principales relacionados con la realización de las pruebas subjetivas y finalmente se presentarán los resultados obtenidos en ellas.

3.3.4.a. Metodología y configuración de la prueba de evaluación subjetiva

En primer lugar, se debe tener en cuenta que la evaluación subjetiva de la QoE de un usuario de vídeo 3D se debe evaluar de forma multidimensional, considerando la calidad del vídeo, la calidad de la profundidad y el confort visual. La evaluación subjetiva fue realizada en dos fases: una primera fase basada en el método de referencia completa DSIS descrito en el Capítulo 2, donde se emplea una escala de degradación MOS (*Mean Opinion Score*) de 5 niveles, y el usuario debía ponderar con un valor de 1 (Very annoying) a 5 (Imperceptible), indicando que tan molesta resulta la degradación de cada una de las secuencias codificadas con las diferentes implementaciones de los codificadores H.264/AVC y H.265/HEVC. La valoración la realiza el usuario teniendo en cuenta la relación con la secuencia original, tal como lo plantea la metodología. En esta etapa, el usuario debía realizar dos ponderaciones: una correspondiente a la calidad del vídeo y otra a la componente de profundidad.

La segunda fase, cuyo objetivo era conocer el grado de *disconfort* o fatiga visual de los usuarios siguiendo las indicaciones dadas en la recomendación ITU-T P.916, fue a su vez dividida en dos etapas. Una etapa inicial, definida como “Test preliminar de fatiga y molestias visuales”, realizada justo antes de la realización de las pruebas de visualización de los vídeos 3D, donde se le realizan al usuario un conjunto de preguntas orientadas a conocer su condición visual previa a la realización de las pruebas (tanto cuestionario empleado como los resultados se presentan en el Anexo 1). Y una segunda etapa, definida como “Test de fatiga y molestias visuales”, que corresponde con una serie de preguntas realizadas al usuario, una vez finalizada la visualización de las secuencias de vídeo 3D. El objetivo nuevamente es conocer su estado y condición visual, esta vez tras la visualización de los contenidos de vídeo 3D.

La prueba se realizó sobre un total de 37 usuarios de los cuales el 77% eran hombres y el 23% mujeres, con edades entre los 20 y los 48 años, estando el 62% de los encuestados por debajo de los 30 años. Las condiciones de la prueba se detallan en la Figura 3.12 y Tabla 3.10.

Figura 3.15. Escenario para la realización de pruebas subjetivas con usuarios.

Tabla 3.10. Características del sistema de visualización empleado para las pruebas subjetivas.

Pantalla	Modelo	BenQ XL2720Z LCD
	Resolución	1920x1080
	Frecuencia actualización	60 Hz
	Tamaño	27” de ancho
Distancia a la pantalla	2,5 veces la altura de la pantalla	
Tecnología 3D	NVIDIA GetForce 3D Vision ready	

Como se puede observar en la imagen de la Figura 3.15, las pruebas subjetivas fueron realizadas en un entorno de laboratorio, bajo unas condiciones de iluminación, distancia a la pantalla y equipamiento controladas. Las secuencias son presentadas al usuario por pares, donde el primer estímulo presentado en cada par corresponde a la secuencia original o vídeo sin procesar, y el segundo estímulo corresponde a la secuencia procesada con uno de los sistemas de codificación bajo estudio. La Figura 3.16 representa el esquema temporal usado para la realización de la prueba.

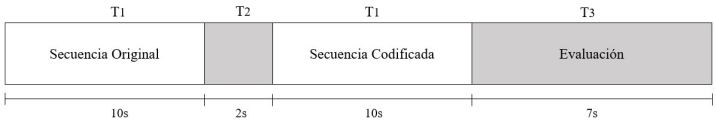


Figura 3.16. Secuencia temporal, prueba subjetiva con metodología DSIS visualización única.

En primer lugar, se presenta al usuario la secuencia original que tiene una duración de 10 s, y tras un descanso de 2s, se le presenta la secuencia codificada. Inmediatamente después, el usuario tiene un periodo de 7 s para realizar su valoración tanto de la calidad de vídeo, como de la profundidad de la secuencia codificada con respecto a la secuencia original.

A continuación, en la Figura 3.17 se presentan los resultados obtenidos para la calidad del vídeo, teniendo en cuenta únicamente las codificaciones simétricas de las 4 secuencias bajo estudio.

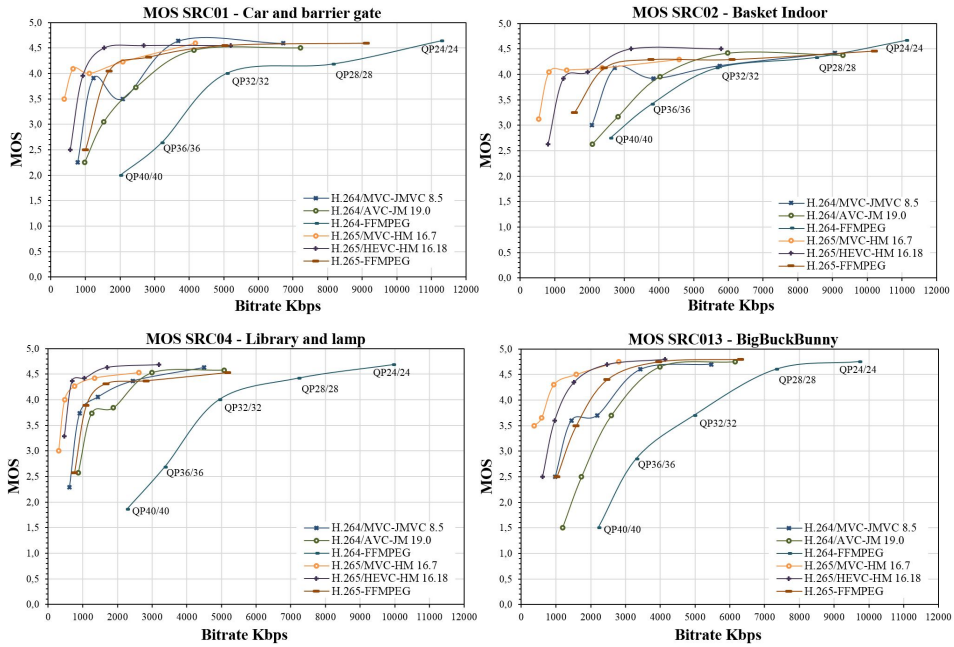


Figura 3.17. Comparación codificadores. Evaluación subjetiva de la calidad en función del MOS.

Considerando que la evaluación subjetiva depende de la opinión del usuario, para eliminar la influencia de las desviaciones individuales de los usuarios en la evaluación, por ejemplo, lo que les gusta y lo que no les gusta, hay que preguntar a muchas personas y promediar sus opiniones (puntuaciones). Sin embargo, es justamente allí donde radica la principal dificultad de las pruebas de evaluación subjetiva, en el poder disponer de un elevado número de usuarios y el tiempo requerido para la realización de las pruebas. En nuestro caso, de acuerdo con las recomendacione de la ITU-T se considera que el número de usuarios encuestados es adecuado y que los resultados obtenidos se encuentran dentro de los limites esperados. No obstante, se ha requerido un pre procesamiento de los datos, orientado a la supresión de los valores críticos, es decir, aquellas ponderaciones que estadísticamente se encuentra fuera del rango.

Según los resultados presentados en la Figura 3.17, se puede observar que, a juicio del usuario, los codificadores basados en H.265 presentan un mejor desempeño en términos de la calidad del vídeo. Así mismo, se observa como secuencias codificadas simétricamente con diferentes parámetros de calidad, por ejemplo, QP=24 y QP=28, ofrecen la misma calidad según la ponderación dada por los usuarios. De esta forma las curvas de distorsión de la calidad con base en el MOS presentan una forma más aplanada en la parte superior y aumentan su pendiente a medida que se aumenta el valor de QP. En general la ponderación de los usuarios es benévola para factores de calidad altos y más exigente a medida que baja la calidad, haciendo que secuencias que según los resultados de las evaluaciones objetivas (60<VMAF<80) podrían estar en un rango de calidad bueno, según los resultados de MOS (MOS<3), las degradaciones percibidas como producto de la codificación resultan, como mínimo, ligeramente molestas.

3.3.4.b. Evaluación subjetiva de las codificaciones asimétricas

En cuanto a la evaluación subjetiva de las codificaciones asimétricas, los resultados obtenidos coinciden con los presentados en trabajos como [82], que indican que si la vista de referencia se codifica con una calidad suficientemente alta y la vista auxiliar se codifica con una calidad inferior, pero por encima de un determinado umbral de VMAF, la degradación de la calidad del vídeo 3D es imperceptible.

Se observa también que, por debajo de este umbral de asimetría, donde empiezan a aparecer sutiles artefactos, la codificación simétrica es mejor que la asimétrica en términos de calidad de vídeo 3D percibida. Por lo tanto, se puede comprobar que la elección entre la codificación asimétrica frente a la simétrica depende del valor de VMAF y, por tanto, de la tasa de bits total disponible.

Con el objetivo de afinar los resultados obtenidos, como parte de la fase de evaluación de la calidad del vídeo, se ha aplicado una segunda prueba basada en el método de comparación por pares (PC) comentado en el Capítulo 2. Con esta prueba se busca medir el confort visual del usuario (y se ha utilizado para comparar entre codificaciones asimétricas y simétricas). En este caso, la metodología consiste en presentar al usuario una secuencia codificada de 10 s, y tras una espera de 2 s presentar una segunda secuencia y dar un tiempo de 7 s para su evaluación. En este caso, empleando las ponderaciones de la Tabla 3.11, el usuario deberá indicar su opinión de la segunda secuencia con respecto a la primera, indicando si la ve mejor, igual o peor, y el grado de mejora o degradación percibida de una secuencia respecto a la otra.

Tabla 3.11. Escala de comparación de pares

Escala de comparación de pares
-3. Much worse
-2. Worse
-1. Slightly worse
0. The same
+1. Slightly better
+2. Better
+3. Much better

Tabla 3.12. Evaluación codificaciones asimétricas FFMPEG H.264 con el método de comparación de pares

	FFMPEG H.264	*SRC01*	*SRC02*	*SRC04*	*SRC013*
	Calidad V0 VMAF	94,08%	97,66%	94,46%	95,60%
	Umbral de asimetría QP24/28	1%	1%	2%	2%
Caso1	% Aceptación respecto a QP24/24	77%	67%	53%	75%
	Calidad V0 VMAF	94,08%	97,66%	94,46%	95,60%
	Umbral de asimetría QP24/32	6%	3%	5%	7%
Caso2	% Aceptación respecto a QP24/24	32%	83%	42%	45%
	Calidad V0 VMAF	92,65%	96,91%	93,25%	94,07%
	Umbral de asimetría QP28/32	5%	3%	4%	4%
Caso3	% Aceptación respecto a QP28/28	36%	62%	59%	61%
	Calidad V0 VMAF	88,14%	94,49%	89,73%	89,42%
	Umbral de asimetría QP32/36	8%	5%	7%	9%
Caso4	% Aceptación respecto a QP32/32	9%	33%	41%	11%

Tabla 3.13. Evaluación codificaciones asimétricas FFMPEG H.265 con el método de comparación de pares

	FFMPEG H.265	*SRC01*	*SRC02*	*SRC04*	*SRC013*
	Calidad V0 VMAF	96,23	98,71	95,64	97,33
	Umbral de asimetría QP24/28	2%	1%	2%	2%
Caso 1	% Aceptación respecto a QP24/24	86%	83%	92%	94%
	Calidad V0 VMAF	96,23	98,71	95,64	97,33
	Umbral de asimetría QP24/32	6%	3%	4%	6%
Caso 2	% Aceptación respecto a QP24/24	77%	79%	75%	72%
	Calidad V0 VMAF	94,00	97,52	94,03	95,41
	Umbral de asimetría QP28/32	4%	2%	3%	4%
Caso 3	% Aceptación respecto a QP28/28	68%	71%	83%	67%
	Calidad V0 VMAF	90,49	95,83	91,76	91,69
	Umbral de asimetría QP32/QP36	6%	4%	4%	7%
Caso 4	% Aceptación respecto a QP32/QP32	45%	58,34	50%	39%

En la Tablas 3.12 y 3.13 se presentan los resultados de la evaluación subjetiva empleando el método de comparación por pares de las codificaciones asimétricas empleando los codificadores FFMPEG H.264 y FFMPEG H.265, respectivamente. El estudio se realiza sobre las secuencias SRC01, SRC02, SRC04 y SRC013, y se comparan las representaciones simétricas con la codificación asimétrica que se forma al degradar la vista derecha V1 con el siguiente valor de QP. Por ejemplo, QP24/28 indica que la vista izquierda se ha codificado con un QP=24 mientras que en la vista derecha se ha usado QP=28. Como parte del análisis, para cada uno de los casos se calcula: la calidad de la V0 en términos del VMAF, el porcentaje de asimetría de calidad en función del VMAF entre V0 y V1 para la representación asimétrica que se está analizando y, finalmente, el porcentaje de aceptación respecto a la representación simétrica correspondiente. Como porcentaje de aceptación se entiende el porcentaje de usuarios que, al ver la representación asimétrica y la representación simétrica correspondiente, considera que la calidad de la representación asimétrica es igual o mejor que la de la representación simétrica.

De esta forma, analizando los resultados de la Tablas 3.12, observamos que para las codificaciones H.264 FFMPEG, siempre que la calidad de la vista izquierda (V0) sea lo suficientemente buena (VMAF=90) y el umbral de asimetría sea inferior al 5%, un porcentaje superior al 50% de los usuarios consideran que la calidad de la representación asimétrica es igual o mejor que la de la representación simétrica correspondiente (estas situaciones corresponden a los campos sombreados dentro de la tabla). Así mismo, se evidencia que, aunque la calidad de V0 sea alta (VMAF superior a 90), si el umbral de asimetría de calidad entre V0 y V1 es igual o superior al 5% el porcentaje de aceptación por parte de los usuarios será inferior al 50%.

Por su parte, en el caso de las codificaciones H.265 FFMPEG, se aprecia en la Tabla 3.13 que a pesar de que las codificaciones se han realizado empleando los mismo valores del parámetro QP que para las codificaciones H.264 FFMPEG, nuevamente la mayor eficiencia del codificador basado en HEVC hace que la calidad tanto de V0 como de V1 sea mayor en términos del VMAF, ampliándose de esta forma el umbral de asimetría y haciendo que solo en los casos donde la calidad de V0 se aproxima a VMAF=90 y el umbral de asimetría supera el 6% el porcentaje de aceptación por parte de los usuarios sea inferior al 50%.

3.3.5. Evaluación subjetiva usando el estándar ITU- T P1.203

La mayoría de las métricas de calidad, como PSNR, SSIM y VMAF, son métricas de referencia completa, lo que significa que comparan el archivo codificado con la fuente para calcular la puntuación. Esto significa que la métrica sólo puede aplicarse cuando el archivo fuente está disponible. Por otra parte, si bien la evaluación subjetiva con usuarios es la mejor forma de medir la QoE, la implementación de pruebas subjetivas resulta muy costosa, especialmente si se tienen en cuenta el tiempo y disponibilidad de usuarios requeridos para el proceso. Por tal motivo, como se comentó en el Capítulo 2, la ITU-T ha publicado recientemente modelos de calidad de vídeo en el contexto del *streaming* adaptativo HTTP, concretamente la Rec. P.1203, que integra puntuaciones de calidad de vídeo y de audio en una puntuación para una secuencia de vídeo de hasta 5 minutos, incluidos los efectos de carga inicial y de las interrupciones, y la Recomendación ITU-T P.1204, que es un conjunto de modelos de alto rendimiento para secuencias UHD/4K de 60 fps codificadas con H.265/HEVC o VP9.

La norma ITU-T P.1203 [37][83] es la primera métrica que intenta medir la calidad de experiencia incorporando tanto la calidad visual como el rendimiento de la calidad de servicio y utiliza un modelo basado en el flujo de bits que puede aplicarse eficazmente en cualquier parte del sistema de distribución. Sin embargo, hasta la fecha no parece haber ninguna herramienta comercial que integre el modelo. No obstante, existe un software de referencia para la norma ITU-T P.1203 que puede utilizarse libremente con fines de investigación y que permite obtener una aproximación de la calidad del vídeo en función del MOS. Al igual que sucede con las métricas objetivas PSNR, SSIM y VMAF, la aplicación de esta herramienta en un sistema de transmisión de vídeo 3D, no deja de ser más que una aproximación, ya que su ponderación se obtiene a partir del promedio del resultado obtenido para cada una de las vistas por separado, y no se tienen en cuenta aspectos como la supresión binocular del sistema de visión humano en el caso de las codificaciones asimétricas.

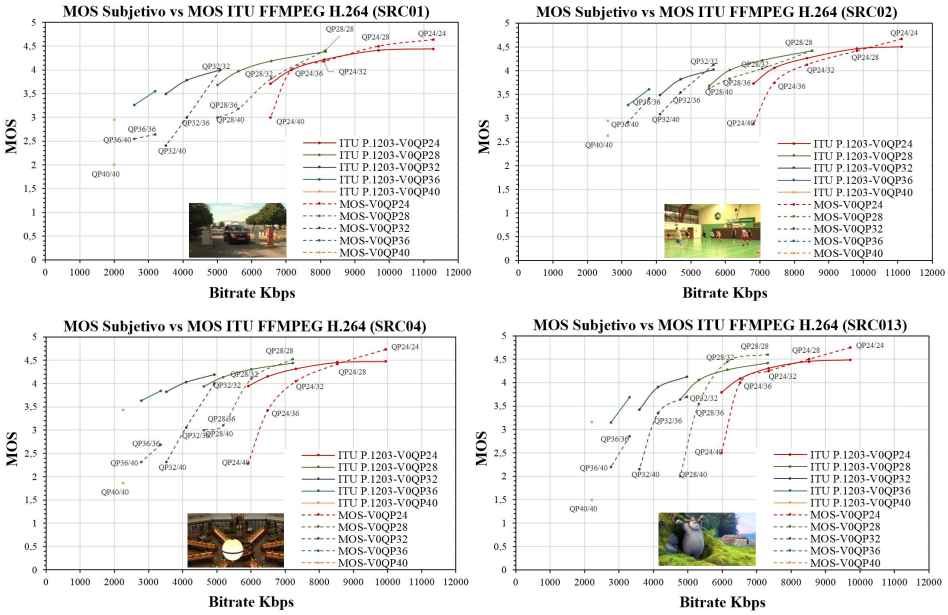


Figura 3.18. Evaluación subjetiva de la calidad en función del MOS vs evaluación subjetiva MOS ITU-T P.1203

En las gráficas de la Figura 3.18 se puede observar como tanto para las codificaciones simétricas como para las asimétricas, para aquellas secuencias donde la vista dominante (V0) ha sido codificada con un valor de QP bajo (24-28), el MOS obtenido mediante la evaluación subjetiva de los usuarios se aproxima o es incluso superior a la ponderación de MOS obtenida mediante el promedio de MOS calculado para las vistas por separado empleando la implementación de la recomendación ITU-T P.1203. Así mismo, en aquellos casos donde QP V0 es bajo (24-28) y la asimetría respecto a V1 no es muy alta, en el MOS subjetivo se observa la prevalencia de la vista de mayor calidad. Por el contrario, en aquellos casos donde la asimetría entre las dos vistas es muy alta, o a medida que el valor de QP se incrementa, se observa que la valoración dada por los usuarios es más exigente, obteniéndose valores de MOS por debajo del MOS promedio que se obtendría empleando la recomendación ITU-T P.1203.

Como se comentó anteriormente, una vez finalizadas las pruebas de evaluación subjetiva, que tuvo una duración total de 1 hora 20 minutos, incluida la sesión de entrenamiento y la presentación de la prueba, a los usuarios se les aplicó una encuesta desarrollada con base en la recomendación ITU-T P.916[34] (*Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video*) según se describe en el Capítulo 2, cuyo objetivo era medir la fatiga visual de los usuarios frente a la visualización de vídeo 3D. Si bien es cierto que el contexto de las pruebas no corresponde con un escenario convencional como podría ser la visualización de una película, partido de fútbol etc, nos da una idea de la predisposición del usuario frente al servicio. Según los resultados obtenidos, los síntomas de fatiga visual más importantes fueron:

- La fatiga ocular, puesto que un 72% de los usuarios presentó tensión ocular, un 64% fatiga en los ojos, un 44% resequedad y un 40% pesadez en los parpados. Así mismo, el

40% de los participantes manifiesta haber tenido que cerrar los ojos en algún momento durante el experimento para aclarar la visión.

- Respecto a otros síntomas, el 40% de los encuestados presentó rigidez en el cuello, un 36% dolor en la parte delantera de la cabeza y un 12% dolor en las sienes. También se resalta que un 56% manifestó sentirse aturdido y un 20% con sensación de mareo.

Los resultados generales obtenidos y el cuestionario empleado se presentan en el Anexo 1.

3.4. Selección de representaciones para la transmisión

Considerando que el siguiente paso en el desarrollo de esta tesis, será la trasmisión del vídeo 3D a través de Internet, con base en los resultados obtenidos de la evaluación de calidad tanto objetiva como subjetiva, se ha implementado un algoritmo que permite la selección automática de las secuencias que, según su calidad, deben estar disponibles en el servidor.

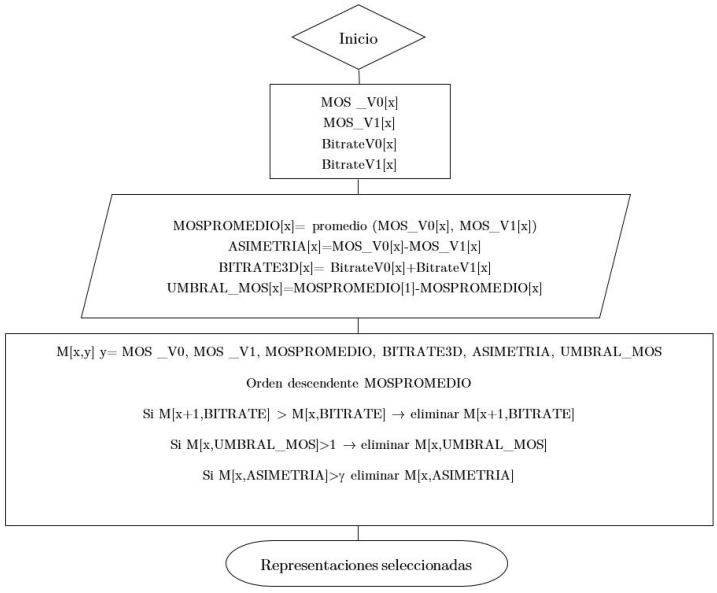


Figura 3.19. Diagrama de flujo del proceso de selección de las representaciones disponibles en el servidor (convex hull).

El diagrama de la Figura 3.19 resume el proceso de selección de representaciones. El primer paso consiste en conocer el *bitrate* de cada una de las secuencias codificadas y calcular, mediante la implementación ITU-T P.1203, el MOS para cada una de las vistas que conforman las distintas representaciones. A partir de estos valores se calculan MOSPROMEDIO, ASIMETRIA, BITRATE3D y UMBRAL_MOS. UMBRAL_MOS se calcula como el porcentaje que representa cada representación, respecto a la de mejor calidad, en nuestro caso QP24/24.

Posteriormente, se conforma la matriz M[x,y] donde x corresponde a cada una de las codificaciones disponibles de la secuencia, en nuestro caso 15. Mientras y=MOS _V0, MOS _V1,

MOSPROMEDIO, BITRATE3D, ASIMETRIA, UMBRAL_MOS. Y a partir de este punto se aplica un procedimiento de 4 pasos:

1. Organizar la matriz `M[x,y]` en orden descendente tomando como referencia la columna `MOSPROMEDIO`.
2. Leer la columna `BITRATE3D` aplicando el criterio:
   Si `M[x+1,BITRATE] > M[x,BITRATE]` → eliminar `M[x+1,BITRATE].`
3. Leer la columna `UMBRAL_MOS` aplicando el criterio:
   si `M[x,UMBRAL_MOS]<80` → eliminar `M[x,UMBRAL_MOS]`, de manera que solo aquellos valores comprendidos comprendidos en un rango<=20% alrededor de la mejor representación serán tenidos en cuenta.
4. Leer la columna `ASIMETRIA` aplicando el criterio:
   si `M[x,ASIMETRIA]`>γ eliminar `M[x,ASIMETRIA]`, donde el valor de γ dependerá del salto entre los distintos valores de QP usados para las codificaciones. En este punto, nos valemos de la tolerancia del HVS a la asimetría, que hace posible que cuando se tiene un vídeo en el que las dos vistas no tienen la misma degradación, si la asimetría es BAJA la calidad percibida por el usuario del vídeo 3D se aproxime a la de la vista de mejor calidad. Así mismo, en el caso de vídeos con ALTA asimetría tiene mayor peso la vista de peor calidad, y la calidad percibida por el usuario del vídeo 3D se aproxima más a ese valor. En nuestro caso hemos empleado un valor de γ>1

La Figura 3.20 muestra las representaciones seleccionadas para cada secuencia de vídeo una vez se aplica el algoritmo de selección. Como se puede observar los resultados obtenidos se corresponden en un alto porcentaje con las representaciones que conforman el denominado *Convex Hull* [84] o Envolvente Convexa.

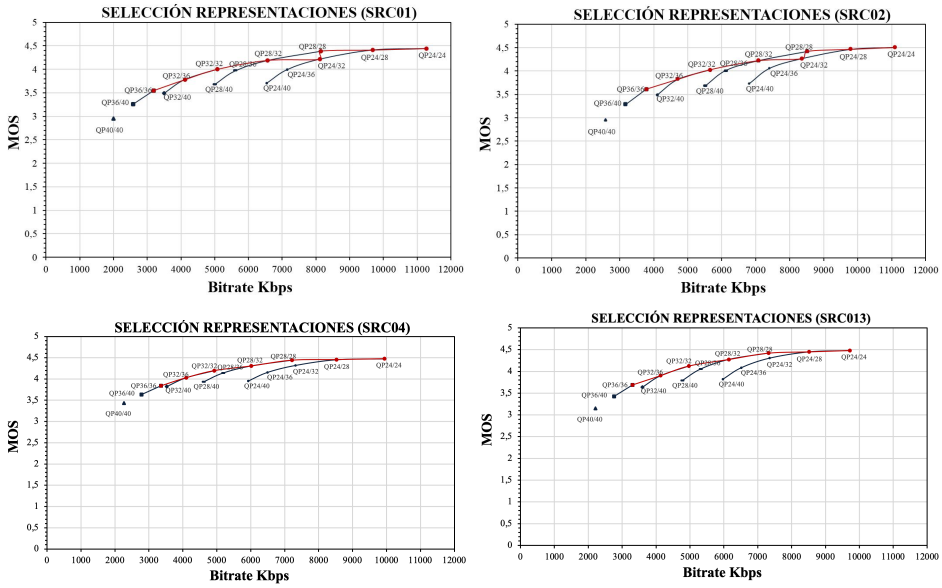


Figura 3.20. Convex Hull. Selección de representaciones.

3.5. Conclusiones

Como resultado del estudio de los esquemas de codificación de vídeo 3D (Multivista, simultcast simétrica y simulcats asimétrica), y la evaluación de su impacto en la QoE del usuario, podemos concluir lo siguiente:

Se ha realizado una comparación de las implementaciones multivista de los estándares H.264 y H.265, con respecto a la codificación estéreo simulcast, es decir, codificando cada vista por separado. En esta comparación se ha podido demostrar que, a pesar de la eficiencia en términos de la reducción de bitrate manteniendo la calidad del vídeo que ofrecen los codificadores MVC gracias al aprovechamiento de la correlación entre vistas, otros factores como el tiempo de procesamiento y el requerimiento de hardware o software específico para la decodificación hacen que el uso de los codificadores de vídeo multivista no sea una buena alternativa en un entorno orientado a la prestación de servicios masivos de streaming de vídeo 3D.

El análisis de la información espacial (SI) y temporal (TI) de las secuencias de vídeo, resulta clave en el proceso previo a la codificación, ya que permite tomar decisiones respecto al grado de degradación soportado por cada secuencia, contribuyendo a una mejor selección de los parámetros de codificación en función de las necesidades del sistema.

Además, se ha comprobado que la eficiencia en términos de la reducción del bitrate, de las codificaciones hechas con H.265/HEVC respecto a las realizadas con H.264/AVC se incrementa a medida que la calidad del vídeo se reduce al aumentar el QP. Así mismo, vemos como para un mismo valor de QP, la ganancia de bitrate de las codificaciones hechas con H.265 FFMPEG respecto a las hechas con H.264 FFMPEG es menor en las secuencias con un nivel de movimiento

alto/medio y se incrementa en las secuencias con nivel de movimiento bajo, confirmándose de esta manera que tanto el grado de movimiento de la secuencia de vídeo como el parámetro de cuantificación inciden directamente en la eficiencia de codificación de un estándar respecto al otro.

Adicionalmente, en el caso particular del vídeo estéreo, se demuestra que el uso de codificaciones asimétricas es una opción válida para conseguir una reducción de bitrate manteniendo la calidad percibida por el usuario. En este sentido, se ha hecho un estudio en profundidad basado en la hipótesis del fenómeno de supresión binocular, y mediante la realización de pruebas subjetivas con usuarios se ha comprobado que, manteniendo unos umbrales de degradación y asimetría entre las dos vistas, la calidad global percibida por el usuario se aproxima a la de la vista de mejor calidad.

Finalmente, utilizando el concepto del convex hull y con base en los resultados tanto de las evaluaciones objetivas como subjetivas de la calidad del vídeo codificado, se ha propuesto una metodología para la selección de las secuencias que deben estar disponibles en el servidor de contenidos para streaming HTTP adaptativo. El algoritmo propuesto se basa en la asimetría entre vistas y el umbral de MOS, para definir las mejores representaciones.

Capítulo 4

Sistema de pruebas para el estudio de la QoE del streaming adaptativo de vídeo sobre HTTP

Es importante tener en cuenta que en un sistema de distribución de vídeo las degradaciones debidas a la producción y la codificación de los contenidos, así como las perdidas y/o errores durante la transmisión, pueden degradar la calidad del vídeo recibida y percibida por el usuario. Por tanto, como parte de este trabajo, además del estudio del rendimiento de los estándares de codificación de vídeo más populares H.264/AVC, H.265/HEVC y sus correspondientes extensiones para vídeo 3D abordada en el Capítulo 3, se ha trabajado en la implementación de un *framework* que permita la evaluación experimental de la QoS (*Quality Of Service*) y la QoE (*Quality Of Experience*) de un sistema de transmisión adaptativa de vídeo sobre HTTP (DASH), enfocado en la distribución de contenidos multimedia 2D y 3D, utilizando diferentes escenarios y condiciones de red, concretamente variaciones de ancho de banda.

Como primer paso, para facilitar el desarrollo, se ha hecho una descomposición del procedimiento en diferentes etapas como se muestra en la Figura 4.1.

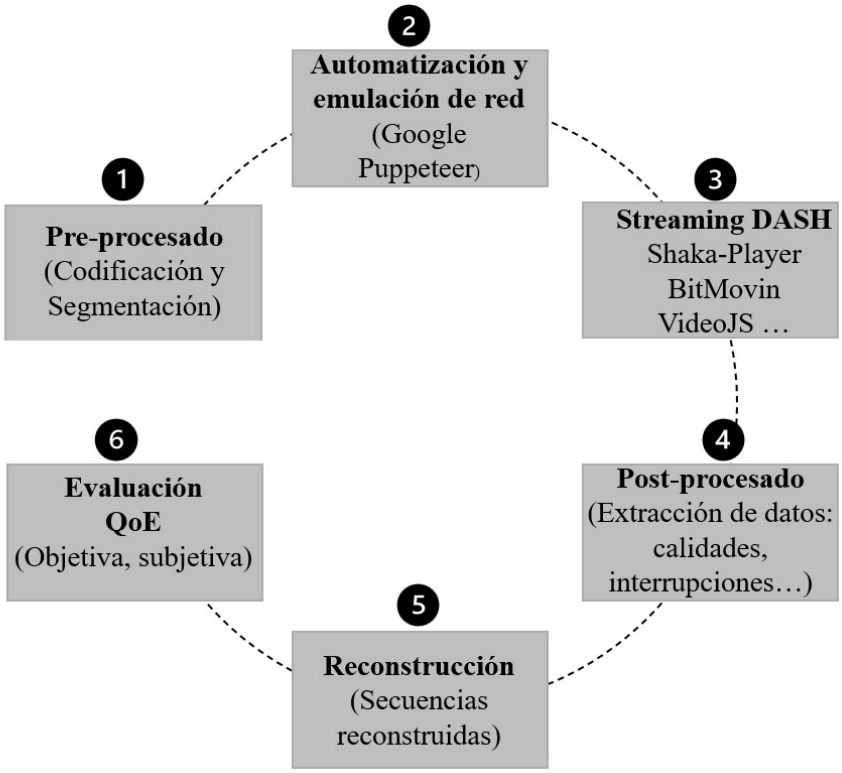


Figure 4.1. Etapas del proceso de transmisión adaptativa de video 3D y evaluación de la QoE.

En primer lugar (etapa 1), encontramos el procesado de los datos referente a la codificación del vídeo y la generación de los segmentos DASH que estarán disponibles en el servidor. El siguiente paso (etapa 2) se refiere a la automatización y emulación de la red; en este punto se definen las condiciones de ancho de banda, los retardos, las pérdidas y las características de los dispositivos que procesan y *renderizan* los contenidos de vídeo. Teniendo en cuenta que actualmente existe una gran variedad de implementaciones de los estándares de transmisión adaptativa, la selección del reproductor de vídeo representa un paso significativo de este proceso (etapa 3). Esta selección se realiza teniendo en cuenta aspectos como: formatos soportados, algoritmo de adaptación, código abierto o propietario, etc. Finalmente, una vez realizada la emulación de la transmisión de vídeo, se ejecutan las etapas de post-procesado, dedicadas al análisis de la información obtenida (etapa 4), y a la reconstrucción del vídeo (etapa 5) para su posterior evaluación tanto objetiva como subjetiva (etapa 6), dando por finalizado el proceso de evaluación.

En resumen, el sistema propuesto permite realizar mediciones de rendimiento de forma automatizada y sistemática para la evaluación de los sistemas DASH en el servicio de transmisión de vídeo 2D y 3D. Se ha utilizado la herramienta Puppeteer, desarrollada por Google, que permite automatizar acciones en el navegador web. Dentro de las acciones automatizadas se encuentran: iniciar la reproducción, generar variaciones de ancho de banda, guardar los resultados de los procesos de cambio de calidad, marcas de tiempo, paradas, etc. A partir de la información

recopilada a partir de las estadísticas de red y demás información proporcionada por el navegador, se realiza la reconstrucción del vídeo visualizado, así como la extracción de métricas de calidad y la evaluación de la QoE de los usuarios.

La distribución de este capítulo es la siguiente. En la sección 4.1 se comentan las plataformas para la transmisión de vídeo 3D, haciendo referencia a los sistemas de transmisión de vídeo estereoscópico en redes IP. La sección 4.2 se presenta la arquitectura del sistema de pruebas (*testbed*) propuesto y cada uno de sus elementos. La sección 4.3 muestra los resultados en términos de métricas objetivas y subjetivas de la evaluación del sistema de transmisión adaptativa de vídeo. Finalmente, en la sección 4.4 se presentan las principales conclusiones del capítulo.

4.1. Transporte de vídeo 3D

La prestación de servicios de vídeo estereoscópico 3D requiere mecanismos de transporte eficientes, que permitan acoplar el volumen de datos de vídeo de las dos vistas en los enlaces de transmisión, que generalmente son canales con pérdidas y con ancho de banda y retardos que varían con el tiempo.

Mientras que el cine en 3D no tenía que enfrentarse a los problemas de transporte, para la consolidación del servicio de 3DTV en los hogares, además de los problemas relacionados con la tecnología de visualización requerida por parte de los usuarios, el sector de la radiodifusión requería de un almacenamiento y distribución eficaz de los contenidos. La Figura 4.2 muestra los diferentes canales mediante los que los contenidos 3D pueden llegar a los hogares.

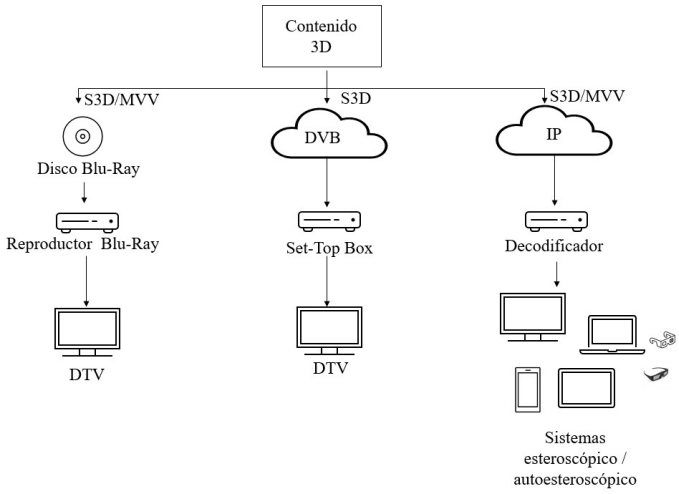


Figure 4.2. Plataformas para el transporte de vídeo 3D.

4.1.1. Sistemas de almacenamiento de vídeo 3D

En primer lugar, encontramos los discos Blu-Ray (BD) como alternativa de almacenamiento para la distribución de los contenidos 3D. Los BD tienen una capacidad de grabación de 50 GB, que es más que suficiente para películas de larga duración en formato de alta definición. Disponen de capacidad suficiente para almacenar flujos adicionales, lo que resulta muy interesante para el almacenamiento de contenidos 3D estereoscópicos o multivista. Como se comentó anteriormente, los formatos *frame-compatible* tienen la ventaja de poder utilizar los dispositivos 2D existentes para las aplicaciones 3D, pero sufren una pérdida de resolución que no puede recuperarse completamente sin alguna información de mejora. Por este motivo, un formato de vídeo estereoscópico *frame-sequential full HD* fue considerado como el principal candidato para la estandarización. En 2009, la BDA (*Blu-ray Disc Association*) anunció oficialmente la creación del estándar 3D para discos Blu-Ray. Las indicaciones de la BDA fueron que, "las especificaciones de Blu-Ray 3D debían codificar el vídeo 3D usando MVC Stereo High Profile", una extensión del estándar ITU-T H.264 Advanced Video Coding (AVC) soportada por los reproductores de discos Blu-ray disponibles [4].

4.1.2. Sistemas de transmisión de vídeo 3D

La transmisión de vídeo 3D a usuarios finales con anchos de banda y terminales de visualización diferentes, es uno de los mayores retos para llevar los contenidos multimedia 3D al hogar y a los dispositivos móviles. Como se puede observar en la Figura 4.1, como alternativa a los sistemas basados en Blu-ray, para la recepción de los contenidos 3D en el hogar existen dos plataformas principales para la difusión de vídeo 3D: las plataformas de DTV (*Digital Television*) y la plataforma IP (*Internet Protocol*).

Según el consenso de los expertos, la 3DTV solo podría perdurar en el tiempo, si era compatible con el servicio de televisión 2D convencional, si requería una baja sobrecarga de transmisión adicional, y si la calidad visual percibida y la comodidad de visualización eran mejores que las de la televisión 2D [85]. Por este motivo, en sus inicios las empresas de radiodifusión eligieron las plataformas de DTV ya existentes para dar inicio al servicio de difusión de 3DTV. Por ejemplo, se utilizaba el estándar DVB (*Digital Video Broadcasting*) para emitir vídeo estereoscópico utilizando formatos compatibles. Aunque el ancho de banda no es un problema importante en la infraestructura de cable, los dispositivos requeridos para decodificar y dar formato al contenido para su visualización sí lo son. Si bien el transporte de vídeo 3D a través de plataformas de TVD queda fuera del ámbito de este trabajo, se puede encontrar más información sobre esto en trabajos como [86] y [87], entre mucho otros.

A diferencia de la difusión tradicional, los servicios basados en redes IP se ofrecen a distintas velocidades y costes a través de diversas infraestructuras físicas, como las redes de telecomunicaciones fijas o inalámbricas. Además, es posible ofrecer una variedad de arquitecturas de servicio, como servidor-cliente (*unicas*t) o peer-to-peer (*multicast*), utilizando diferentes opciones de protocolo de transporte, como HTTP/TCP o RTP/UDP, sobre la plataforma IP. En resumen, IP proporciona un canal más flexible para transmitir tantas vistas como requiera el terminal de visualización del usuario.

Como se ha mencionado anteriormente, el creciente interés por mejorar la QoE de los usuarios, haciendo un mejor uso de los recursos de la red, ha dado lugar al *streaming* adaptativo de vídeo

sobre HTTP (HAS), siendo el estándar DASH (*Dynamic Adaptive Streaming over HTTP*) su ejemplo más representativo [88][89]. DASH permite una adaptación flexible de la calidad del vídeo a los recursos de red disponibles y a las capacidades del dispositivo cliente. Por lo tanto, permite una mejor gestión del estado del *buffer*, el control de las interrupciones durante la reproducción y una mejor gestión del ancho de banda. Con DASH, los servidores de contenidos ofrecen múltiples versiones (representaciones) del mismo vídeo, que pueden estar asociadas a diferentes tasas de bits, resoluciones de vídeo y tasas de muestreo de audio entre otros factores. A continuación, cada representación se divide en segmentos temporales (chunks) y se almacena en un servidor HTTP. Toda la información asociada a los segmentos de vídeo, como la resolución, duración y la tasa de bits media, se especifican en un archivo denominado MPD (*Media Presentation Description*). El reproductor multimedia, en el lado del cliente, ejecuta un algoritmo de tasa de bits adaptativa (ABR) o algoritmo de adaptación, para seleccionar los segmentos de la representación más adecuada para su descarga en cada momento, en función del ancho de banda estimado y/o el estado del *buffer*, de modo que se puedan evitar las interrupciones y el ancho de banda disponible se pueda aprovechar de la mejor manera. La tecnología DASH ha sido adoptada por una amplia gama de aplicaciones y proveedores de contenidos de vídeo, como YouTube [90] o Netflix [91]. Esto ha hecho que el estudio de las prestaciones de los sistemas de transmisión adaptativa de vídeo en Internet y su impacto en la QoE del usuario sean temas de investigación con numerosas aportaciones en los últimos años [92][93].

En la bibliografía se pueden encontrar varias publicaciones centradas en el estudio teórico o experimental de los diversos reproductores y algoritmos de adaptación. En [94], los autores presentan un estudio sobre el estado del arte de los algoritmos de adaptación de *bitrate* para HAS. Los esquemas de adaptación de *bitrate* se clasifican en función de la entidad del sistema donde se implementa la lógica: adaptación basada en el servidor, adaptación basada en el cliente, adaptación asistida por la red, así como adaptación híbrida, que utiliza información de cualquier combinación del cliente, el servidor o los servidores y la red. En [95] y [96] los autores llevaron a cabo evaluaciones experimentales de diferentes reproductores/algoritmos adaptativos de transmisión sobre HTTP, comerciales y de código abierto. Como en la mayoría de los trabajos de este tipo, el objetivo se centró en los mecanismos de adaptación de la tasa de bits y en cómo se enfrentan a la evolución del ancho de banda de la red.

Hasta ahora, la metodología experimental se ha basado en la conexión entre dos ordenadores (Servidor y Cliente) e implica que el dispositivo que ejecuta el reproductor de vídeo también ejecute un *sniffer* de paquetes y un emulador de red, como DummyNet, NetEm o Netlimiter, entre otros.

Asimismo, además de la necesidad de entender la caja negra que representa el algoritmo de adaptación, el objetivo en muchos casos es su optimización [97] considerando, por ejemplo, la variación del tamaño del segmento, el ancho de banda estimado de la ruta y la ocupación actual del *buffer*, para predecir con exactitud el tiempo necesario para descargar el siguiente segmento, o su valoración con base en los resultados de la evaluación subjetiva de los usuarios [98].

Sin embargo, existen pocos trabajos orientados a la implementación de un sistema de referencia común que permite realizar y replicar, de forma fácil y accesible, pruebas de rendimiento de los reproductores multimedia. Por ejemplo, encontramos una propuesta en [99], que apunta a un objetivo equivalente al nuestro, y se centra en el estudio de reproductores multimedia para

entornos web. Sin embargo, como diferencia remarcable, la solución propuesta implica el uso de tres servidores, uno para el almacenamiento de contenidos (Web Server), otro para la emulación de las variaciones de la red (Mininet) y, finalmente, un tercer servidor (Selenium Server) que automatizará el acceso a los diferentes reproductores.

En un contexto de condiciones de red variables y transmisión adaptativa, la evaluación de la calidad del vídeo 3D en el terminal del usuario final es crucial. Se ha trabajado mucho en la evaluación de la QoE para la transmisión de vídeo 2D y ahora podemos encontrar algunos trabajos para los servicios de transmisión de vídeo 3D. La evaluación de la calidad se hace empleando tanto métricas de evaluación objetiva, como métodos subjetivos: (i) La evaluación objetiva se basa en métricas bien conocidas que han sido presentadas en el Capítulo 2 y empleadas para la evaluación de la calidad del vídeo durante la etapa de codificación en el Capítulo 3. Nos referimos al PSNR [26], SSIM [28] y VMAF [100]. Sabemos que, si bien la implementación de las métricas objetivas es sencilla, tiene un bajo coste computacional y pueden ser replicadas, se requiere disponer de la secuencia original para su implementación; (ii) En cuanto a la evaluación subjetiva, uno de los objetivos es conseguir integrar tantas métricas como sea posible para mejorar la estimación de la QoE, tanto en tiempo real como para el análisis fuera de línea. Así, nos basamos en la implementación de la Recomendación UIT-T P.1203 [37][83] que se ha convertido en el primer modelo estandarizado para la evaluación de la QoE de los servicios de transmisión de audio y vídeo adaptativos.

4.2. Arquitectura del sistema de pruebas para el estudio automatizado del rendimiento de un sistema DASH de transmisión de vídeo 3D

La Figura 4.3, presenta la arquitectura y los diferentes componentes que conforman el sistema de pruebas (*testbed*) propuesto.

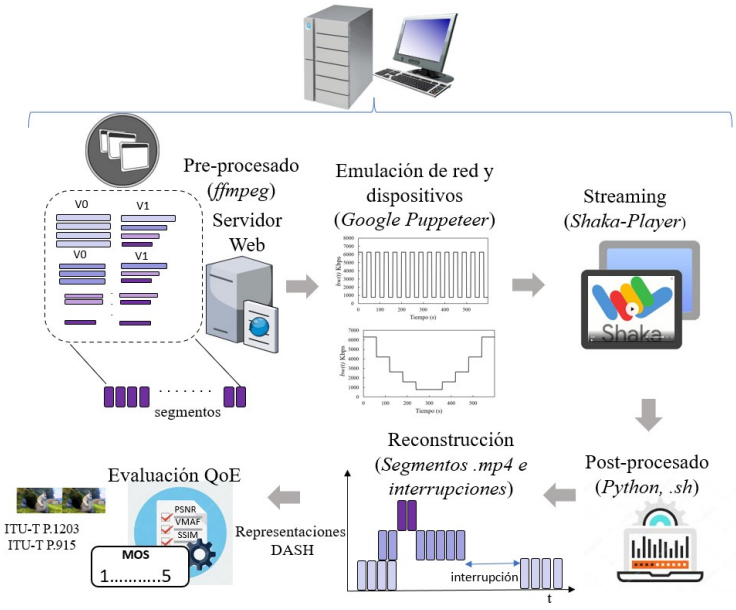


Figure 4.3. Arquitectura del sistema de pruebas propuesto.

Como se puede ver en la Figura 4.3, dentro de la arquitectura del sistema de pruebas, podemos identificar 6 componentes:

- Pre-procesado y codificación de vídeos en formato 3D.
- Servidor web basado en HTTP que aloja los contenidos pre-procesados (vídeo codificado y segmentado).
- Herramienta para la emulación de condiciones de red (variaciones de ancho de banda y/o latencia)
- Cliente con el reproductor DASH.
- Herramientas para la obtención y, post-procesado de los datos relativos a la transmisión y la extracción de métricas.
- Estimación de la calidad de experiencia (QoE).

Todos los elementos del sistema propuesto, tanto el servidor como el cliente, se han implementado en un PC Intel Core i7 de novena generación con Ubuntu (versión 18.04.2 LTS). Los bloques son modulares y pueden ser ejecutados en la misma máquina. Así, se podría utilizar un sistema de virtualización o basado en contenedores, como Docker, para desplegar cada uno de los módulos en el mismo ordenador. El hecho de implementar todo el proceso del *testbed* en un único equipo permite que el sistema desarrollado sea versátil y fácilmente exportable y replicable por la comunidad científica. En las siguientes subsecciones se describen detalladamente cada uno de los componentes.

4.2.1. Codificación de vídeo 3D y servidor web

El servidor web aloja el contenido de vídeo para la transmisión adaptativa a través de HTTP. En nuestro caso se ha configurado un servidor web Apache sobre la plataforma Linux. Como paso previo a la generación de los contenidos DASH y su almacenamiento en el servidor web, se debe realizar el proceso de codificación del vídeo 3D. Como se comentó en el capítulo anterior, las secuencias serán procesadas en formato estereoscópico, lo que permite codificar cada una de las vistas de forma independiente y facilita la obtención de codificaciones tanto simétricas como asimétricas.

Atendiendo a los resultados de la comparación entre los dos principales codificadores de referencia H.264 y H.265 realizado en el Capítulo 3, el proceso de codificación se realiza empleando la implementación para FFMPEG del estándar H.264 (libx264) y el parámetro de cuantización (QP) como parámetro de restricción de codificación. La principal ventaja de H.264 es su universalidad ya que es soportado por todos los navegadores web y, plataformas de OTT/Smart TV. En cuanto a su implementación en FFMPEG, ofrece significativas ventajas en cuanto al coste computacional y tiempo de procesamiento, respecto a las implementaciones de referencia. No obstante, el sistema está abierto a otros codificadores y a cualquier otra configuración de codificación, incluyendo otros parámetros como el *Bitrate* o el CRF (*Constant Rate Factor*), que según el tipo de servicio que se desee ofrecer pueden resultar de mayor interés. El uso de otros codificadores, como HEVC, VP9 o AV1, dependerá de su compatibilidad con el navegador web, el reproductor DASH y la implementación para la evaluación subjetiva.

Las representaciones que estarán disponibles en el servidor se eligieron empleando el algoritmo de selección de representaciones para transmisión, presentado en la sección 3.4 del Capítulo 3. Las secuencias fueron codificadas con valores de QP en un rango de 18 a 42. Se realizó una prueba preliminar de calidad perceptiva para seleccionar los parámetros de cuantificación que abarcaban un amplio rango de calidad visual. Nuevamente, las secuencias de vídeo estereoscópico se construyeron empleando dos esquemas de codificación diferentes, que en adelante se denominarán HRC (Circuitos de Referencia Hipotéticos) [101]:

- Secuencias estereoscópicas codificadas simétricamente. Las vistas izquierda y derecha se codifican con el mismo parámetro de cuantificación QP en el rango de 18 a 42 con dos pasos de longitud. En lo sucesivo, esta condición se denominará SYM.
- Vista izquierda (V0) se codifica con QP en el rango de 18 a 42 con dos pasos de longitud y para la vista derecha (V1) el parámetro de cuantificación varía en el rango de 20 a 42. Estas condiciones se denominarán ASYM.

Una vez codificadas las secuencias de vídeo, se generan los segmentos DASH para los que se ha elegido una duración de 5 segundos, así como del archivo MPD, que contiene toda la información sobre las diferentes calidades de vídeo usadas y el ancho de banda de cada una. Como formato de entrega se empleará MPEG-DASH, aunque también existe la opción de utilizar HLS (*HTTP Live Streaming*), que es el protocolo de *streaming* multimedia basado en HTTP implementado por Apple.

4.2.2. Emulación de condiciones de red y terminales cliente (Puppeteer)

La transmisión adaptativa de vídeo se puede dar en entornos heterogéneos y un solo cambio en las condiciones de contexto puede tener un gran impacto en el comportamiento del reproductor y, muy probablemente, en la experiencia de visualización del usuario final. Como herramienta para la automatización de las pruebas extremo a extremo, incluida la emulación de las variaciones de red, se empleará Puppeteer [102]. Puppeteer es la nueva librería desarrollada por Google que ofrece una interfaz basada en node.js que permite ejecutar y controlar Chrome (o Chromium) en modo *headless* a través del protocolo DevTools mediante la ejecución de un script desde la línea de comandos. En concreto, para esta etapa se establece una sesión CDP (*Chrome Devtools Protocol*) con la web en la que se encuentra alojada la implementación del Shaka Player. Mediante el acceso a los recursos `Network.emulateNetworkConditions` y `Emulation.setCPUThrottlingRate` se proporcionan los parámetros necesarios para la activación del *throttling* y la definición de las condiciones de la CPU del cliente. Dentro de las condiciones de red que se deben definir se encuentran: *downloadThroughput* (byte/s), *uploadThroughput* (byte/s) y *latency* (ms), que serán seleccionadas según las condiciones de contexto deseadas [103]. Cabe resaltar que la versión actual de Puppeteer no permite emular entornos de red complejos. Por tal motivo, esta primera versión del sistema de pruebas se centra en la evaluación de la calidad a partir de la emulación de variaciones del ancho de banda (cambios rápidos, cambios lentos, escalonados, etc.) y/o la latencia del enlace.

La Figura 4.4 muestra los cuatro escenarios de red o de variación de ancho de banda considerados en este estudio. Teniendo en cuenta que la frecuencia, el tipo y la ubicación de los cambios de calidad de vídeo durante una sesión de transmisión de vídeo pueden perturbar la atención visual del usuario y, por lo tanto, afectar a su QoE, los escenarios emulados representan fluctuaciones de ancho de banda persistentes y no persistentes que se corresponden en algunos casos con los *Network Presets* disponibles para la gestión de las variaciones de ancho de banda en Chrome.

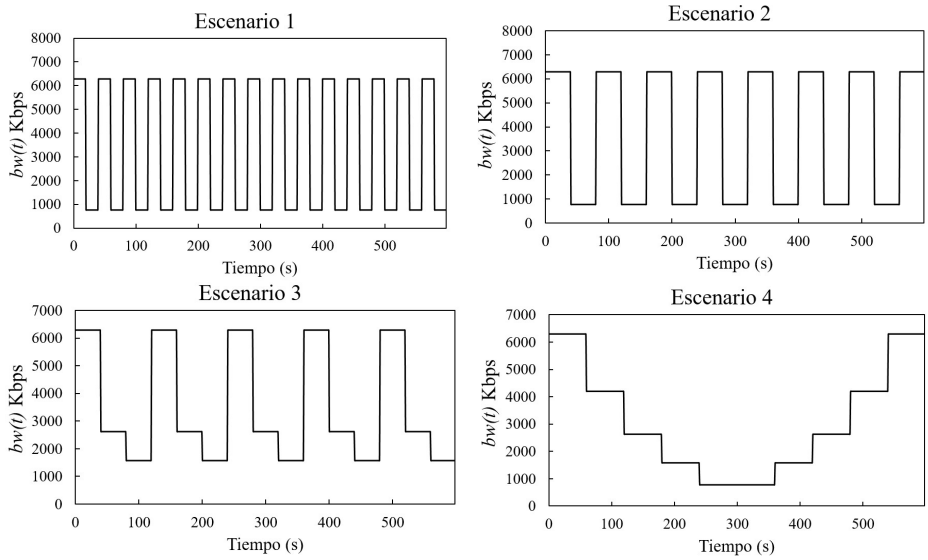


Figura 4.4. Escenarios de red o de variación de ancho de banda emulados.

4.2.3. Cliente reproductor DASH

Entre los diferentes reproductores de vídeo compatibles con DASH disponibles en la actualidad, para el desarrollo de este trabajo hemos seleccionado el reproductor Shaka Player [104]. Shaka Player es una librería de código abierto de JavaScript que permite la reproducción de contenidos multimedia tanto en formato DASH como HLS en un navegador estándar, sin requerir el uso de ningún tipo de *plugin* adicional. El reproductor debe estar alojado en un servidor web, que en nuestro caso es el ordenador local. Mediante el uso de Puppeteer, se ha desarrollado un script en node.js que permite las siguientes funciones: acceder a la web donde se encuentra el reproductor Shaka Player; seleccionar el vídeo; activar la consola de registro (donde se recogerán los datos para su post-procesado) e iniciar la reproducción del vídeo haciendo clic programáticamente en el botón correspondiente.

Todos los reproductores de vídeo para los formatos modernos de transmisión (por ejemplo, HLS y MPEG-DASH) tienen un conjunto de características comunes. Muchas de las características están sujetas a diversos compromisos entre la QoE y otros parámetros, lo que significa que a menudo es posible mejorar la QoE mediante una mejor heurística. En el caso de algunos reproductores (incluido Shaka Player), es más fácil mejorar algunas métricas de QoE a expensas de otras, ya que la heurística puede ajustarse mediante opciones de configuración en el reproductor. Las dos características más importantes en los reproductores de vídeo modernos que tienen un impacto en la QoE son:

- Selección de la tasa de bits: para elegir una tasa de bits adecuada cuando hay múltiples representaciones en diferentes calidades para un flujo de vídeo. Esta función se conoce con muchos nombres, por ejemplo, estrategia de tasa de bits adaptativa (ABR), estrategia de tasa de bits múltiple (MBR) o selección automática de tasa de bits.

- Estrategia de almacenamiento en *buffer*: para decidir la cantidad de datos multimedia que se mantendrán en el *buffer* interno del reproductor, cuándo se obtendrán los datos multimedia y cuántos datos multimedia se deben tener disponibles antes de que se inicie la reproducción.

El sistema de *buffering* de Shaka Player tiene tres parámetros: *bufferingGoal, rebufferingGoal* y *bufferBehind.* Todos se expresan en segundos. El parámetro *bufferingGoal* es la cantidad de contenido que intentamos almacenar en el *buffer*, *rebufferingGoal* es la cantidad de contenido que se precisa tener en el *buffer* antes de poder reproducir, y *bufferBehind* es la cantidad de contenido que mantenemos en el *buffer* detrás del *playhead.*

La modificación de la estrategia de *buffering* está fuera de los objetivos de este trabajo, que no pretende realizar cambios sobre el algoritmo de adaptación del reproductor. Sin embargo, se han realizado pruebas modificando los valores asociados al *bufferingGoal*, en concreto se han realizado simulaciones utilizando valores de *bufferingGoal* de 30, 20 y 10 segundos. Utilizando el valor de 10 s, establecido por defecto en el reproductor, el número de interrupciones se multiplicó en cada uno de los escenarios estudiados, pero en particular la situación fue crítica en los escenarios con una alta tasa de variación del *bitrate*. Finalmente, se seleccionó el valor de *bufferingGoal* igual a 20 segundos, que permite tener un comportamiento fluido durante el proceso de transmisión de vídeo, ya que mantiene una buena relación proporcional con el tamaño del segmento utilizado (5 s).

La Figura 4.5. muestra una instantánea de lo que el usuario vería al ejecutar el sistema y desactivar el modo *headless.* En la esquina superior izquierda de la pantalla se muestra un mensaje que indica que un software de pruebas automatizado está controlando Chrome. El marco derecho de la pantalla muestra que el acceso a DevTools está activo, lo que permite acceder a limitar el ancho de banda y a las estadísticas de la red.

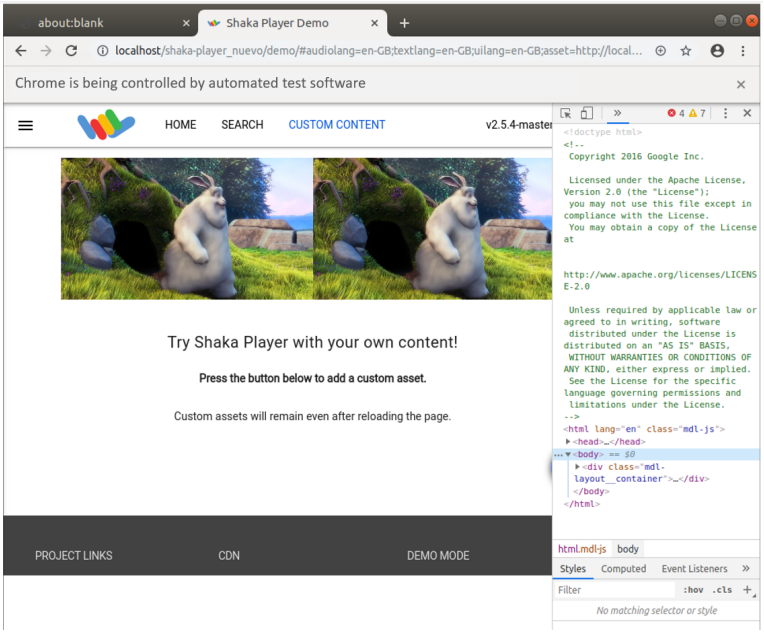


Figure 4.5. Chrome controlado por Puppeteer. Modo Headless desactivado.

4.2.4. Post-procesado y extracción de estadísticas de red

La posibilidad de acceder a través de Puppeteer a todos los recursos o herramientas de desarrollo orientados a la evaluación del rendimiento de los servicios web permite, una vez iniciada la reproducción de vídeo, la captura de métricas y registros relacionados con las estadísticas de red, nivel de *buffer*, número de interrupciones, duración de las mismas, tiempo de reproducción y versiones transmitidas, entre otros. Esta información se genera mediante métodos que permiten definir un *callback* en el que podemos especificar los elementos de la página a extraer de forma programada. Estos datos se recogen a través de la salida por consola, la generación de un fichero JSON que registra la interacción entre el cliente y el servidor o extrayendo las estadísticas y variables del reproductor a través de métodos JavaScript. Estos archivos se procesan para recuperar los datos necesarios para la reconstrucción del vídeo que se utilizará en las etapas de evaluación objetiva y subjetiva. La Figura 4.6 muestra un ejemplo de vídeo reconstruido (Q_1 corresponde a la representación de mayor calidad disponible y Q_n a la de menor calidad). En los escenarios libres de paradas, la reconstrucción se realiza concatenando los segmentos descargados según sus marcas de tiempo de reproducción. Sin embargo, en los escenarios en los que hay interrupciones, también es necesario tener la hora de inicio y la duración de las interrupciones.

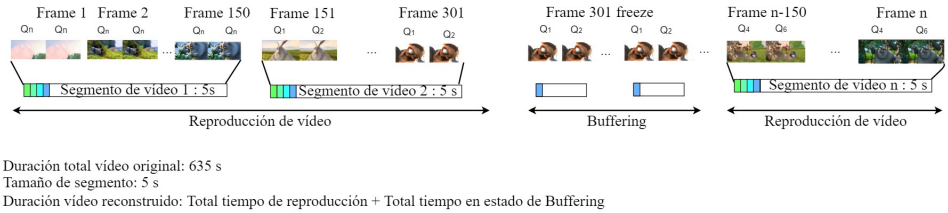


Figure 4.6. Proceso de reconstrucción del video recibido.

4.2.5. Evaluación de la calidad de experiencia (QoE)

Aunque la posibilidad de alternar entre diferentes representaciones de un vídeo reduce considerablemente el riesgo de interrupciones, las variaciones de calidad que esto conlleva pueden reducir la calidad global percibida del vídeo y resultar molestas para el usuario. Los usuarios de los servicios de transmisión de vídeo suelen tener unas expectativas diferentes a las de otros tipos de servicios web. Por ello, conocer la QoE percibida por el usuario es un factor clave para evaluar el rendimiento de los algoritmos de transmisión adaptativa. Si hay un retraso notable después de hacer clic en un enlace de vídeo, o si hay una interrupción durante la reproducción, o cualquier degradación perceptible en la calidad visual del vídeo, la QoE percibida por el usuario puede verse afectada.

En esta etapa, como se ha mencionado anteriormente, la evaluación de la calidad del vídeo recibido se realizará mediante métricas objetivas: PSNR, SSIM y VMAF. Asimismo, se usará la implementación en Python del estándar ITU-T P.1203, que se ha convertido en el primer modelo estandarizado para la evaluación de la calidad de los servicios de transmisión adaptativa de audio y vídeo. Finalmente, evaluará cómo afecta a la calidad percibida por el usuario la respuesta del algoritmo de adaptación frente a aspectos como: la frecuencia de conmutación de calidades, el rango de variación y el tipo de variación (ascendente o descendente).

Tal como se comentó en el Capítulo 2, la implementación ITU-T P.1203 consta de tres módulos, uno para la estimación de la calidad del vídeo (P.1203, Pv), otro para la estimación de la calidad del audio (P.1203, Pa) y uno más que proporciona una percepción global de la calidad (P.1203.3, Pq). Así mismo, la herramienta tiene 4 modos de operación, desde el modo 0 hasta el modo 3. Los modos se distinguen según la cantidad de información disponible, que va desde sólo los metadatos (códec, resolución, tasa de bits, tasa de *frames*, resolución del *display*, duración del segmento) en modo 0, hasta el acceso al flujo de bits completo en el modo 3.

En este trabajo nos centramos en el modo 3 y en la salida O.46, que corresponde con la evaluación integral del MOS. La recomendación ITU-T P.1203 tiene en cuenta la información sobre las interrupciones para la predicción de la calidad y proporciona un valor en una escala de 1 a 5 por cada segundo del vídeo. Como limitación tenemos que la implementación disponible actualmente de la recomendación P.1203, solo soporta vídeos codificados en H264 con una resolución de hasta 1080p.

4.3. Evaluación del rendimiento de la transmisión de vídeo 3D empleando DASH y presentación de resultados

Para los experimentos se ha utilizado la secuencia de vídeo Big Buck Bunny [77] con una resolución de 1080p y 30 fps. Esta secuencia es comúnmente utilizada [105], ya que su contenido y duración (635 s) resulta adecuada en este tipo de experimentos y permite reducir los sesgos en el análisis. El vídeo de cada vista (V0 y V1) se ha codificado utilizando FFMPEG libx264, preset "medium", una estructura GoP (*Group of Pictures*) cerrada y QP variable. Siguiendo el procedimiento para la selección de representaciones propuesto e incluyendo algunas restricciones respecto a la diferencia de ancho de banda mínimo que debe existir entre ellas, del total de 91 secuencias codificadas disponibles, se seleccionaron 19 HRC para ser segmentadas y alojadas en el servidor. La longitud de los segmentos (Sd) fue de 5 s. Una vez generados los segmentos, se ubican junto con el MPD dentro del Servidor Web para que puedan ser accedidos por el reproductor DASH. La Tabla 4.1 detalla la información sobre la conformación de cada flujo de vídeo estereoscópico.

Tabla 4.1. Representaciones disponibles en el servidor de streaming de vídeo

Tipo de representación	*HRC*	*V0*	*V1*	*Tasa de bits promedio (Kbps)*
SYM	HRC1	QP20	QP20	5544,12
ASYM	HRC2	QP20	QP22	5394,88
ASYM	HRC3	QP20	QP24	5183,74
ASYM	HRC4	QP20	QP26	4948,15
ASYM	HRC5	QP20	QP28	4698,32
ASYM	HRC6	QP20	QP30	4383,89
SYM	HRC7	QP28	QP28	3860,92
ASYM	HRC8	QP28	QP30	3546,49
ASYM	HRC9	QP28	QP32	3235,4
ASYM	HRC10	QP28	QP34	2977,57
SYM	HRC11	QP32	QP32	2610,81
ASYM	HRC12	QP32	QP34	2352,98
ASYM	HRC13	QP32	QP36	2160,4
SYM	HRC14	QP36	QP36	1708,85
ASYM	HRC15	QP36	QP38	1548,13
ASYM	HRC16	QP36	QP40	1429,78
SYM	HRC17	QP40	QP40	1148,55
ASYM	HRC18	QP40	QP42	1051,41
SYM	HRC19	QP42	QP42	954,07

Los experimentos se llevan a cabo utilizando el sistema de pruebas para la transmisión automática de vídeo desarrollado en este trabajo. El sistema propuesto se basa en el uso de Puppeteer, del

que se habló en la sección 4.2.2. Puppetteer está orientado a la automatización de pruebas funcionales en entornos web. Utilizando Puppeteer, podemos acceder automáticamente al servidor web y al contenido de vídeo, mientras emulamos las condiciones de red correspondientes en cada uno de los escenarios propuestos. Al final del proceso de simulación de la transmisión, el sistema también nos permite realizar la reconstrucción del vídeo reproducido en el cliente y realizar el análisis del rendimiento del proceso de transmisión, teniendo acceso a valores relacionados con parámetros como las marcas de tiempo de reproducción por segmento, el retardo inicial del *buffer*, el tamaño y tiempo de descarga de cada segmento, el estado del *buffer* por segundo, las representaciones solicitadas y descargadas, los estados de reproducción (*buffering*, pausa o reproducción), y los niveles de conmutación del ancho de banda, la frecuencia y la ubicación. La Figura 4.7 representa la interfaz del sistema y las diferentes opciones disponibles y que deben ser configuradas por el usuario según la información que desee extraer. En los siguientes apartados se presenta el análisis de los principales aspectos evaluados en este tipo de estudios.

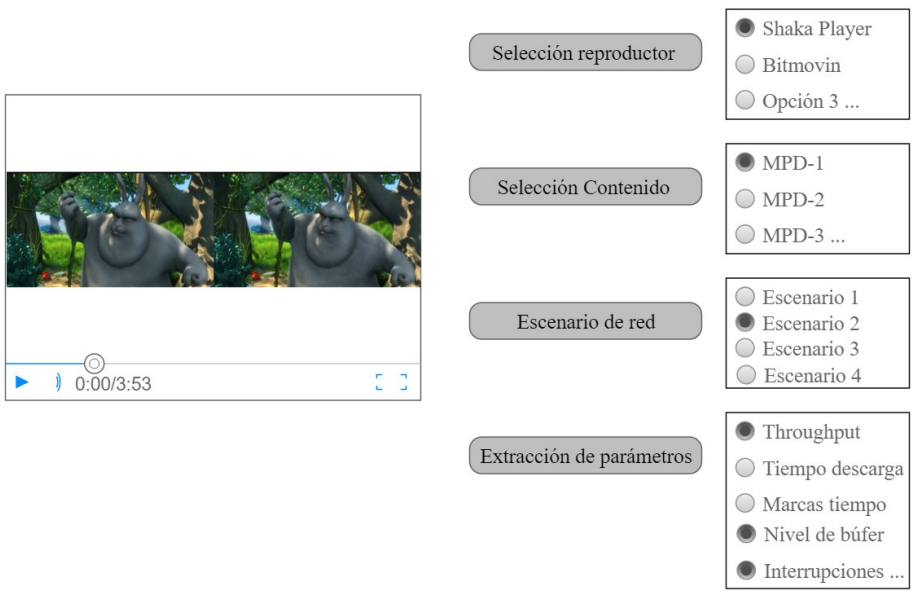


Figure 4.7. Interfaz de configuración del sistema de pruebas propuesto.

La Tabla 4.2 resume la notación y la definición de los parámetros utilizados para la evaluación del rendimiento durante el proceso de transmisión adaptativa de vídeo. Como muestran la Figura 4.8 y la Figura 4.9, se considerarán el Escenario 2 y el Escenario 4 para mostrar los resultados, siendo los más interesantes (más realistas) para extraer conclusiones. El Escenario 2 (Figura 4.8) emula las variaciones periódicas a corto plazo del ancho de banda disponible. En este contexto, los cambios a corto plazo son fluctuaciones producidas en intervalos de 40 s. Por otro lado, el Escenario 4 (Figura 4.9) produce variaciones a largo plazo del ancho de banda disponible. Este escenario presenta dos partes: desde t=0 s hasta t=300 s, el *bitrate* disminuye de 7 Mbps a 2 Mbps

cada minuto, y desde t=300 s hasta t=635 s el *bitrate* aumenta de 2 Mbps a 7 Mbps cada minuto. La latencia se ha fijado en 5 ms en todos los casos.

Tabla 4.2. Representaciones disponibles en el servidor de streaming de vídeo

Notación	Unidades	Definición
T	s	Tiempo total de la sesión
S_d	s	Duración del segmento
S	segmentos	Número total de segmentos en la sesión. $S= T/S_d$
R	nivel	Número total de representaciones en el servidor
r_i	kbps	Niveles representaciones $1<i<R$. r_R representa el nivel de calidad más alto
$bw_{available}(t)$	kbps	Ancho de banda disponible (emulado).
$bw_{availablemax}(t)$	kbps	Máximo ancho de banda disponible.
$\tau(t)$	kbps	*Throughput*, tasa de bit medida en el lado del cliente.
$b(t)$	kbps	Tasa de bits seleccionada en el lado del cliente.
T_{buf}	s	Tiempo total de *buffering*, tiempo total que la reproducción es interrumpida por infrautilización de *buffer*.
ε_{buf}		Eficiencia de *buffer*. $\varepsilon_{buf} = {T_{buf}}/{T}$

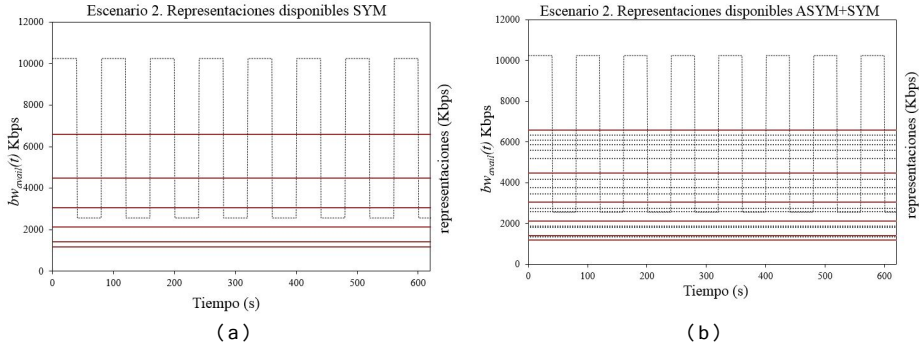


Figure 4.8. Variaciones de ancho de banda y representaciones disponibles para el Escenario 2 (a) Representaciones Simétricas (SYM). (b) Representaciones Simétricas y Asimétricas (ASYM+SYM).

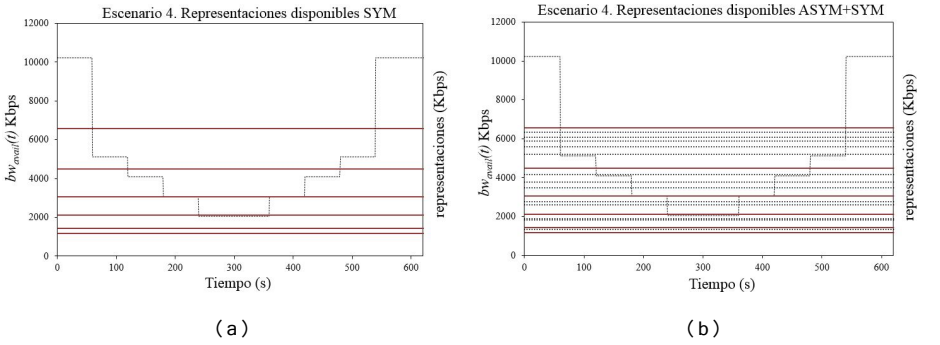


Figure 4.9. Variaciones de ancho de banda y representaciones disponibles para el Escenario 4 (a) Representaciones Simétricas (SYM). (b) Representaciones Simétricas y Asimétricas (ASYM+SYM).

Las Figuras 4.10. y 4.13. muestran el ancho de banda disponible $(bw_{available}(t))$, la tasa de bits solicitada b(t) en función de las variaciones de ancho de banda y el *throughput* $\tau(t)$ o tasa de bits medida en el lado del cliente para el Escenario 2 y el Escenario 4, respectivamente. $\tau(t)$ se obtiene calculando el *throughput* efectivo de cada segmento, que es el tamaño del segmento dividido entre el tiempo de descarga del mismo. Aunque no se puede apreciar debido a la escala de las gráficas, en todos los casos el algoritmo de adaptación descarga los dos primeros segmentos utilizando el nivel de calidad más bajo y, a continuación, cambia al nivel de calidad más alto disponible dentro de los primeros 12 segundos, lo que suele ser el caso de la mayoría de los algoritmos de adaptación. Los resultados obtenidos son coherentes con lo que se espera del algoritmo de adaptación de Shaka Player, que según la literatura emplea un mínimo de dos medias móviles ponderadas exponencialmente (EWMA). Tal como se puede ver en las Figuras 4.10 y 4.13, el algoritmo de adaptación consigue adaptarse rápidamente a las caídas de la tasa de bits, pero aumenta gradualmente la calidad cuando el ancho de banda sube. Como se puede observar en ambos escenarios, la inclusión de las representaciones asimétricas junto con las simétricas (ASYM+SYM) permite optimizar el uso del ancho de banda disponible, ya que las calidades descargadas utilizando las representaciones ASYM+SYM permiten experimentar una mejor calidad de vídeo global. Por ejemplo, en la Figura 4.13 Escenario 4, se puede observar en t=200 s que, para un ancho de banda disponible de 3000 kbps, la representación elegida utilizando SYM corresponde a una calidad de 2118 Kbps de *bitrate* frente a una calidad de 2750 Kbps de *bitrate* cuando se eligen las representaciones simétricas y asimétricas.

La Figura 4.11 y la Figura 4.14 muestran el tiempo entre peticiones y el tiempo de descarga de un segmento para los dos escenarios bajo estudio, respectivamente. Se puede apreciar cómo el comportamiento del Tiempo de Descarga del Segmento se ve afectado por el estado del *buffer* y el ancho de banda disponible, ya que su valor aumenta significativamente cuando la ocupación del *buffer* disminuye o se está produciendo una infrautilización del mismo.

En los experimentos, se muestra el comportamiento del sistema bajo severos descensos de la ocupación del *buffer*. Así, se observa en el Escenario 2 que el tiempo de descarga del segmento alcanza valores en torno a 1 s cuando el ancho de banda disponible es de 10 Mbps. También puede alcanzar hasta 14 s cuando el ancho de banda disponible disminuye bruscamente a 2,5 Mbps.

La mejora de las representaciones codificadas con ASYM+SYM sobre las representaciones codificadas con SYM es más evidente en el Escenario 4. En este caso, debido a que el reproductor tiene un mayor número de representaciones disponibles utilizando ASYM+SYM, se puede realizar un proceso de adaptación más suave, seleccionando representaciones de mayor calidad, en cada una de las transiciones de cambio de calidad requeridas como consecuencia de las variaciones de ancho de banda. La Figura 4.12 y la Figura 4.15 muestran respectivamente el comportamiento del *buffer* en función del ancho de banda disponible. Utilizando un *bufferingGoal* de 20 segundos, hemos conseguido minimizar e incluso evitar las interrupciones de la reproducción.

Los algoritmos de adaptación utilizan la ocupación del *buffer* y el *throughput* como parámetros para elegir el siguiente segmento de descarga. Se puede demostrar que las representaciones SYM+ASYM permiten reproducir segmentos de mejor calidad manteniendo un nivel de ocupación del *buffer* similar al de las representaciones SYM en la mayoría de los escenarios.

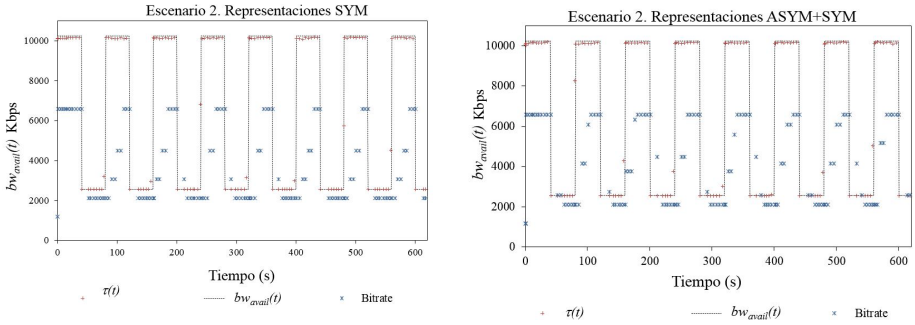


Figure 4.10. Escenario 2. *Throughput* por segmento, ancho de banda disponible y bitrate solicitado. (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

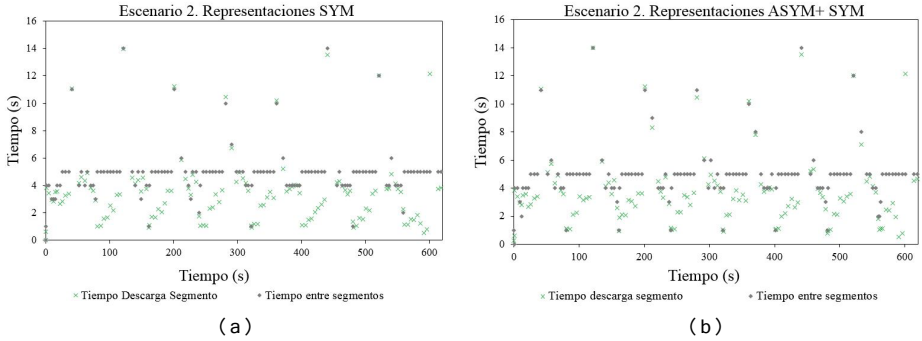


Figure 4.11. Escenario 2. Tiempo de descarga y Tiempo entre solicitud de segmentos (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

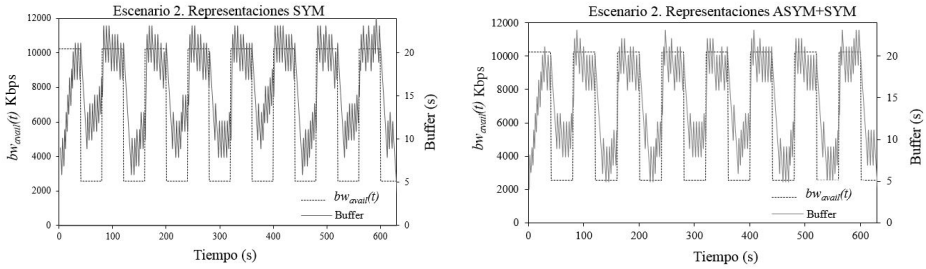


Figure 4.12. Escenario 2. Ancho de banda disponible $bw_{avail}(t)$ y estado del *buffer*. (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

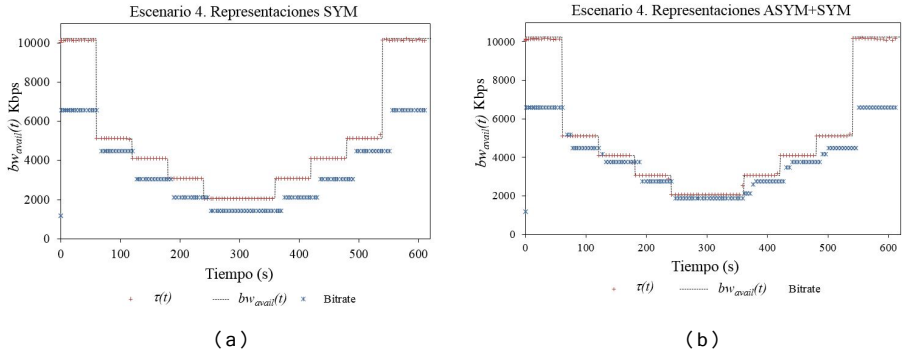


Figure 4.13. Escenario 4. *Throughput* por segmento, ancho de banda disponible y bitrate solicitado. (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

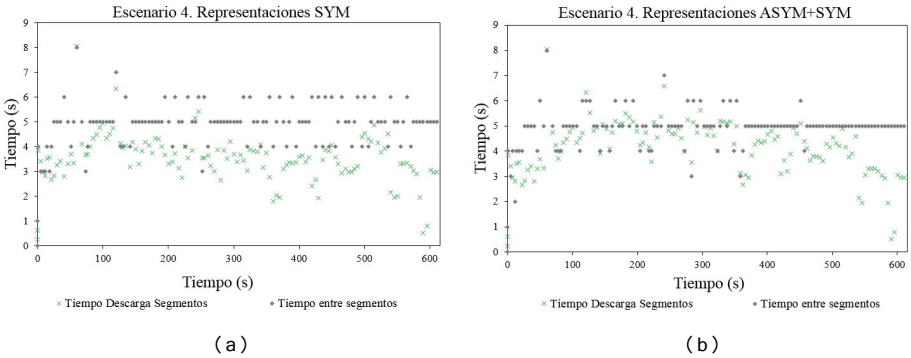


Figure 4.14. Escenario 4. Tiempo de descarga y Tiempo entre solicitud de segmentos (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

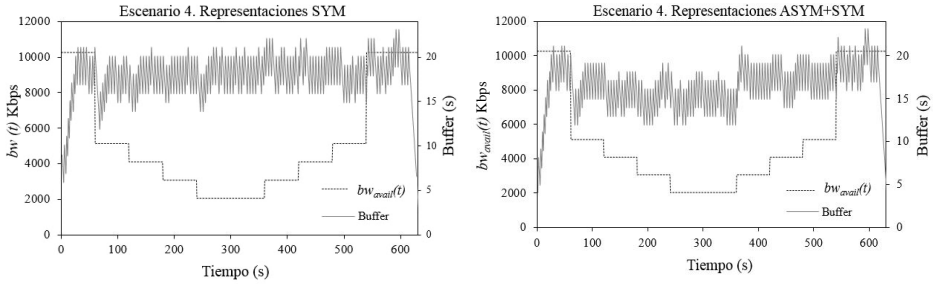


Figure 4.15. Escenario 4. Ancho de banda disponible y estado del *buffer*. (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

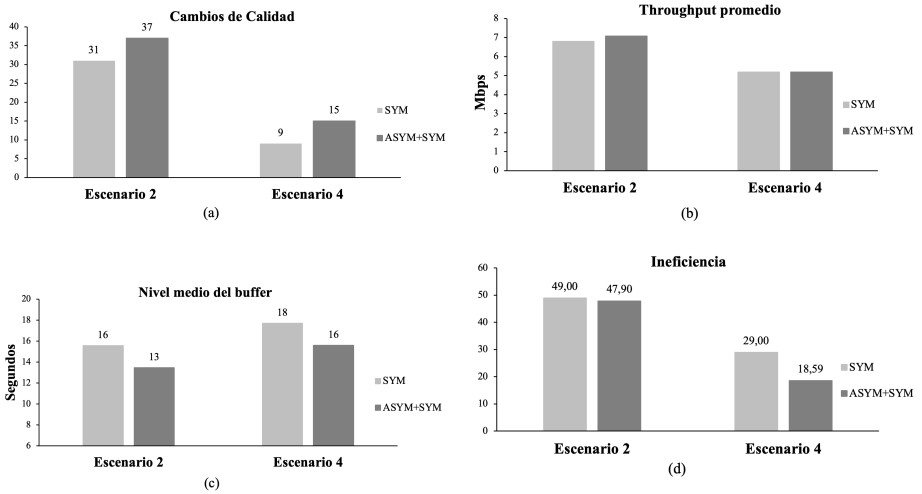


Figura 4.16. Resultados de rendimiento para Escenario 2 y Escenario 4. (a) Cambios de Calidad, (b) *Throughput* promedio, (c) Nivel medio del *buffer* (d) Ineficiencia

La Figura 4.16 ilustra algunos de los resultados de rendimiento que proporciona el sistema de pruebas, como el número de cambios de calidad, el *throughput* promedio, el nivel medio de *buffer* y la ineficiencia. Sin embargo, a partir del análisis de los datos proporcionados por el sistema pueden calcularse otros, como el número de paradas y el porcentaje de tiempo en estado de *buffer* con respecto al tiempo total de reproducción (estos casos no se muestran porque en nuestro caso presentan un valor cero). En particular, la ineficiencia se define según la ecuación (3) [106] y determina hasta qué punto el algoritmo utiliza correctamente el ancho de banda disponible en la red.

$$\boldsymbol{Inefficiency} = \sum_{\boldsymbol{t}} \frac{|\boldsymbol{b}(\boldsymbol{t}) - \boldsymbol{\tau}(\boldsymbol{t})|}{\boldsymbol{\tau}(\boldsymbol{t})} \qquad (3)$$

En cuanto a la métrica *Cambios de calidad*, podemos ver que tanto para el Escenario 2 como para el Escenario 4, se observan más cambios de calidad en el contexto de las representaciones ASYM+SYM porque hay más representaciones disponibles. Aunque en la mayoría de los algoritmos de evaluación de la calidad se castiga un mayor número de cambios de representación, el uso de una lista de representaciones con mayor granularidad utilizando vistas asimétricas, permite que las transiciones entre las diferentes tasas de bits disponibles sean más suaves, mejorando así la calidad de la experiencia del usuario. También podemos observar que el *Throughput promedio* es ligeramente superior en el contexto ASYM+SYM respecto al uso exclusivo de representaciones SYM, aunque no es una diferencia representativa en este caso.

Por otro lado, los parámetros *Nivel Medio de Buffer* e *Ineficiencia* muestran que el uso de las representaciones ASYM+SYM permite un mejor uso del ancho de banda. Esta tendencia es especialmente notable en el Escenario 4 debido al patrón de variación del ancho de banda (teniendo en cuenta la frecuencia, la duración y la magnitud del cambio en el ancho de banda disponible).

4.4. Evaluación objetiva de la calidad de vídeo

Esta sección se centra en el estudio de los resultados ofrecidos por las principales métricas de evaluación objetiva de la calidad de vídeo (VMAF, PSNR, SSIM, VIF), comparando el uso de las representaciones simétricas (SYM) con el uso de representaciones asimétricas y simétricas (ASYM+SYM) para los dos escenarios estudiados (Escenario 2 y Escenario 4).

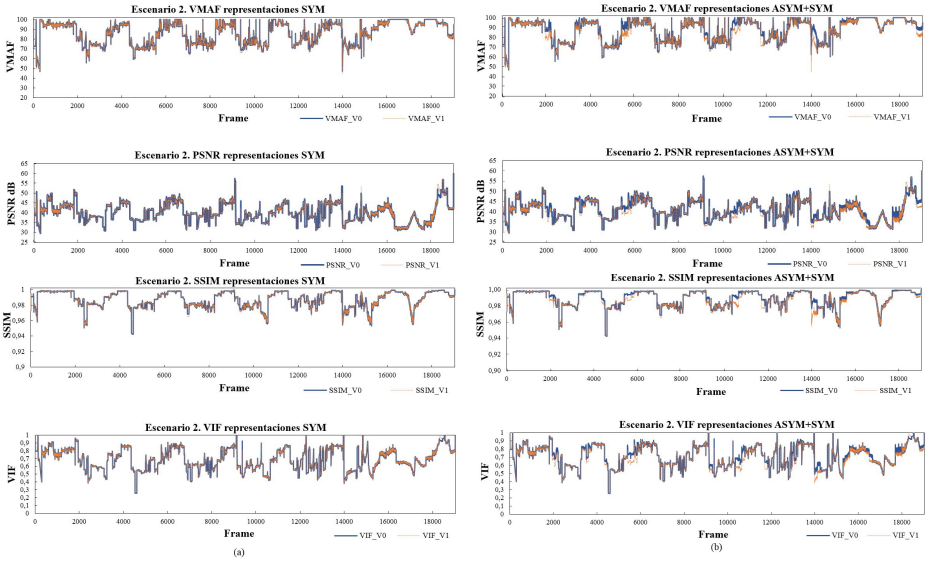


Figura 4.17. Escenario 2. VMAF, PSNR, SSIM y VIF (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

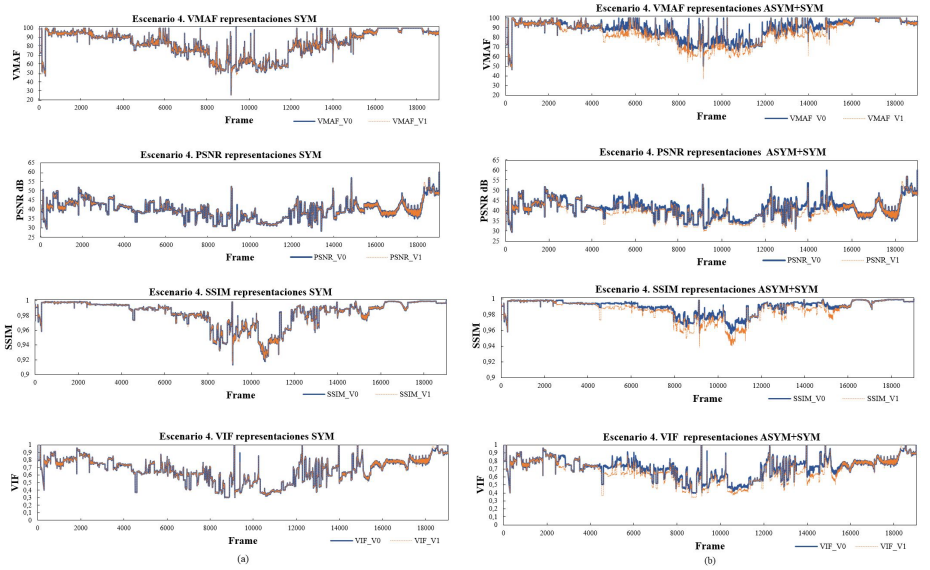


Figura 4.18. Escenario 4. VMAF, PSNR, SSIM y VIF (a) Representaciones Simétricas (SYM). (b) Representaciones Asimétricas y Simétricas (ASYM+SYM)

Como se puede observar en las Figuras 4.17 y 4.18, el comportamiento de las curvas VMAF y SSIM tienen una clara correlación con el patrón de variación del ancho de banda en los dos escenarios evaluados. Asimismo, aunque los valores obtenidos para el resto de las métricas objetivas (PSNR, VIF) parecen tener comportamientos similares en ambos casos (SYM y ASYM+SYM), si analizamos el vídeo estereoscópico reproducido encontramos que el *bitrate* medio es mayor cuando se utilizan representaciones ASYM+SYM en comparación con el uso de representaciones SYM únicamente. Analizando las cifras correspondientes al VMAF para el Escenario 2 y el Escenario 4 utilizando únicamente las representaciones SYM respecto al VMAF obtenido utilizando las representaciones ASYM+SYM, se observa que se obtiene un mejor resultado para la segunda condición tanto para el Escenario 2 como para el Escenario 4. Desde un punto de vista cuantitativo, para el Escenario 2 se obtiene un valor medio de VMAF de 85,73 cuando se utilizan sólo codificaciones SYM frente a 87,13 cuando se utilizan codificaciones ASYM+SYM. El mismo comportamiento, pero aún más evidente, se representa en la Figura 4.18 para el Escenario 4, donde tenemos un valor VMAF medio para el vídeo 3D de 82,47 cuando se utilizan sólo codificaciones SYM, frente a 86,61 cuando se utilizan codificaciones ASYM+SYM (en este caso la diferencia puede ser apreciable).

4.5. Evaluación subjetiva de la calidad del vídeo

Teniendo en cuenta los inconvenientes de la realización de pruebas subjetivas, como el número de usuarios necesarios y el tiempo para realizar las pruebas, encontramos que el requisito de un equipo específico y de condiciones de visualización controladas durante la prueba para el contexto de vídeo 3D hace inviable el uso de metodologías como el *crowdsourcing*, que habría sido una forma altamente rentable, rápida y flexible de realizar experimentos con usuarios. Por esta razón, además de la evaluación de la QoE a través de pruebas subjetivas, en este trabajo nos centramos en cómo predecir la calidad de un vídeo 3D estereoscópico a partir de la evaluación automática tanto objetiva como subjetiva de cada una de las vistas que conforman el par estereoscópico.

La evaluación subjetiva de cada una de las vistas V0 y V1 que conforman la secuencia estereoscópica se llevó a cabo con la implementación en Python de la recomendación ITu-T P.1203, tal y como se ha comentado en la sección 4.2.5. Los resultados de las simulaciones se muestran en la Tabla 4.3 y la Tabla 4.4, y en las Figuras 4.19 y 4.20.

La Tabla 4.3 muestra los resultados obtenidos para el Escenario 2 y el Escenario 4 en ambos contextos, representaciones SYM y ASYM+SYM. En primer lugar, se muestra la salida O.23, que es una indicación perceptiva del *buffering*, como una única puntuación expresada en una escala de calidad (de 1 a 5) para una sesión determinada. En este caso, O.23 tiene el valor máximo, teniendo en cuenta que no hubo paradas durante la transmisión en ninguno de los escenarios analizados. Además, también se presentan O.35 y O.46. Por un lado, O.35 da la versión integrada en el tiempo de O.34 y representa la puntuación final de la calidad de la codificación audiovisual, también dada utilizando una escala de calidad MOS (1-5). Por último, O.46 denota la puntuación de calidad global que tiene en cuenta el retardo inicial de la memoria intermedia. A partir del resultado O.35, se puede observar que las codificaciones ASYM+SYM mejoran los resultados de calidad de vídeo en comparación con el uso de sólo representaciones SYM, aunque depende del escenario. Es

decir, en el escenario 1, la mejora es de alrededor del 1% mientras que en el escenario 2 la mejora es de más del 4%.

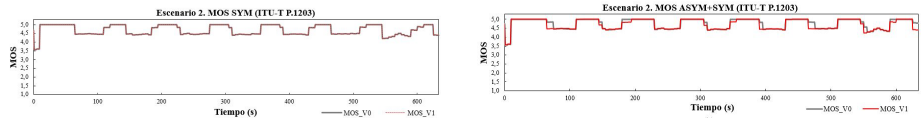
Table 4.3. ITU-T P.1203 Resultados MOS ASYM+SYM

	Salida	MOS V0	MOS V1	Promedio V0V1
Escenario 2	O.23	5,00	5,00	5,00
	O.35	4,68	4,63	4,66
	O.46	4,63	4,60	4,62
Escenario 4	O.23	5,00	5,00	5,00
	O.35	4,81	4,69	4,75
	O.46	4,74	4,64	4,69

Table 4.4. ITU-T P.1203 Resultados MOS SYM

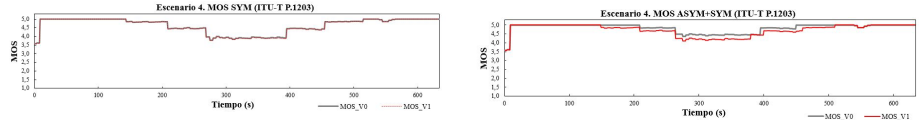
	Salida	MOS V0	MOS V1	Promedio V0V1
Escenario 2	O.23	5,00	5,00	5,00
	O.35	4,60	4,60	4,60
	O.46	4,58	4,58	4,58
Escenario 4	O.23	5,00	5,00	5,00
	O.35	4,56	4,56	4,56
	O.46	4,52	4,52	4,52

Por otro lado, las Figuras 4.19. y 4.20. representan el comportamiento de la salida O.34, que corresponde con la calidad audiovisual por intervalo de muestreo de salida. Sólo se está considerando el vídeo, ya que el audio ha sido eliminado de la secuencia original y está fuera del alcance de este estudio. Se puede observar en las figuras como el parámetro MOS disminuye, lo que significa una menor QoE, cuando el reproductor descarga versiones de menor calidad para contrarrestar la disminución del ancho de banda y así evitar las interrupciones.



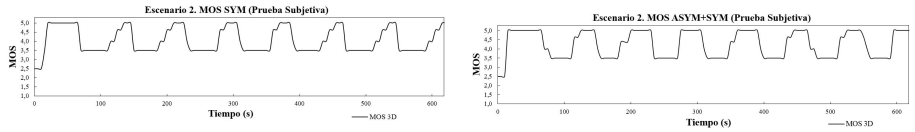
(a) (b)

Figure 4.19. ITU-T P.1203 O.34 salida Escenario 2. (a) MOS por segundo usando representaciones SYM. (b) MOS por segundo usando representaciones ASYM+ SYM)



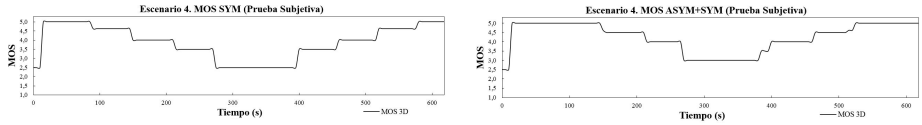
(a) (b)

Figure 4.20. ITU-T P.1203 O.34 salida Escenario 4. (a) MOS por segundo usando representaciones SYM. (b) MOS por segundo usando representaciones ASYM+ SYM.



(a) (b)

Figure 4.21. MOS prueba subjetiva con usuarios Escenario 2. MOS por Segundo representaciones SYM. (b) MOS por Segundo representaciones ASYM+SYM.



(a) (b)

Figure 4.22. MOS prueba subjetiva con usuarios Escenario 4. MOS por Segundo representaciones SYM. (b) MOS por segundo representaciones ASYM+SYM.

Los estudios existentes (por ejemplo, [107] y [73]) sugieren que el promedio de la calidad de las vistas 2D izquierda y derecha predice bien la calidad de los vídeos estereoscópicos con distorsión simétrica, pero genera un sesgo de predicción sustancial cuando se aplica a los vídeos estereoscópicos con distorsión asimétrica. De acuerdo con esto, se llevó a cabo una evaluación subjetiva basada en la recomendación ITU-T P.915 [33], utilizando el método DS (*Double-Stimulus Quality Rating*) con referencia explícita usando una escala de calidad MOS de 5 puntos, de acuerdo a lo descrito en el Capítulo 2.

En las pruebas subjetivas de calidad, se evaluaron 10 segundos de cada una de las 19 secuencias de vídeo estéreo (SYM y ASYM) disponibles en el servidor de transmisión DASH. Durante la sesión de evaluación, las secuencias se mostraron en un orden aleatorio y, antes de cada secuencia de prueba, se mostró un vídeo correspondiente al vídeo original sin ninguna degradación. Cada secuencia de prueba se evaluó individualmente justo después de ser presentada al evaluador, que disponía de 5 s para evaluar la QoE del vídeo 3D en términos de calidad visual dando una puntuación basada en la escala de calidad MOS de cinco niveles. Los evaluadores vieron cada secuencia de vídeo utilizando el sistema de visión 3D de NVIDIA y un monitor LCD de 17 pulgadas con una frecuencia de refresco de 120. La Figura 4.21 y la Figura 4.22 muestran los valores MOS obtenidos en las pruebas subjetivas para el Escenario 2 y el Escenario 4, respectivamente. Con estos resultados, y en consonancia con los resultados de la evaluación objetiva, a partir de los valores de MOS para la secuencia 2D (vista izquierda, vista derecha) proporcionados por la implementación de la recomendación ITU-T P.1203, se obtiene una muy buena predicción del MOS de la secuencia estereoscópica simétrica. Sin embargo, en el caso de las representaciones asimétricas, la predicción de los valores MOS de las vistas separadas no es sencilla y se requiere un estudio más profundo.

4.6. Conclusiones

Referente a la evaluación de la QoE en un escenario de transmisión adaptativa de vídeo 3D, podemos decir lo siguiente:

La evaluación de la QoE 3D en sistemas DASH incluye muchos aspectos que, en general, no hacen viable la replicación de los experimentos y complica la posibilidad de comparar soluciones o mejoras. Es preciso disponer de una herramienta de pruebas versátil, que permita automatizar, escalar, replicar y simplificar el proceso de evaluación. El sistema desarrollado en este trabajo comprende desde la descripción del proceso de codificación hasta la evaluación subjetiva utilizando la recomendación UIT-T P.1203. La implementación proporcionada de la recomendación UIT-T P.1203 es una buena primera aproximación para la predicción de la calidad de los vídeos estereoscópicos simétricamente distorsionados. Sin embargo, genera un importante sesgo de predicción cuando se aplica a vídeos estereoscópicos distorsionados asimétricamente, como se demuestra al comparar los resultados con los obtenidos mediante la evaluación subjetiva de los usuarios.

Para el proceso de emulación de la red y la automatización de las pruebas, se ha utilizado la herramienta Google Puppeteer y las opciones que ofrece Chrome DevTools, que han resultado útiles para gestionar el cliente web y el reproductor. Con el uso de la tecnología Docker, se pretende incluir todas las herramientas implementadas en contenedores. Esto facilitará la replicación del sistema por parte de desarrolladores e investigadores. El sistema evolucionará y se ampliará a medida que se incorporen nuevos codificadores (como AV1, etc.) tanto en el proceso de codificación como en la sección de reproducción y evaluación según el estándar, así como que se desarrollen nuevos reproductores o se propongan diferentes algoritmos de adaptación. Sin embargo, todas estas nuevas propuestas no implican cambios en todos los módulos, ya que la adaptación puede realizarse de forma independiente cuando sea necesario, siguiendo la evolución de la tecnología de transmisión de vídeo. De hecho, la principal ventaja de este desarrollo modular es que cualquier investigador pueda desarrollar modificaciones y mejoras tanto a nivel de nuevos algoritmos de adaptación como con la inclusión de nuevos reproductores, o en cualquier otro módulo, enriqueciendo así el desarrollo presentado en esta tesis en particular y, a su vez en general, mejorando una herramienta que puede ser de gran utilidad para el resto de la comunidad científica.

Referencias

[1] A. Television and S. Committee, "3D-TV Terrestrial Broadcasting , Part 1-6," 2015. [Online]. Available: https://www.atsc.org/atsc-documents/a104-atsc-3d-tv-terrestrial-broadcasting/. [Accessed: 02-Feb-2021].

[2] Intenational Telecommunication Union. Report ITU-R BT.2160-3, "Features of three-dimensional television video systems for broadcasting," 2012.

[3] L. Lucas, C. Loscos, and Y. Remion, *3D Video: From Capture to Diffusion*. Wiley, 2013.

[4] A. Vetro, A. M. Tourapis, K. Muller, and T. Chen, "3D-TV Content Storage and Transmission," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 384–394, Jun. 2011.

[5] C. G. Gürler, B. Gorkemli, G. Saygili, and a. M. Tekalp, "Flexible transport of 3-d video over networks," *Proc. IEEE*, vol. 99, no. 4, pp. 694–707, 2011.

[6] A. Vetro, "Frame compatible formats for 3D video distribution," *Proc. - Int. Conf. Image Process. ICIP*, no. October 2010, pp. 2405–2408, 2010.

[7] R. E. Patterson, *Human factors of stereoscopic 3D displays*. Springer US, 2015.

[8] K. Brunnström *et al.*, "Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad," *Eur. Netw. Qual. Exp. Multimed. Syst. Serv. (COST Action IC 1003)*, no. March, p. 26, 2013.

[9] M. Urvoy, M. Barkowsky, and P. Le Callet, "How visual fatigue and discomfort impact 3D-TV quality of experience: A comprehensive review of technological, psychophysical, and psychological factors," *Ann. des Telecommun. Telecommun.*, vol. 68, no. 11–12, pp. 641–655, 2013.

[10] M. T. M. Lambooij, W. A. IJsselsteijn, M. F. Fortuin, and I. E. J. Heynderickx, "Visual discomfort and visual fatigue in stereoscopic displays : a review," *J. Imaging Sci. Technol.*, vol. 53, no. 3, pp. 1–14, 2009.

[11] W. J. Tam *et al.*, "Stereoscopic 3D-TV : Visual Comfort," IEEE Transactions on Broadcasting, (2011), 335-346, 57(2). July, 2011.

[12] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, "Depth cues in human visual perception and their realization in 3D displays," no. September 2017, p. 76900B, 2010.

[13] B. G. Cumming and G. C. Deangelis, "THE PHYSIOLOGY OF STEREOPSIS," 2001.

[14] S. Pastoor and M. Wijpking, "DISPLAYS ELSEVIER Displays 17 (1997) 100-l IO 3-D displays: A review of current technologies."

[15] H. Urey, K. V. Chellappan, E. Erden, and P. Surman, "State of the art in stereoscopic and autostereoscopic displays," in *Proceedings of the IEEE*, 2011, vol. 99, no. 4, pp. 540–555.

[16] C. de Estudio and del Uit-t, "UIT-T Rec. E.800 (09/2008) Definiciones de términos relativos a la

calidad de servicio," 2008.

[17] *Qualinet White Paper on Definitions of Quality of Experience: Output Version of the Dagstuhl Seminar 12181*. 2012.

[18] U. Reiter *et al.*, "Factors Influencing Quality of Experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Cham: Springer International Publishing, 2014, pp. 55–72.

[19] K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, "Quality of Experience for Streaming Services: Measurements, Challenges and Insights," *IEEE Access*, vol. 8, pp. 13341–13361, 2020.

[20] T. Zhao, Q. Liu, and C. W. Chen, "QoE in Video Transmission: A User Experience-Driven Strategy," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 1, pp. 285–302, 2017.

[21] E. Dumi and S. Grgi, "3D video subjective quality : a new database and grade comparison study," 2016.

[22] P. Aflaki, M. M. Hannuksela, and M. Gabbouj, "Subjective quality assessment of asymmetric stereoscopic 3D video," *Signal Image Video Process.*, vol. 9, no. 2, pp. 331–345, 2015.

[23] T. Tian, X. Jiang, and X. Du, "Subjective quality assessment of compressed 3D video," in *2014 7th International Congress on Image and Signal Processing*, 2014, pp. 606–611.

[24] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[25] F. Lewandowski, M. Paluszkiewicz, T. Grajek, and K. Wegner, "Subjective quality assessment methodology for 3D video compression technology," in *2012 International Conference on Signals and Electronic Systems (ICSES)*, 2012, pp. 1–5.

[26] Q. Chen, Yanjiao ; Wu, Kaishun ; Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Commun. Surv. Tutorials*, vol. 17(2), pp. 1126–1165, 2015.

[27] A. Hore and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[29] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 3, no. 2, pp. 430–444, 2004.

[30] ITU-R, "ITU-R. BT.500-13 Methodology for the subjective assessment of the quality of television pictures," vol. 13, 2012.

[31] ITU-R, "ITU-R. BT.2021-1. Subjective methods for the assessment of stereoscopic 3DTV systems," vol. 1, 2015.

[32] ITU-T, "ITU-T P914. Display requirements for 3D video quality assessment," 2016.

[33] ITU-T, "ITU-T. P915. Subjective assessment methods for 3D video quality," 2016.

[34] ITU-T, "ITU-T. P916. Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video," 2016.

[35] T. Kawano, K. Yamagishi, and T. Hayashi, "Performance comparison of subjective assessment methods for 3D video quality," in *2012 4th International Workshop on Quality of Multimedia Experience, QoMEX 2012*, 2012, pp. 218–223.

[36] J. Kuze and K. Ukai, "Subjective evaluation of visual fatigue caused by motion images," *Displays*, vol. 29, no. 2, pp. 159–166, Mar. 2008.

[37] ITU Telecommunication Standardization Sector, "ITU-T Rec P 1203 Models and tools for quality assessment of streamed media," 2017.

[38] A. Raake, M. N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable

video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," *2017 9th Int. Conf. Qual. Multimed. Exp. QoMEX 2017*, vol. 12, pp. 0–5, 2017.

[39] P. Merkle, K. Müller, and T. Wiegand, "3D video: acquisition, coding, and display," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 946–950, May 2010.

[40] A. Vetro, "Representation and coding formats for stereo and multiview video," *Stud. Comput. Intell.*, vol. 280, pp. 51–73, 2010.

[41] A. Vetro *et al.*, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[42] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-View Video Plus Depth Representation and Coding," in *2007 IEEE International Conference on Image Processing*, 2007, vol. 1, pp. I-201-I–204.

[43] W. Mathias, *Eficiency Video Coding. Coding Tools and Specification*. Springer Berlin Heidelberg, 2015.

[44] UIT-T, "H.120 (03/93) Códecs para videoconferencia con transmisión de grupo digital primario," 1993.

[45] UIT-T, "H.261 (03/93) Códec de vídeo para servicios audiovisuales a p × 64 kbit/s," 1993.

[46] ISO/IEC, "ISO/IEC 13818-2:2013 Information technology — Generic coding of moving pictures and associated audio information — Part 2: Video," 2013.

[47] ITU-T, "H.262 : Information technology - Generic coding of moving pictures and associated audio information: Video."

[48] T. Wiegand and G. J. Sullivan, "The H.264/AVC Video Coding Standard," no. August 1999, pp. 148–153, 2007.

[49] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[50] G. Tech, Y. Chen, K. Muller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 35–49, 2016.

[51] G. J. Sullivan *et al.*, "Standardized Extensions of High Ef fi ciency Video Coding (HEVC)," vol. 7, no. 6, pp. 1001–1016, 2013.

[52] J. Bankoski, P. Wilkins, and Y. Xu, "Technical overview of VP8, an open source video codec for the web," *Proc. - IEEE Int. Conf. Multimed. Expo*, no. July, 2011.

[53] D. Mukherjee *et al.*, "A technical overview of VP9-the latest open-source video codec," *SMPTE Motion Imaging J.*, vol. 124, no. 1, pp. 44–54, 2015.

[54] "WebM: an open web media project." [Online]. Available: https://www.webmproject.org/. [Accessed: 18-Jun-2021].

[55] J. Han *et al.*, "A Technical Overview of AV1," *Proc. IEEE*, 2021.

[56] B. Bross *et al.*, "Overview of the Versatile Video Coding (VVC) Standard and its Applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, 2021.

[57] G. B. and A. L. T. Wiegand, G. J. Sullivan, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576.

[58] ITU-T and ISO/IEC, "Advanced video coding for generic audiovisual services, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC)," 2019.

[59] T. Wiegand and G. . Sullivan, "H.264/AVC video coding standard," *IEEE Signal Process. Mag.*, pp. 148–153, 2007.

[60] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 786015, Jan.

2009.

[61] ITU-T, "High Efficiency Video Coding. Recommendation ITU-T H.265," *ITU Telecommun. Stand. Sect.*, vol. 265, p. 712, 2019.

[62] ISO/IEC, "ISO/IEC 23008-2:2020 Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 2: High efficiency video coding," 2020.

[63] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[64] P. Akyazi and T. Ebrahimi, "Comparison of Compression Efficiency between HEVC/H.265, VP9 and AV1 based on Subjective Quality Assessments," in *2018 10th International Conference on Quality of Multimedia Experience, QoMEX 2018*, 2018.

[65] K. Muller *et al.*, "3D High-Efficiency Video Coding for Multi-View Video and Depth Data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[66] L. Zhang, G. Tech, K. Wegner, and S. Yea, "Test Model 6 of 3D-HEVC and MV-HEVC," 2013.

[67] A. A. Mazhar, "Performance Evaluation of h.265/mpeg-hevc, vp9 and h.264/mpeg-avc Video Coding," *Int. J. Multimed. Its Appl.*, vol. 8, no. 1, pp. 35–44, 2016.

[68] V. Baroncini, J. Ohm, and G. J. Sullivan, "Report of Subjective Test Results of Responses to the Joint Call for Proposals on Video Coding Technology for High Efficiency Video Coding (HEVC)," 2010.

[69] V. Baroncini, K. Muller, and S. Shimizu, "MV-HEVC Verification Test Report. JCT3V-N1001," 2016.

[70] K. Wang *et al.*, "Subjective evaluation of HDTV stereoscopic videos in IPTV scenarios using absolute category rating," *Stereosc. Displays Appl. XXII*, vol. 7863, p. 78631T, 2011.

[71] ITU-T, "ITU-T.P910. Subjective video quality assessment methods for multimedia applications," 2008.

[72] G. Saygili, C. G. Gurler, and A. M. Tekalp, "Evaluation of Asymmetric Stereo Video Coding and Rate Scaling for Adaptive 3D Video Streaming," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 593–601, 2011.

[73] F. Battisti, M. Carli, P. Le Callet, and P. Paudyal, "Toward the assessment of quality of experience for asymmetric encoding in immersive media," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 392–406, 2018.

[74] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived Quality of Compressed Stereoscopic Images: Effects of Symmetric and Asymmetric JPEG Coding and Camera Separation," *ACM Trans. Appl. Percept.*, vol. 3, no. 2, pp. 95–109, 2006.

[75] M. Urvoy *et al.*, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 109–114.

[76] E. Cheng, P. Burton, J. Burton, A. Joseski, and I. Burnett, "RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database," *2012 4th Int. Work. Qual. Multimed. Exp. QoMEX 2012*, pp. 212–217, 2012.

[77] BlenderFoundation, "Big Buck Bunny," 2008. [Online]. Available: https://peach.blender.org/. [Accessed: 01-Dec-2018].

[78] "H.264/AVC multi-view coding (MVC) extension JMVC reference software." [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/jmvc.

[79] "H.264/AVC Reference Software (JM 19.0)." [Online]. Available: https://github.com/RenfengLiu/JM-19.0.

[80] "HEVC Reference Software," 2011. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/.

[81] A. Vetro and Tian, "Analysis of 3D and Multiview Extensions of the Emerging HEVC Standard," 2012.

[82] G. Saygili, C. G. Gurler, and A. M. Tekalp, "Evaluation of Asymmetric Stereo Video Coding and Rate

Scaling for Adaptive 3D Video Streaming," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 593–601, Jun. 2011.

[83] W. Robitza *et al.*, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P. 1203: Open Databases and Software," *Proc. 9th ACM Multimed. Syst. Conf.*, pp. 466–471, 2018.

[84] J. D. C. and D. R. Anne Aaron, Zhi Li, Megha Manohara, "Per-Title Encode Optimization," 2015. .

[85] R. Sand, "3-DTV-a review of recent and current developments," in *IEE Colloquium on Stereoscopic Television*, 1992, pp. 1/1-1/4.

[86] G. B. Akar, A. M. Tekalp, C. Fehn, and M. R. Civanlar, "Transport Methods in 3DTV—A Survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1622–1630, Nov. 2007.

[87] T. Schierl and S. Narasimhan, "Transport and storage systems for 3-D video using MPEG-2 systems, RTP, and ISO file format," *Proc. IEEE*, vol. 99, no. 4, pp. 671–683, 2011.

[88] ISO/IEC, "ISO/IEC 23009-1 Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats Technologies, 2019," 2019.

[89] T. Su, A. Javadtalab, A. Yassine, and S. Shirmohammadi, "A DASH-based 3D multi-view video rate control system A DASH-Based 3D Multi-view Video Rate Control System," 2014.

[90] D. K. Krishnappa, D. Bhat, and M. Zink, "DASHing YouTube: An analysis of using DASH in YouTube video service," in *38th Annual IEEE Conference on Local Computer Networks*, 2013, pp. 407–415.

[91] J. Martin, Y. Fu, N. Wourms, and T. Shaw, "Characterizing Netflix bandwidth consumption," *2013 IEEE 10th Consum. Commun. Netw. Conf. CCNC 2013*, pp. 230–235, 2013.

[92] H. K. Yarnagula, S. Luhadia, S. Datta, and V. Tamarapalli, "Quality of Experience Assessment of Rate Adaptation Algorithms in DASH : An Experimental Study," pp. 1–8, 2016.

[93] Jingteng Xue, Dong-Qing Zhang, Heather Yu, and Chang Wen Chen, "Assessing quality of experience for adaptive HTTP video streaming," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[94] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over HTTP," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 1, pp. 562–585, Jan. 2019.

[95] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," in *Proceedings of the second annual ACM conference on Multimedia systems - MMSys '11*, 2011, p. 157.

[96] C. Timmerer, M. Maiero, and B. Rainer, "Which Adaptation Logic ? An Objective and Subjective Performance Evaluation of HTTP-based Adaptive Media Streaming Systems."

[97] P. Juluri, V. Tamarapalli, and D. Medhi, "SARA: Segment aware rate adaptation algorithm for dynamic adaptive streaming over HTTP," in *2015 IEEE International Conference on Communication Workshop, ICCW 2015*, 2015.

[98] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 1, pp. 469–492, Jan. 2015.

[99] A. Zabrovskiy, E. Kuzmin, E. Petrov, C. Timmerer, and C. Mueller, "AdViSE: Adaptive Video Streaming Evaluation Framework for the Automated Testing of Media Players," *Proc. 8th ACM Multimed. Syst. Conf. - MMSys'17*, pp. 217–220, 2017.

[100] Z. L. A. et al Li, "Toward A Practical Perceptual Video Quality Metric," *Netflix TechBlog*, 2016. .

[101] R. I.-T. J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference.," 2008.

[102] "Puppeteer." [Online]. Available: https://github.com/GoogleChrome/puppeteer. [Accessed: 01-May-2022].

[103] P. G. Castillo, P. A. Vila, and J. C. G. Cebollada, "Automatic QoE evaluation of DASH streaming using

ITU-T standard P.1203 and google puppeteer," *PE-WASUN 2019 - Proc. 16th ACM Int. Symp. Perform. Eval. Wirel. Ad Hoc, Sensor, Ubiquitous Networks*, pp. 79–86, 2019.

[104] "Shaka Player Demo." [Online]. Available: https://shaka-player-demo.appspot.com. [Accessed: 01-Apr-2022].

[105] S. Lederer, C. Müller, and C. Timmerer, "Dynamic Adaptive Streaming over HTTP Dataset," pp. 89–94, 2012.

[106] N. Barman and M. G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming-A Survey and Open Challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[107] J. Wang, S. Wang, and Z. Wang, "Asymmetrically Compressed Stereoscopic 3D Videos: Quality Assessment and Rate-Distortion Performance Evaluation," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1330–1343, 2017.

Apéndice A

Listado de publicaciones

Las publicaciones generadas como resultado de este trabajo de investigación, se listan a continuación.

Publicaciones en revistas

1. P. Guzmán, P. Arce, and J. C. Guerri, "Automatic QoE evaluation for asymmetric encoding of 3D videos for DASH streaming service," Ad Hoc Networks, vol. 106, article 102184, 2020.
2. S. González, W. Castellanos, P. Guzmán, P. Arce, and J. C. Guerri, "Simulation and Experimental Testbed for adaptive video streaming in ad hoc networks," Ad Hoc Networks, vol. 52, pp. 89-105, 2016.

Publicaciones en congresos

3. P. Guzmán, P. Arce, and J. C. Guerri, "Automatic QoE Evaluation of DASH Streaming using ITU-T Standard P.1203 and Google Puppeteer," in Proc. of Int. Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks (PE-WASUN), Miami Beach, FL (USA), Nov. 2019, pp. 79-86.
4. P. Guzmán, P. Arce, and J. C. Guerri, "Evaluación automática de la QoE del streaming DASH utilizando el estándar ITU-T P.1203 y Google Puppeteer," in Proc. of Jornadas de Ingeniería Telemática (JITEL), Zaragoza (Spain), oct. 2019.
5. P. Guzmán, P. Arce, and J. C. Guerri, "Evaluación de un sistema DASH para el streaming de vídeo 3D," in *Proc. of Jornadas de Ingeniería Telemática (JITEL)*, Valencia (Spain), Sep. 2017, pp. 224-228.
6. W. Castellanos, P. Guzmán, P. Arce, and J. C. Guerri, "Mechanisms for improving the scalable video streaming in mobile ad hoc networks," in Proc. of ACM Int. Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks (PE-WASUN), Cancun (Mexico), Nov. 2015, pp. 33-40.

7. P. Arce, I. de Fez, F. Fraile, S. González, P. Guzmán, and J. C. Guerri, “QoE en redes adhoc, descarga adaptativa de contenidos y vídeo 3D,” in Proc. of Jornadas de Ingeniería Telemática (JITEL), Palma de Mallorca (Spain), oct. 2015, pp. 339-346.

Anexo 1

Test fatiga y molestias visuales

A continuación, se presentan los cuestionarios aplicados a los usuarios antes y después de la realización de las pruebas subjetivas. El objetivo es conocer si los sujetos presentan síntomas de fatiga o molestias visuales previos a la realización de la prueba de evaluación subjetiva, y si estos síntomas se mantienen, incrementan, desaparecen o aparecen otros nuevos, una vez finalizada la evaluación.

En primer lugar, se presenta la información relativa al denominado "Test preliminar de fatiga y molestias visuales" y posteriormente la correspondiente al "Test de fatiga y molestias visuales".

Como referencia para la elaboración de la prueba se empleó la Recomendación ITU-T P.916 - Información y directrices para evaluar y minimizar la incomodidad y la fatiga visuales producidas por el vídeo 3D, presentada en la sección 2.4.3.b de este documento.

La prueba fue aplicada sobre una muestra de 37 sujetos (29 hombres – 8 mujeres), con edades entre los 20 y 48 años, con una media de edad de 27 años. Un 59% de los sujetos usa gafas para corregir problemas de miopía o astigmatismo.

Nombre: Edad:

Hombre ☐ Mujer ☐

Gafas Sí ☐ No ☐ Observaciones: _____

Test preliminar fatiga y molestias visuales

¿Estás parpadeando tus ojos con más frecuencia de lo habitual hoy?

No	Ligeramente	Un poco	En mayor medida	Mucho más

¿Tienes dificultades con tu visión de cerca hoy?

No	Ligeramente	Un poco	En mayor medida	Mucho

¿Tienes dificultades con la visión nocturna?

No	Ligeramente	Un poco	En mayor medida	Mucho

¿Sientes dolor en la parte posterior de tu cabeza?

No	Ligeramente	Un poco	En mayor medida	Mucho

¿Tienes problemas para enfocar?

No	Ligeramente	Un poco	En mayor medida	Mucho

¿Tiene dificultades para conducir de noche?

No	Ligeramente	Un poco	En mayor medida	Mucho más

¿Estás preocupado por tu visión?

Nunca	Rara vez	A veces	A menudo	Siempre

¿Te están llorando los ojos?

No	Ligeramente	Un poco	En mayor medida	Mucho

En este momento, la visión de tus dos ojos es:

Excelente	Muy buena	Buena	Pasable	Pobre

¿Tienes los ojos secos?

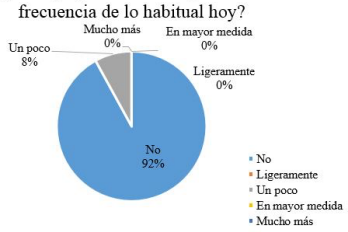
No	Ligeramente	Un poco	En mayor medida	Mucho

¿Tienes rigidez en el cuello?

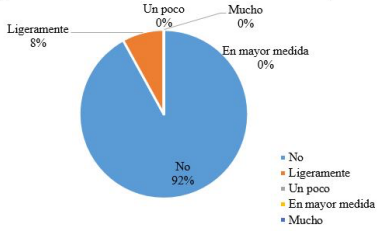
No	Ligeramente	Un poco	En mayor medida	Mucho

Resultados Test preliminar fatiga y molestias visuales

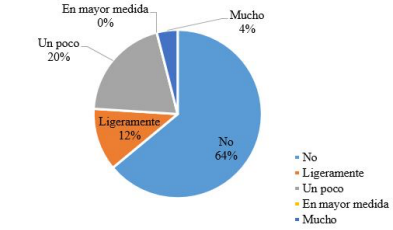
1. ¿Estás parpadeando tus ojos con más frecuencia de lo habitual hoy?



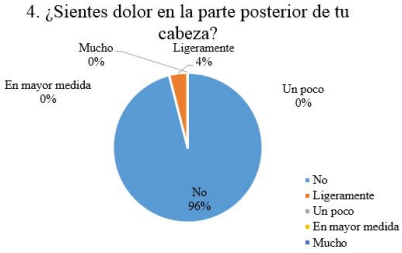
2. ¿Tienes dificultad con tu visión de cerca hoy?



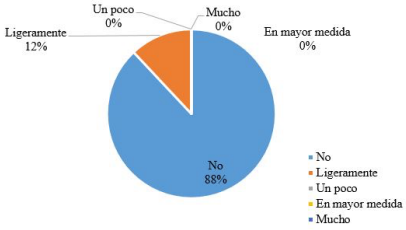
3. ¿Tienes dificultades con la visión nocturna?



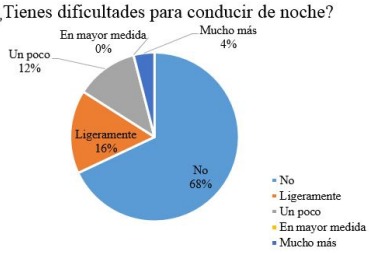
4. ¿Sientes dolor en la parte posterior de tu cabeza?



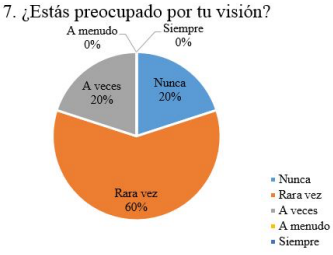
5. ¿Tienes problemas para enfocar?



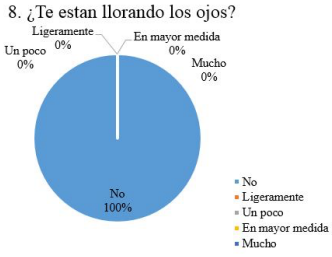
6. ¿Tienes dificultades para conducir de noche?



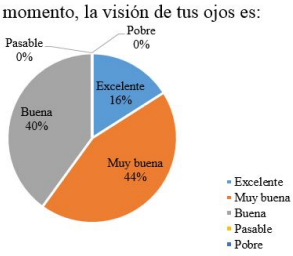
7. ¿Estás preocupado por tu visión?



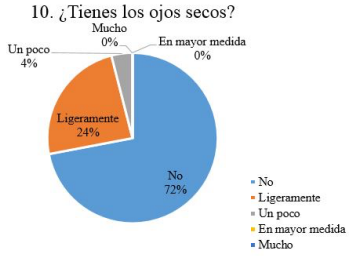
8. ¿Te estan llorando los ojos?



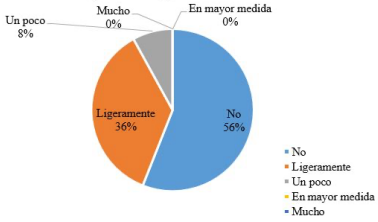
9. En este momento, la visión de tus ojos es:



10. ¿Tienes los ojos secos?



11. ¿Tienes rigidez en el cuello?



Nombre: Edad:

Hombre ☐ Mujer ☐

Gafas Sí ☐ No ☐ Observaciones: _____

Test fatiga y molestias visuales

¿Tienes la sensación de que el movimiento de tus ojos está desacoplado?

No	Si

Con respecto al comienzo del experimento, ¿experimentas rigidez de tu cuello?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas tensión ocular?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas mareos?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas aturdimiento?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas pesadez en los párpados?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

¿Te han llorado los ojos durante el experimento?

No	Si

Con respecto al comienzo del experimento, ¿experimentas dolor punzante en los ojos?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas dolor en la parte delantera de la cabeza?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas fatiga en los ojos?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas problemas para concentrarte?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas somnolencia?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas dolor en las sienes?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas nauseas?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas visión borrosa?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

¿Tienes la sensación de que tus ojos están mirando en diferentes direcciones?

No	Si

Con respecto al comienzo del experimento, ¿Tus ojos parpadean?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

¿Durante el experimento, cerraste los ojos para restablecer una visión clara?

No	Si

¿Experimentas una visión doble?

No	Si

Con respecto al comienzo del experimento, ¿experimentas dolor en la parte trasera de tu cabeza?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas dificultades para enfocar?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas ojos secos?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas ojos llorosos?

Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Con respecto al comienzo del experimento, ¿experimentas dolor en los hombros?

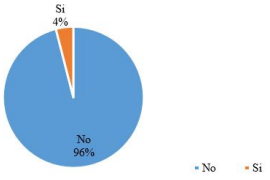
Mucho menos	Un poco menos	Lo mismo	Un poco más	Mucho más

Durante el experimento, ¿necesitó mirar un objeto diferente de la pantalla? ¿Cuál?

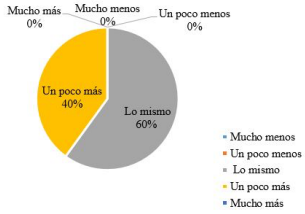
No	Si

Resultados test fatiga y molestias visuales

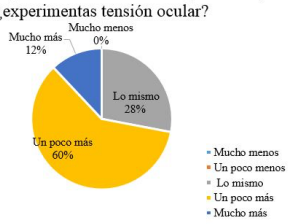
1.¿Tienes la sensación de que el movimiento de tus ojos está descoplado?



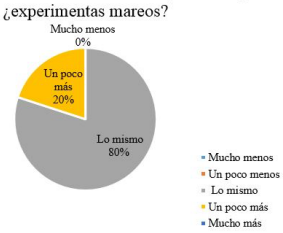
2.Con respecto al comienzo del esperimento, ¿experimentas rigidez en el cuello?



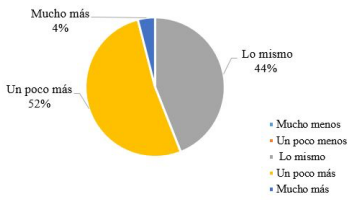
3. Con respecto al comienzo del experimento, ¿experimentas tensión ocular?



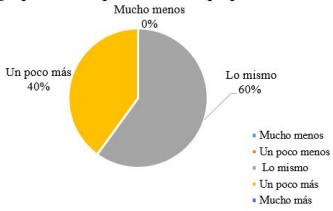
4.Con respecto al comienzo del experimento, ¿experimentas mareos?



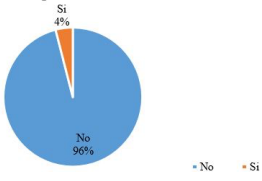
5.Con respecto al comienzo del experimento, ¿experimentas aturdimiento?



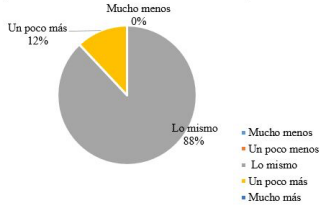
6.Con respecto al comienzo del experimento, ¿experimentas pesadez en los párpados?



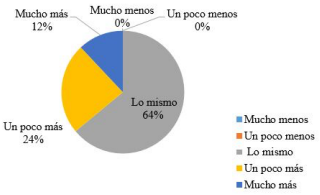
7.¿Te han llorado los ojos durante el experimento?



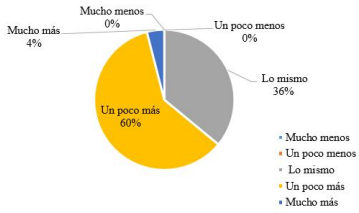
8.Con respecto al comienzo del experimento, ¿experimentas dolor punzante en los ojos?



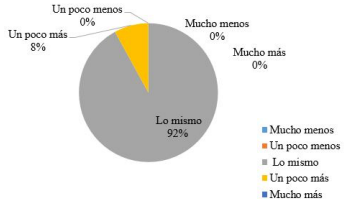
9. Con respecto al comienzo del experimento, ¿experimentas dolor en la parte delantera de la cabeza?



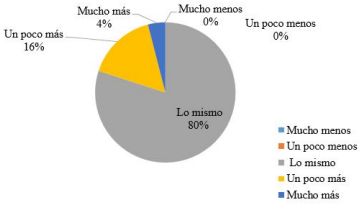
10.Con respecto al comienzo del experimento, ¿experimentas fatiga en los ojos?



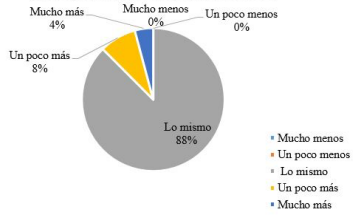
11. Con respecto al comienzo del experimento, ¿experimentas problemas para concentrarte?



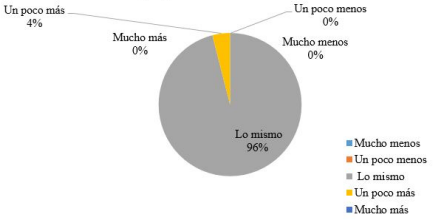
12. Con respecto al comienzo del experimento, ¿experimentas somnolencia?



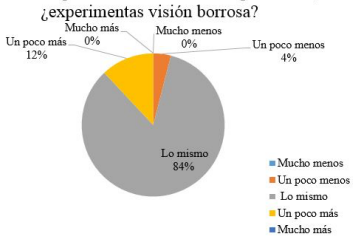
13.Con respecto al comienzo del experimento, ¿experimentas dolor en las sienes?



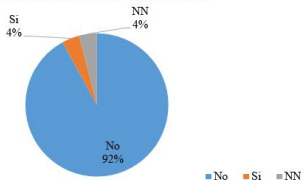
14. Con respecto al comienzo del experimento, ¿experimentas nauseas?



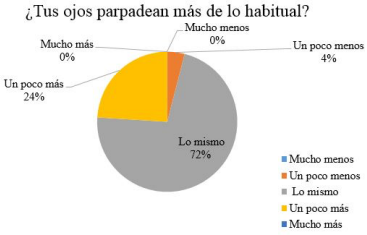
15. Con respecto al comienzo del experimento, ¿experimentas visión borrosa?



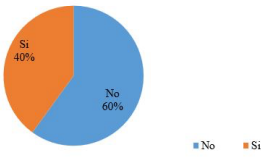
16. ¿Tienes la sensación de que tus ojos están mirando en diferentes direcciones



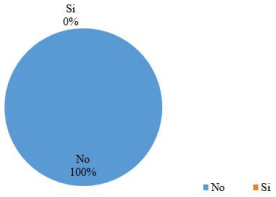
17. Con respecto al comienzo del experimento, ¿Tus ojos parpadean más de lo habitual?



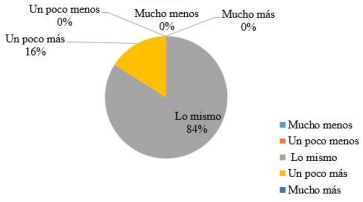
18. ¿Durante el experimento, cerraste los ojos para restablecer una visión clara?



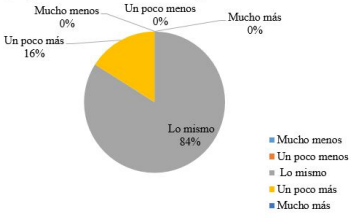
19. ¿Experimentas una visión doble?



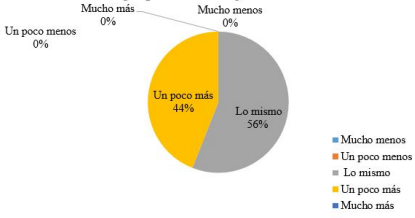
20. Con respecto al comienzo del experimento, ¿experimentas dolor en la parte trasera de tu cabeza?



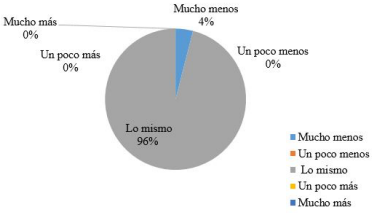
21. Con respecto al comienzo del experimento, ¿experimentas dificultades para enfocar?



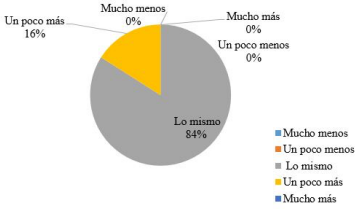
22. Con respecto al comienzo del experimento, ¿experimentas ojos secos?



23. Con respecto al comienzo del experimento, ¿experimentas ojos llorosos?



24. Con respecto al comienzo del experimento, ¿experimentas dolor en los hombros?



25. Durante el experimento, necesito mirar un objeto diferente a la pantalla para poder enfocar?