UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Implementación de estrategias de optimización computacional para la estimación precisa en tiempo real de la calidad de experiencia en vídeo

T E S I S D O C T O R A L

Presentada para optar al título de Doctor por:

Álvaro Llorente Gómez
Máster Universitario en Ingeniería de Telecomunicación

Madrid, 2025


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Telecomunicación

Doctorado en Tecnologías y Sistemas de Comunicaciones

Implementación de estrategias de optimización computacional para la estimación precisa en tiempo real de la calidad de experiencia en vídeo

T E S I S D O C T O R A L

Presentada para optar al título de Doctor por:

Álvaro Llorente Gómez
Máster Universitario en Ingeniería de Telecomunicación

Bajo la dirección de:
Dr. José Manuel Menéndez García
Dr. David Jiménez Bermejo

Madrid, 2025

Título: Implementación de estrategias de optimización computacional para la estimación precisa en tiempo real de la calidad de experiencia en vídeo

Autor: Álvaro Llorente Gómez

Programa de Doctorado: Tecnologías y Sistemas de Comunicaciones

Dirección de Tesis:

Dr. José Manuel Menéndez García, Catedrático de Universidad, Universidad Politécnica de Madrid

Dr. David Jiménez Bermejo, Profesor Ayudante Doctor, Universidad Politécnica de Madrid

Revisores Externos:

Tribunal de Tesis:

Fecha de Defensa de Tesis:

« El éxito no es un accidente. Es trabajo duro, perseverancia, aprendizaje, estudio, sacrificio y sobre todo, amar lo que estás haciendo » - Pelé

Agradecimientos

Gracias a mi familia (a mi padre, a mi madre, a mi hermana, a mi abuelo) y a mis amigos, por el apoyo constante y por estar siempre ahí, en los momentos buenos y en los no tan buenos. Gracias María, compañera de vida, por ese último empujón final y esa motivación que necesitaba para terminar de escribir este documento. A Mazapán, por hacerme sonreír y alegrarme los días.

A todos mis compañeros del GATV, gracias por el día a día y por compartir tantas y tantas horas de trabajo. En especial, a mis dos directores de tesis, José Manuel y David, gracias por vuestra paciencia, por el apoyo, por guiarme y ayudarme durante todos estos años.

Gracias a la Cátedra RTVE, a la Cátedra Video-MOS y al programa de doctorado del departamento SSR de la UPM, por haber permitido la realización de esta tesis doctoral.

Abstract

The consumption of audiovisual content has grown significantly in recent years, driven by the massive use of social networks and technological advances in the sector. In this context, the quality perceived by users, known as QoE (*Quality of Experience*), has emerged as a critical factor in the success or failure of audiovisual services. However, its evaluation is a major challenge due to the subjective nature of the user experience and the complexity of the multiple factors that influence it.

The reduction of computational cost in objective QoE evaluation represents a fundamental challenge in the field of audiovisual processing. The main objective of this doctoral thesis is to optimize video processing in objective QoE evaluation models, minimizing the computational cost without compromising the accuracy of the quality estimation. To this end, this research is based on the use of the Video-MOS tool, a proprietary solution that uses non-reference video metrics and allows real-time quality evaluation. The proposed optimization methods focus on the efficient exploitation of the spatial and temporal redundancy of the videos in order to improve the computational efficiency without compromising the reliability of the evaluation.

In order to exploit the spatial redundancy of the images, several approaches are proposed to select specific regions within the image, such as the central region or areas of interest identified by saliency detection methods. The possibility of reducing the size of the images is also investigated. These approaches imply a significant reduction in computational cost, but affect, to a greater or lesser extent, the accuracy in estimating the final MOS (*Mean Opinion Score*) value.

For the temporal redundancy of videos, uniform temporal sampling approaches are used, which allow the processing of a reduced subset of images. In addition, the use of the SSIM (*Structural Similarity Index Measure*) metric to identify significant changes between consecutive images and the use of different image types in video coding (type I, type P and type B) as a strategy to determine the relevance of each image within the temporal context of the video are evaluated.

The combination of spatial and temporal redundancy allows a significant optimization of the computational cost, with promising results. The proposed mode is based, depending on the video metric, on feature extraction with low and original resolution images. It is further complemented by a uniform temporal sampling selecting only the first frame of each video measurement, and by type I images. The mode limits the maximum number of images to be processed, ensuring real-time operation for all measurements, regardless of the complexity and type of content.

On the test video dataset, composed of 1123 three-second sequences, the proposed mode achieves an average computational cost reduction of more than 95.32%, with a MOS estimation error of 0.09. On the other hand, with the validation dataset composed of more than 174000 video sequences of the main HD DTT channels in Spain, a computational cost saving of more than 94.96% is obtained, with a MOS error of 0.11. This exhaustive validation with 144 hours of audiovisual content confirms the reliability and validity of the proposed solution.

Therefore, this research presents an effective proposal for an objective quality assessment that achieves a balance between accuracy and efficiency in the estimation of QoE with the Video-MOS tool. Although the proposal is applied to the Video-MOS solution, it is expected that the conclusions obtained can be extrapolated to other models of objective quality assessment in images and videos, and make a positive contribution to the scientific community and the entire audiovisual sector.

Resumen

El consumo de contenido audiovisual ha experimentado un crecimiento significativo en los últimos años, impulsado por el uso masivo de las redes sociales y los avances tecnológicos en el sector. En este contexto, la calidad percibida por los usuarios, conocida como Calidad de la Experiencia (QoE, *Quality of Experience*), se ha consolidado como un factor determinante en el éxito o fracaso de los servicios audiovisuales. Sin embargo, su evaluación es todo un reto por la naturaleza subjetiva de la experiencia y la complejidad de los múltiples factores involucrados.

La reducción del coste computacional en la evaluación objetiva de la QoE representa un desafío en el ámbito del procesamiento audiovisual. El objetivo principal de esta tesis doctoral es optimizar el procesamiento de vídeo en modelos objetivos de evaluación de QoE, minimizando el coste computacional sin afectar la precisión en la estimación de calidad. Esta investigación se fundamenta en el uso de la herramienta Video-MOS, una solución propia que emplea métricas de vídeo sin referencia y permite la evaluación de calidad en tiempo real. Los métodos de optimización propuestos se centran en la explotación eficiente de la redundancia espacial y temporal de los vídeos, para mejorar la eficiencia computacional sin comprometer la fiabilidad de la evaluación.

Para aprovechar la redundancia espacial de las imágenes, se proponen diversos enfoques orientados a seleccionar regiones específicas dentro de la imagen, como la región central o las áreas de interés identificadas mediante métodos de detección de saliencia. También se investiga la posibilidad de reducir el tamaño de las imágenes. Estos enfoques implican una disminución significativa del coste computacional, pero impacta, en mayor o en menor medida, en la precisión del valor de MOS (*Mean Opinion Score*).

Para la redundancia temporal del vídeo, se emplean enfoques de muestreo temporal uniforme que permiten procesar un subconjunto reducido de imágenes. Además, se evalúa el uso de la métrica SSIM (*Structural Similarity Index Measure*) para identificar cambios significativos entre imágenes consecutivas, y el uso de diferentes tipos de imagen (I, P y B) en la codificación de vídeo como estrategia para determinar su relevancia dentro del contexto temporal del vídeo.

La combinación de redundancia espacial y temporal permite una optimización sustancial del coste computacional, con resultados prometedores. El modo propuesto se basa, dependiendo de la métrica de vídeo, en la extracción de características con imágenes a baja resolución y a resolución original. Se complementa con un muestreo temporal uniforme que selecciona únicamente la primera imagen de cada medida de vídeo, e imágenes tipo I. El modo limita el número máximo de imágenes a procesar, lo que garantiza su funcionamiento en tiempo real para todas las medidas, independientemente de la complejidad y tipo de contenido.

Con el conjunto de vídeos de prueba, compuesto por 1123 secuencias de tres segundos de duración, el modo propuesto logra una reducción media del coste computacional superior al 95.32%, con un error en la estimación de MOS de 0.09. Por otro lado, con el conjunto de datos de validación compuesto por más de 174000 secuencias de vídeo de los principales canales HD

de la TDT en España, se obtiene un ahorro computacional superior al 94.96%, con un error de MOS de 0.11. Esta validación exhaustiva con 144 horas de contenido audiovisual confirma la fiabilidad y validez de la solución.

Por todo ello, esta investigación presenta una propuesta efectiva para la evaluación objetiva de calidad, logrando un equilibrio entre precisión y eficiencia en la estimación de QoE con la herramienta Video-MOS. Aunque la solución se aplica directamente sobre esta herramienta, se espera que las conclusiones obtenidas se puedan extrapolar a otros modelos de evaluación de calidad objetiva, y contribuir positivamente a la comunidad científica y al sector audiovisual.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Abreviaturas y acrónimos

3D-DCT NR-VQA	3D Discrete Cosine Transform No-Reference Video Quality Assessment
ACR	Absolute Category Rating
AV1	AOMedia Video 1
AVC	Advanced Video Coding
BIQA	Blind Image Quality Assessment
BIQI	Blind Image Quality Index
BLIINDS	Blind Image Integrity Notator using DCT Statistics
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
BVQA	Blind Video Quality Assessment
C3DVQA	Convolutional Neural Network with 3D kernels Video Quality Assessment
CCR	Comparison Category Rating
ChipQA	Chip Quality Assessment
CNN	Convolutional Neural Networks
CNN-TLVQM	Convolutional Neural Network - Two Level Video Quality Model
COINVQ	Comprehensive Interpretation Network for Video Quality
CONTRIQUE-FR	Contrastive Image Quality Evaluator - Full Reference
CORNIA	Codebook Representation for No-reference Image Assessment
CSF	Contrast Sensitivity Function
CSIQ	Categorical Subjective Image Quality
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DeepQA	Deep Image Quality Assessment
DEEPSTQ	Deep SpatioTemporal Video Quality Assessor
DeepVQA	Deep Video Quality Assessor

DeepVQUE Deep Video Quality Evaluator

DIIVINE Distortion Identification-based Image Verity and Integrity Evaluation

DisCoVQA Distortion-Content Transformers for Video Quality Assessment

DISTS Deep Image Structure and Texture Similarity

DOVER Disentangled Objective Video Quality Evaluator

DTS Decoding Time Stamp

EBU European Broadcasting Union

EPG Electronic Program Guide

ETSI European Telecommunications Standards Institute

FAST Full-reference Assessor Along Salient Trajectories

FAST-VQA Fragment Sample Transformer for VQA

FG Fine Grained

FLOSIM-FR Flow-based Similarity Index Measure - Full Reference

FMSE Foveated Mean Squared Error

FR Full Reference

FRIQUEE Feature Maps–based Referenceless Image Quality Evaluation Engine

FSIM Feature Similarity Index Measure

GM-LOG Gradient Magnitude and Laplacian Of Gaussian

GOP Group of Pictures

GPU Graphics Processing Unit

GSTVQA Generalized Spatial-Temporal Deep Feature Representation for No-reference Video Quality Assessment

HAS Human Auditory System

HD High Definition

HDR High Dynamic Range

HEVC High Efficiency Video Coding

HIGRADE HDR Image Gradient based Evaluator

HSV Hue, Saturation, Value

HVS Human Visual System

IA Inteligencia Artificial

IQA Image Quality Assessment

IQR Interquartile Range

JND Just Noticeable Difference

MAD Most Apparent Distortion

MAE Mean Absolute Error

MLSP VQA Multi-Level Spatially Pooled Deep-features Video Quality Assessment

MOS Mean Opinion Score

MOVIE Motion-based Video Integrity Evaluation

MSE Mean Squared Error

MS-SSIM Multi-Scale Structural Similarity Index Measure

NIQE Natural Image Quality Evaluator

NR No Reference

NSS Natural Scene Statistics

OTT Over The Top

PATCH VQ Patch Video Quality

PSNR Peak Signal to Noise Ratio

PTS Presentation Time Stamp

PVM Perceptual Video Metric

QoE Quality of Experience

QoS Quality of Service

RankDVQA Ranking-Inspired Hybrid Training Deep VQA

RAPIQUE Rapid and Accurate Video Quality Evaluator

RMSE Root Mean Square Error

RR Reduced Reference

RRED Reduced Reference Entropic Differences

RTVE Radiotelevisión Española

SaaS Software as a Service

SACONVA Shearlet- and CNN-based NR VQA

SI Spatial Information

SPEED QA Spatial Efficient Entropic Differencing for Quality Assessment

SR Spectral Residual

SSIM Structural Similarity Index Measure

STD Standard Deviation

ST-GREED Space-Time Generalized Entropic Difference

ST-RRED Spatio-Temporal Reduced Reference Entropic Differences

STFC SpatioTemporal Feature Combination model

STFEE SpatioTemporal Feature Extraction and Evaluation

ST-MAD Spatio-Temporal Most Apparent Distortion

ST-RRED Spatio-Temporal Reduced Reference Entropic Differences

SVM Support Vector Machine

SVR Support Vector Regression

TCP Transmission Control Protocol

TDT Televisión Digital Terrestre

TI Temporal Information

TLVQM Two Level Video Quality Model

TRRED Temporal Reduced Reference Entropic Differences

TS Transport Stream

UDP User Datagram Protocol

UGC User Generated Content

UHD Ultra High Definition

UIT Unión Internacional de Telecomunicaciones

UPM Universidad Politécnica de Madrid

VBLIINDS Video Blind Image Integrity Notator using DCT Statistics

VIDEVAL Video Quality Evaluator

VIF Visual Information Fidelity

VIIDEO Video Intrinsic Integrity and Distortion Evaluation Oracle

VIS3 Video Quality Assessment Via Analysis of Spatial and Spatiotemporal Slices

VMAF Video Multi-Method Assessment Fusion

V-MEON Video Multi-task End-to-end Optimized Neural Network

VoD Video On Demand

VQA Video Quality Assessment

VQEG Video Quality Experts Group

VSNR Visual Signal to Noise Ratio

VVC Versatile Video Coding

WCG Wide Color Gamut

XR Extended Reality

Capítulo 1

Introducción

El tráfico de contenido audiovisual a nivel mundial ha experimentado un crecimiento exponencial en los últimos años [1]. Este aumento se debe, en gran medida, a la masificación de las redes sociales, la mejora en la velocidad y conectividad de Internet, los avances tecnológicos en dispositivos audiovisuales y la proliferación de las plataformas digitales de contenidos. Estos factores han impulsado tanto la evolución y diversificación de aplicaciones y servicios multimedia como la transformación de los hábitos de consumo, intensificado especialmente desde la pandemia COVID-19 [2]. En particular, sectores como la videovigilancia, la realidad extendida (XR, *Extended Reality*), los servicios de *streaming* de vídeo y los videojuegos han consolidado al vídeo como el principal componente del tráfico global de Internet. Se estima que en 2017 el tráfico de vídeo representaba el 75 % del total en Internet, aumentando hasta el 82 % en 2022 [3]. Entre las plataformas digitales de contenidos, YouTube, Netflix, Facebook Video y TikTok han sido identificadas como las principales impulsoras de este crecimiento en el tráfico de vídeo a través de la red [4].

En la actualidad, las plataformas de vídeo bajo demanda (VoD, *Video on Demand*), los servicios OTT (*Over The Top*) y los radiodifusores tradicionales de TDT (Televisión Digital Terrestre) ofrecen al público una amplia gama de programas y servicios multimedia. En este escenario de creciente diversidad en la oferta audiovisual, la calidad del contenido se posiciona como un factor crítico para el éxito o el fracaso de un servicio específico, ya que incide directamente en la experiencia y satisfacción del usuario final [1]. Como resultado, la evaluación de la calidad percibida por los usuarios se ha convertido en una de las principales áreas de interés dentro de la industria audiovisual en los últimos años [5].

A lo largo de la cadena de valor de la producción y explotación de contenidos audiovisuales, que abarca desde la adquisición inicial del contenido multimedia hasta su consumo final por parte del usuario, cada etapa del proceso puede introducir distorsiones que afectan la calidad percibida en distintos grados. Entre las alteraciones más comunes se encuentran problemas de contraste y color derivados de las características inherentes a la escena, desenfoques, la presencia de artefactos visuales generados por las técnicas de compresión con pérdidas del vídeo, congelaciones en la reproducción, variaciones en la tasa binaria de transmisión y la pérdida

[1] https://bitmovin.com/blog/qoe-why-quality-video-matters/

de información durante la distribución del contenido [6], [7]. Estas degradaciones pueden impactar significativamente en la experiencia del usuario, haciendo esencial la implementación de estrategias efectivas para su evaluación.

La calidad percibida de un contenido audiovisual desde el punto de vista del usuario consumidor se conoce como Calidad de Experiencia (QoE, *Quality of Experience*). Según la UIT (Unión Internacional de Telecomunicaciones) [2], la QoE se define como: *la aceptabilidad general de una aplicación o servicio según la percepción subjetiva del usuario final* [8]. La evaluación de la QoE involucra múltiples factores, como el tipo de contenido, la presencia de artefactos en la señal, la experiencia previa del usuario, así como la respuesta del sistema visual humano (HVS, *Human Visual System*) y del sistema auditivo humano (HAS, *Human Auditory System*). Además, considera el impacto de las condiciones de la red y las capacidades del dispositivo de reproducción.

A pesar de los avances en la comprensión de la QoE, todavía existen numerosas cuestiones abiertas en la evaluación de la calidad percibida por los usuarios, debido a la influencia de factores humanos, de sistema y de contexto [9]. Entre los factores más relevantes que afectan a la calidad percibida de un contenido específico se encuentran los aspectos espaciales de las imágenes, como el realismo de las formas, el color y la textura, así como las características temporales del vídeo, tales como el movimiento, la fluidez y la trayectoria de los objetos [10].

Sin embargo, la percepción de la QoE está fuertemente influenciada por la variabilidad en la interpretación individual de la calidad del contenido, lo que genera diferencias significativas en la evaluación subjetiva de la experiencia. Aspectos como las expectativas previas, la familiaridad con la tecnología o las condiciones de visualización juegan un papel determinante en esta variabilidad [11]. Aunque el contenido representa aproximadamente un 77 % de la contribución total en la percepción de QoE, otros factores como el precio (84 %), la facilidad de uso del servicio (81 %) y la disponibilidad del contenido (79 %) también son esenciales para la satisfacción general del usuario [12]. En el ámbito específico de la señal de vídeo, la valoración sobre cómo de bien o cómo de mal se ve un contenido depende de la percepción subjetiva y las características del sistema visual humano. Factores como la fisiología ocular, el movimiento de los ojos y los procesos cognitivos desempeñan un papel clave en la interpretación de la calidad visual [13], [14].

La evaluación de la calidad de imagen (IQA, *Image Quality Assessment*) y de vídeo (VQA, *Video Quality Assessment*) ha sido ampliamente estudiada en las últimas décadas para estimar la QoE [15]. En ambos casos, la evaluación de QoE se clasifica en dos enfoques principales: evaluación subjetiva y evaluación objetiva. La evaluación subjetiva es el método más fiable, ya que involucra a usuarios reales que califican la calidad de imágenes o vídeos en una escala numérica. Sin embargo, su aplicación es limitada en entornos de tiempo real debido a los altos requerimientos de tiempo y rigor para llevar a cabo estas pruebas. Por esta razón, la evaluación objetiva se presenta como la alternativa más eficiente en este contexto. Este enfoque estima automáticamente la calidad del contenido audiovisual mediante modelos matemáticos de IQA y VQA, permitiendo su aplicación en escenarios donde la rapidez y la automatización son esenciales.

[2] https://www.itu.int/en/Pages/default.aspx#/es

Los principales desafíos en la evaluación objetiva de la QoE en contenidos audiovisuales siguen siendo la incorporación de elementos subjetivos en los modelos de estimación y la reducción del elevado coste computacional asociado al procesamiento de los contenidos [16], [17]. En la actualidad, los modelos objetivos propuestos en la literatura [3] tienen como objetivo mejorar tanto el rendimiento como la precisión de las estimaciones, procurando emular de manera fiel la percepción subjetiva. También se busca reducir los costes computacionales asociados al procesamiento de vídeo en la evaluación de calidad.

1.1 Motivación

El desarrollo de un modelo objetivo para la evaluación de la calidad del vídeo, capaz de estimar con precisión la percepción humana de la calidad audiovisual, sigue siendo un desafío clave en la investigación actual. Esta dificultad no solo radica en la complejidad de diseñar algoritmos matemáticos que emulen fielmente las evaluaciones subjetivas de los observadores humanos [18], sino también en la constante evolución del sector audiovisual. Los avances en tecnologías de codificación, formatos de vídeo y dispositivos de visualización exigen una adaptación continua de los métodos de evaluación para garantizar su relevancia y precisión en entornos cada vez más dinámicos. Las métricas deben detectar no solo los artefactos generados por codificadores tradicionales como H.264/AVC (*Advanced Video Coding*), sino también aquellos propios de formatos más avanzados, como H.265/HEVC (*High Efficiency Video Coding*), AV1 (*AOMedia Video 1*) o H.266/VVC (*Versatile Video Coding*) [19].

El auge de nuevos tipos de contenido audiovisual y formatos que se alejan de lo tradicional ha requerido una re-evaluación de las métricas de vídeo [20], [21]. Contenidos generados por usuarios (UGC, *User Generated Content*) [22], vídeos de alto rango dinámico (HDR, *High Dynamic Range*) [23], [24], vídeos 360º u omnidireccionales [25], [26], [27], [28], videojuegos [29] y contenidos sintéticos [30], [31], [32] presentan características visuales únicas que dificultan el análisis de calidad mediante métricas convencionales. Además, parámetros fundamentales de la señal de vídeo, como la resolución y la tasa de refresco de imagen, han evolucionado significativamente, proporcionando experiencias visuales más inmersivas y realistas. Tras la adopción masiva de la alta definición (HD, *High Definition*), el formato de ultra alta definición (UHD, *Ultra High Definition*) [33] se ha consolidado como un estándar presente y futuro en aplicaciones de vídeo. Actualmente, plataformas de *streaming* como YouTube, Netflix, Disney+ y Amazon Prime Video ya ofrecen contenido en resolución 4K UHD, marcando el camino hacia una mayor exigencia en la evaluación de la calidad audiovisual.

En este contexto, la evolución de la TDT en España juega un papel crucial para adaptarse a las crecientes expectativas de los espectadores y competir con las plataformas de *streaming* de vídeo y televisión de pago. En febrero de 2024, todas las transmisiones televisivas adoptaron el formato HD, lo que representó un avance significativo en la calidad audiovisual. Este cambio no solo mejora la calidad visual, sino que también abre la puerta a nuevas tecnologías, servicios y la emisión de canales en UHD. Desde entonces, todas las emisiones en la TDT en España son en HD, excepto La1 UHD, que transmite en 4K (3840x2160 píxeles) a 50 imágenes por

[3] https://paperswithcode.com/task/video-quality-assessment

segundo [4].

El desarrollo de métricas objetivas para evaluar la calidad de vídeo en resoluciones UHD, como 4K y 8K, plantea un desafío técnico considerable debido al alto coste computacional del procesamiento de estos formatos. Un vídeo en 4K UHD [34] tiene una resolución espacial cuatro veces mayor que un vídeo HD [35] (3840x2160 píxeles vs. 1920x1080 píxeles), mientras que el 8K UHD [34] maneja cuatro veces la información de la resolución 4K UHD y dieciséis veces la resolución HD (7680x4320 píxeles vs. 1920x1080 píxeles). Además, duplicar la tasa de refresco de 25 a 50 imágenes por segundo también duplicaría la cantidad de datos a procesar.

Dado el creciente volumen de contenidos audiovisuales y el aumento en la cantidad de información necesaria para evaluar la calidad, resulta fundamental desarrollar soluciones que reduzcan significativamente el coste computacional en la evaluación de la QoE. Optimizar el procesamiento permitiría analizar un mayor número de contenidos en paralelo con los recursos disponibles, facilitar aplicaciones en tiempo real y disminuir el consumo energético, contribuyendo así a la sostenibilidad ambiental.

La demanda global de electricidad continúa en ascenso. Según el informe *Electricity 2024* de la Agencia Internacional de Energía [36], se espera un crecimiento anual del 3.4 % hasta el año 2026. En este contexto, el sector audiovisual ha emergido como una de las áreas con mayor impacto energético, especialmente por las emisiones de gases de efecto invernadero asociadas a la transmisión de contenidos digitales. Actualmente se estima que el tráfico de vídeo a través de Internet representa cerca del 3.7 % de las emisiones globales de carbono [37]. Además, más de 5000 millones de personas (aproximadamente el 65 % de la población mundial) utilizan redes sociales y plataformas de entretenimiento, lo que conlleva un elevado gasto energético. Aunque el consumo varía según la aplicación y tiempo de uso, se estima un promedio de 10.73 mAh por usuario. Por ejemplo, el videoclip musical de *Despacito*, con más de 5000 millones de visualizaciones en la plataforma de YouTube, habría consumido una cantidad de energía comparable al gasto anual de varios países africanos [5]. Asimismo, se calcula que el consumo energético anual de plataformas como Netflix podría abastecer unas 37000 viviendas diferentes [38].

Estos datos ponen de manifiesto la urgencia de implementar estrategias sostenibles para reducir el impacto ambiental del consumo audiovisual. Diversos estudios han abordado este desafío desde una perspectiva regulatoria. Por ejemplo, el estudio [39] explora diferentes escenarios de políticas públicas aplicables a servicios de vídeo en Europa entre 2020 y 2030. Entre las medidas más eficaces destacan la limitación de la resolución de vídeo en dispositivos donde no se perciban mejoras significativas, la promoción de codificadores de vídeo más eficientes y el impulso a redes de distribución con menor consumo energético. Otras investigaciones recientes han demostrado cómo variables como la resolución del vídeo, la frecuencia de refresco, la tasa binaria, el tipo de pantalla o el dispositivo de visualización inciden directamente en el consumo energético [40], [41], [42]. Además, el proyecto europeo LoCaT [43] ha evidenciado que la TDT consume considerablemente menos energía que la distribución de contenidos por Internet, consolidándose como la opción más sostenible.

[4] https://television.digital.gob.es/ayuda-ciudadano/sala-prensa/Paginas/por-fin-llega-el-4k-para-toda-espana.aspx

[5] https://www.linkedin.com/pulse/despacito-streaming-energy-consumption-rabih-bashroush/

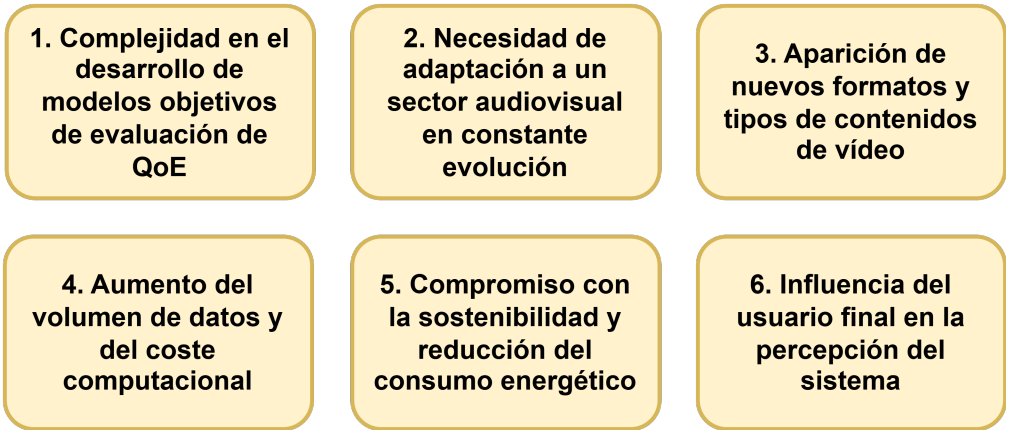

Figura 1.1: Motivación de la tesis doctoral.

Además de los aspectos técnicos y regulatorios, el comportamiento de los usuarios también influye en la sostenibilidad del sector. Algunos estudios han explorado la disposición del público a aceptar una reducción de calidad si ello conlleva a una menor huella energética, proponiendo un equilibrio entre sostenibilidad y calidad percibida [44]. También ha aparecido el concepto de *usuario ecológico*, un tipo de usuario caracterizado por una mayor conciencia ambiental que tiende a valorar más positivamente los contenidos audiovisuales energéticamente optimizados [45].

Para finalizar esta sección, la Figura 1.1 recoge de forma estructurada los factores fundamentales que motivan el planteamiento de esta investigación.

1.2 Objetivos y contribución

Considerando todo lo expuesto hasta ahora, la reducción del coste computacional se presenta como un argumento clave en el diseño de nuevos métodos y modelos para la evaluación objetiva de la calidad de vídeo. Impulsada por esta necesidad y los desafíos que conlleva, esta tesis doctoral tiene como objetivo principal optimizar el procesamiento de vídeo en la evaluación de la calidad, buscando una eficiencia máxima en el uso de los recursos computacionales, sin comprometer la precisión en la estimación de la calidad percibida.

Para reducir el coste computacional en el procesamiento de vídeo, se pueden considerar varias estrategias. Entre las más comunes se incluyen la disminución del tamaño de la imagen, el recorte de la imagen, el muestreo temporal de imágenes dentro de una secuencia de vídeo, el uso de *hardware* especializado como las Unidades de Procesamiento Gráfico (GPUs, *Graphics Processing Units*), y la paralelización y optimización de los algoritmos. No obstante, esta investigación se enfoca principalmente en explorar y aprovechar tanto la redundancia espacial como temporal presente en las imágenes consecutivas que componen una secuencia de vídeo. En el procesamiento de imágenes y vídeo, la redundancia espacial y temporal son dos características clave que permiten una representación más eficiente de la información visual. La redundancia espacial se refiere a la similitud entre píxeles cercanos dentro de una misma imagen, mientras que la redundancia temporal hace referencia a la repetición de información

entre imágenes sucesivas en una secuencia de vídeo. El objetivo es reducir el tamaño de las imágenes a procesar y seleccionar un subconjunto específico de imágenes dentro de la secuencia, con el fin de lograr ahorros significativos en el coste computacional. Diversos estudios previos, con objetivos similares, han evaluado el submuestreo espacial y temporal en el procesamiento de vídeo, obteniendo resultados muy prometedores [46], [47], [48], [49], [50], [51], [52], [53], [54].

En esta investigación, se utiliza como modelo de evaluación de la QoE la solución Video-MOS SaaS (*Software as a Service*), gracias al acuerdo de colaboración entre la empresa europea Video-MOS [6] y la UPM (Universidad Politécnica de Madrid) [7], formalizado como una cátedra universitaria de investigación [8]. Aunque en un capítulo posterior se ofrece una descripción detallada de la solución Video-MOS, esta herramienta es un sistema híbrido de métricas de vídeo sin referencia y características de la señal, que permite la monitorización, evaluación y análisis en tiempo real de la calidad de contenidos audiovisuales para diversas plataformas de contenido digital y redes de comunicación. Utiliza sofisticados algoritmos basados en IA (Inteligencia Artificial) [55], [56]. Gracias a una amplia variedad de metadatos, parámetros y características extraídas de la señal de vídeo, la solución Video-MOS es capaz de caracterizar el tipo de secuencia, detectar y cuantificar las distorsiones presentes en la señal, y proporcionar una estimación de la QoE dentro de un rango numérico de 1 a 5 en la escala MOS (*Mean Opinion Score*). Acompañada de una extensa base de datos con más de 10.000 secuencias de vídeo, la solución busca estimar la calidad subjetiva de un contenido audiovisual tal y como la percibiría un usuario promedio.

La herramienta Video-MOS está registrada en el Registro Territorial de la Propiedad Intelectual de la Comunidad de Madrid [9], y está protegida mediante cuatro módulos *software* que conforman el sistema completo:

1. Módulo de extracción de características del vídeo para generación de base de datos de entrenamiento. Identificador M-002018/2023.
2. Módulo de adquisición, captura y procesamiento de vídeo en tiempo real para medición de calidad. Identificador M-002033/2023.
3. Módulo de estimación de calidad subjetiva de vídeo mediante inteligencia artificial. Identificador M-002037/2023.
4. Módulo de gestión de datos en tiempo real para procesado de vídeo y estimación de calidad subjetiva. Identificador M-002039/2023.

Además, se está gestionando la solicitud de una patente. La solicitud más reciente se presentó en la Oficina Española de Patentes y Marcas [10] el 4 de diciembre de 2023, bajo el número Nº202331007, con el título: *Sistema y método para la estimación de la calidad percibida y su impacto en el consumidor de contenidos audiovisuales multimedia*. Aunque la solicitud fue inicialmente rechazada, actualmente se encuentra en proceso de apelación.

[6] https://www.video-mos.com/
[7] https://www.upm.es/
[8] https://www.upm.es/upm?id=CON03229&prefmt=articulo&fmt=detail
[9] https://www.comunidad.madrid/gobierno/informacion-juridica-legislacion/registro-territorial-propiedad-intelectual
[10] https://oepm.es/es/

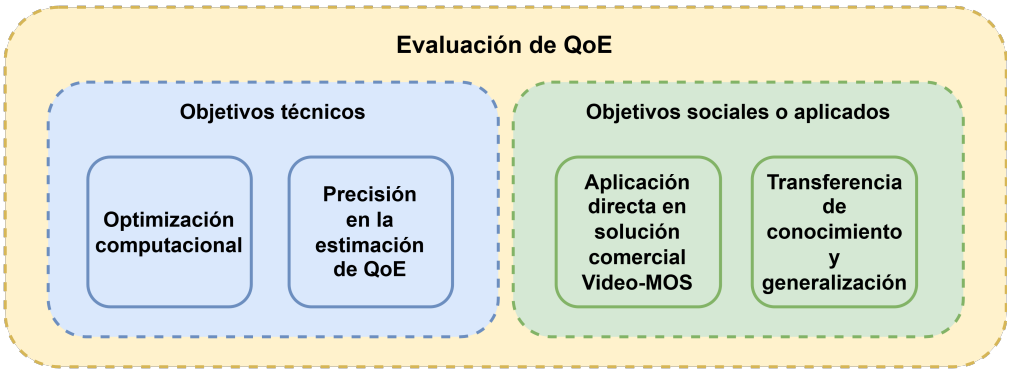

Figura 1.2: Objetivos de la tesis doctoral.

Aunque esta investigación se centra en la solución específica de Video-MOS y en el desarrollo de estrategias y soluciones para reducir el coste computacional de esta herramienta, se espera que los resultados obtenidos puedan ser aplicables y extrapolables a otras soluciones de evaluación de calidad de imagen y de vídeo.

Finalmente, en la Figura 1.2 se ofrece una representación gráfica de los objetivos principales de esta tesis doctoral.

1.3 Estructura del tomo

Esta tesis doctoral consta de seis capítulos, cuya estructura se describe a continuación.

El Capítulo 2 proporciona una descripción detallada del estado del arte en la evaluación de la calidad, abordando el concepto de QoE y sus principales desafíos, con un enfoque en la subjetividad y la percepción visual humana. Este capítulo culmina con una revisión de los estudios y trabajos más recientes en el campo de la evaluación de QoE, destacando las investigaciones más innovadoras en este ámbito.

El Capítulo 3 describe la metodología adoptada en la investigación, proporcionando una explicación exhaustiva de los materiales utilizados en las pruebas, así como el diseño y desarrollo del plan de pruebas propuesto en la tesis doctoral.

El Capítulo 4 presenta los resultados obtenidos tras aplicar diversas estrategias que exploran la redundancia espacial y temporal en las secuencias de vídeo. Para cada estrategia, se analizan sus ventajas y desventajas, el coste computacional, la extracción de características de imagen y la precisión en la estimación de QoE. Este capítulo incluye gráficos y tablas que facilitan la visualización de los datos y permiten realizar comparaciones entre los diferentes enfoques, lo que posibilita discutir la validez de cada solución propuesta.

El Capítulo 5 presenta una discusión detallada y exhaustiva de la solución propuesta para reducir el coste computacional, sin comprometer la precisión en la estimación de QoE. Para validar el modo propuesto, se utilizan más de 144 horas de contenido audiovisual de los seis principales canales de TDT en HD en España.

Finalmente, el Capítulo 6 presenta las conclusiones más significativas de esta tesis doctoral y

las líneas de trabajo futuro. En este capítulo, se sintetizan los hallazgos clave y se destacan las principales contribuciones científicas al campo de la evaluación objetiva de QoE. El capítulo concluye con una reflexión sobre el futuro de la QoE, sugiriendo diversas direcciones futuras para continuar con la investigación y expandir el conocimiento en este área de interés.

Capítulo 2

Estado de la cuestión

2.1 Introducción a la Calidad de Experiencia

El concepto de calidad hace referencia al grado de excelencia de un elemento o servicio. En el ámbito audiovisual, el concepto de calidad está estrechamente vinculado con la satisfacción del usuario durante el consumo de un determinado contenido, un proceso que depende de diferentes parámetros. A lo largo de la historia, la evaluación de la calidad audiovisual se ha centrado principalmente en la ausencia de distorsiones perceptuales, considerado un factor primordial en la experiencia final del espectador.

Tradicionalmente, la calidad del contenido audiovisual se ha valorado mediante la evaluación subjetiva, en la que los usuarios valoran directamente el contenido a través de diferentes formatos de análisis. Estas valoraciones, que se engloban dentro del campo de la evaluación de calidad subjetiva, han sido la base para comprender la percepción del usuario sobre el contenido. Sin embargo, la complejidad de realizar este tipo de evaluación, ha generado la necesidad de implementar los procesos de evaluación mediante el uso de métricas objetivas que puedan aproximarse fielmente a los juicios humanos de manera eficiente.

En sus inicios, la evaluación de calidad objetiva se basaba principalmente en parámetros técnicos objetivos, como el análisis de la imagen, el vídeo y la fidelidad de la señal. No obstante, debido a la rápida evolución tecnológica y al aumento del conocimiento sobre los procesos de percepción humana, han surgido enfoques más complejos que no solo consideran los aspectos técnicos del contenido, sino que también incorporan la experiencia global del usuario. Este enfoque, que da forma al concepto de QoE, se reconoce actualmente como un componente esencial en la evaluación integral de la calidad de los servicios audiovisuales.

En las primeras etapas, la medición de calidad en contenidos audiovisuales se basó en la evaluación de parámetros de la imagen y del vídeo a través de técnicas IQA y VQA. Estas técnicas comenzaron con la comparación pixel a pixel entre imágenes, y evolucionaron para evaluar características como la resolución, la nitidez, la fluidez del movimiento o la presencia de artefactos en la imagen. Sin embargo, la experiencia audiovisual no depende únicamente de la calidad visual. La calidad del audio también juega un papel determinante en la percepción del contenido, especialmente en contenidos con baja tasa binaria [57]. Se ha demostrado que la

calidad del audio puede compensar ciertas deficiencias en la imagen, mejorando la percepción global del usuario [58]. Además, la sincronización entre las señales de audio y de vídeo es crucial, ya que un desfase superior a 80 ms entre ambas señales puede generar una percepción negativa de la calidad del contenido [59].

Con el desarrollo de los servicios digitales y las plataformas de transmisión de contenidos, se hizo evidente que la calidad del contenido no solo dependía de sus propias características, sino también de las condiciones de entrega del servicio. En este contexto se introdujo el concepto de Calidad de Servicio (QoS, *Quality of Service*), ampliamente utilizado en telecomunicaciones para evaluar el desempeño de los sistemas de transmisión. Según la UIT-T E.800 [60], el concepto QoS se define como: *la totalidad de características de un servicio de telecomunicaciones que influyen en su capacidad para satisfacer las necesidades declaradas e implícitas del usuario del servicio*. Este enfoque permitió definir la calidad cuantificando factores clave para la correcta transmisión del contenido audiovisual como la latencia, la pérdida de paquetes y la estabilidad de la conexión. Siguiendo esta línea técnica, la especificación ETSI TR 101 290 [61], ampliamente adoptada en entornos de radiodifusión, definió un conjunto de parámetros objetivos estandarizados para evaluar la QoS en servicios de televisión digital, facilitando la monitorización de la señal y la identificación de posibles degradaciones en la transmisión del contenido, como la pérdida de información, la discontinuidad de la señal, errores de sincronización o interrupciones en los flujos de audio y vídeo. No obstante, a pesar de que una medida de QoS óptima es un requisito esencial para garantizar el funcionamiento de un servicio, su cumplimiento no siempre garantiza una experiencia satisfactoria en el usuario final. Un ejemplo claro de esta situación fue el capítulo de *Game of Thrones* titulado *The Long Night*, emitido por la plataforma HBO (ahora conocida como Max). Aunque cumplía con los estándares técnicos establecidos, fue ampliamente criticado por los espectadores debido a su excesiva oscuridad, la cual dificultaba la visibilidad y perjudicaba la experiencia de visualización. Si bien en los estudios de edición el contenido se apreciaba correctamente, factores como la compresión del vídeo, la calibración deficiente de muchas pantallas de televisión y las limitaciones propias de las plataformas de distribución afectaron negativamente la percepción final del usuario [1].

La necesidad de evaluar de manera más precisa la percepción del usuario llevó al desarrollo del concepto de QoE, que amplió el enfoque de medición más allá de parámetros técnicos. Según el organismo de la UIT, la QoE se define como: *la aceptabilidad general de una aplicación o servicio según la percepción subjetiva del usuario final* [8]. Esta definición integra múltiples factores, incluyendo la calidad del contenido, el deterioro de la señal, las expectativas y experiencias del usuario, las condiciones de la red y las capacidades del dispositivo de reproducción. En 2013, en el marco de la *COST Action* de Qualinet (Red Europea sobre Calidad de Experiencia en Sistemas y Servicios Multimedia) [2] [62] se redefinió el concepto de QoE como: *el grado de deleite o molestia del usuario de una aplicación o servicio. Resulta del cumplimiento de sus expectativas respecto a la utilidad y/o disfrute de la aplicación o servicio a la luz de la personalidad y estado actual del usuario*. Esta evolución refleja la transición desde un enfoque puramente técnico hacia una perspectiva más centrada en la percepción y

[1] https://collider.com/game-of-thrones-too-dark-to-see/
[2] https://www.qualinet.eu/resources/qualinet-white-paper/

satisfacción del usuario, incluyendo elementos como la personalidad y el estado emocional.

La evaluación de la QoE se ha convertido en un aspecto crucial para los proveedores de servicios de contenidos, como plataformas de *streaming* de vídeo o emisoras de televisión, ya que su monitorización resulta esencial para garantizar la competitividad en un mercado en constante evolución. Aunque la medición en el punto de recepción final proporciona un evaluación precisa de la experiencia del usuario, la monitorización continua en puntos intermedios de la cadena audiovisual también es relevante para detectar degradaciones en la señal y prevenir fallos incluso antes de que afecten al usuario final.

Uno de los principales desafíos en la evaluación de la QoE es la dificultad de cuantificar la percepción subjetiva del usuario. La calidad percibida no solo está determinada por factores técnicos y objetivos, sino también por variables individuales como la iluminación ambiental, la distancia de visionado, el tamaño de pantalla el tipo de contenido, el precio del servicio o la familiaridad del usuario con la tecnología. Debido a esta complejidad, la QoE se ha convertido en un campo de estudio interdisciplinario, que abarca tanto ciencias técnicas como aspectos psicológicos y sociales.

Actualmente, la evaluación de la QoE sigue siendo un área de investigación activa. La industria audiovisual continúa desarrollando métodos de evaluación objetiva que permitan estimar la calidad percibida de manera automatizada y precisa, en aplicaciones en tiempo real. A medida que los servicios audiovisuales evolucionan, la medición de la QoE se ha convertido en un elemento clave para mejorar la experiencia del usuario y garantizar la competitividad en entornos digitales cada vez más exigentes.

2.2 Desafíos en la evaluación de la Calidad de Experiencia

La calidad del contenido audiovisual es susceptible a todas las distorsiones producidas a lo largo de la cadena audiovisual, que abarca desde la etapa de adquisición del contenido hasta la reproducción en el dispositivo del usuario final. Las distorsiones pueden comprometer significativamente la percepción de la calidad por parte del usuario. La evaluación de QoE representa un desafío multidimensional debido a la complejidad de la percepción visual humana y a la interacción de diversos factores técnicos, psicológicos y contextuales. En los últimos años, disciplinas como la psicología, la sociología y las humanidades han comenzado a investigar cómo los usuarios perciben la calidad audiovisual, considerando tanto aspectos afectivos como variables socioeconómicas, como el precio del servicio o la situación económica del usuario.

2.2.1 Distorsiones producidas en el contenido audiovisual

Las señales de audio y vídeo pueden experimentar degradaciones en diversas etapas, incluyendo la adquisición, la codificación, la transmisión y la reproducción final del contenido audiovisual [63]. Estas distorsiones impactan en la calidad percibida. Las más comunes se producen durante la etapa de codificación y transmisión de la señal, aunque también pueden

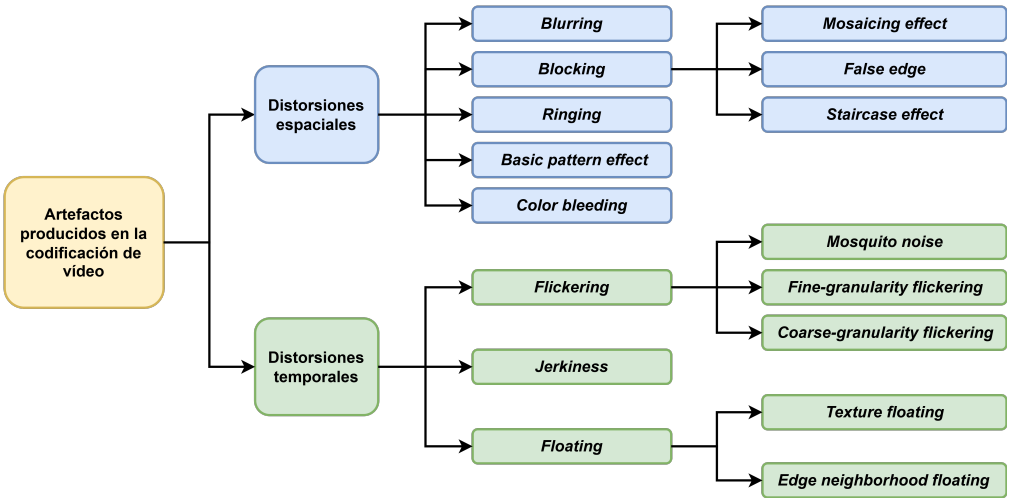

Figura 2.1: Artefactos visuales en la codificación de vídeo.

surgir problemas significativos en la adquisición y visualización del contenido. Entre los factores que afectan a la etapa de adquisición se encuentran el ruido del sensor, el desenfoque debido al movimiento de la cámara y los problemas de exposición de la luz en la escena. En la visualización, destacan las distorsiones como las aberraciones cromáticas, el ruido en la imagen y los errores de escalado o submuestreo de la información.

En entornos de transmisión, con redes de comunicaciones susceptibles a errores, los protocolos de comunicación juegan un papel crucial en la medida de QoE [64]. En función del tipo de protocolo utilizado en la transmisión de contenidos a través de Internet, se pueden producir diferentes tipos de distorsiones [65]. La pérdida de paquetes es una distorsión frecuente al utilizar el protocolo UDP (*User Datagram Protocol*), protocolo orientado a las comunicaciones en tiempo real, dado que no garantiza la retransmisión de datos perdidos y puede generar artefactos visuales y auditivos [66]. Por otro lado, el protocolo TCP (*Transmission Control Protocol*) prioriza la integridad de los datos mediante retransmisiones, lo que puede ocasionar interrupciones en la reproducción del contenido debido a eventos de *rebuffering* [67], [68]. La frecuencia de estas interrupciones constituye un factor determinante que afecta muy negativamente a la calidad de los contenidos consumidos a través de las transmisiones por Internet [69], [70].

La codificación del vídeo representa una de las principales fuentes de degradación de la calidad, ya que los algoritmos de compresión emplean esquemas con pérdidas de información para optimizar la tasa binaria. Estándares como H.264/AVC [71] y H.265/HEVC [72] utilizan técnicas de compensación de movimiento y la Transformada Discreta del Coseno (DCT, *Discrete Cosine Transform*) para la cuantificación de los coeficientes, lo que introduce artefactos perceptibles por el usuario. El proceso de cuantificación consiste en disminuir la precisión de los coeficientes transformados de una imagen, después de aplicar la transformada DCT, para reducir la cantidad de información de las imágenes. Este procedimiento trata de priorizar las pérdidas de información para los elementos menos importantes a nivel visual de la imagen, que se corresponden con los coeficientes más altos y con la información de más altas frecuencias de la imagen. Introduce un nivel de pérdidas que varía según la complejidad

Figura 2.2: Aspecto visual de las distorsiones de *blurring*, *blocking*, *flickering* y *jerkiness* en un contenido emitido por la TDT.

del contenido audiovisual y la cantidad de tasa binaria disponible [73], [74]. Muchas de las distorsiones producidas en la codificación surgen debido a este proceso de cuantificación. El resultado es la presencia de artefactos comúnmente perceptibles por el ojo humano [75]. Los artefactos generados durante la codificación del vídeo pueden ser de naturaleza espacial o temporal. Mientras que los artefactos espaciales se manifiestan como distorsiones en imágenes individuales dentro de una secuencia de vídeo, los artefactos temporales se perciben durante la reproducción del vídeo. La Figura 2.1 presenta los artefactos más comunes que se producen en el proceso de codificación del vídeo [63]. Para facilitar la comprensión y mantener la coherencia con la literatura existente, se ha decidido mantener los términos de las distorsiones en inglés. El aspecto visual de alguna de estas distorsiones se muestra en la Figura 2.2, donde las capturas corresponden a instantes diferentes de un mismo contenido deportivo emitido por la TDT.

2.2.1.1 Distorsiones espaciales

Las distorsiones espaciales surgen de la segmentación de la imagen y la cuantificación por bloques en la codificación del vídeo. Entre las distorsiones más relevantes se encuentran: *blurring* (desenfoque o emborronamiento), *blocking* (efecto de bloques), *ringing* (efecto timbre), *basic pattern effect* (efecto de patrón base) y *color bleeding* (sangrado de color). Muchas de estas distorsiones están relacionadas entre sí, se identifican a nivel de imagen y dependen principalmente de su apariencia visual.

- ***Blurring***. La codificación de vídeo consiste en transformar la señal en el dominio de la frecuencia y aplicar un proceso de cuantificación que elimina determinados coeficientes (específicamente, se eliminan los coeficientes relacionados con las altas frecuencias). Las

imágenes más nítidas tienen una mayor proporción de componentes de alta frecuencia y, por el contrario, las imágenes borrosas presentan una menor proporción de estas componentes. Por ello, la eliminación de las altas frecuencias provoca un efecto de desenfoque y pérdida de nitidez, que se refleja comúnmente en la falta de detalle en los bordes y en las áreas con texturas dentro de la imagen [76].

- ***Blocking***. Distorsión muy común que a menudo se observa en vídeos codificados a baja tasa binaria. Se produce por la cuantificación independiente de bloques individuales dentro de una imagen de vídeo, típicamente bloques de 8x8 píxeles [77]. Esta distorsión se presenta como una discontinuidad entre bloques adyacentes de una imagen, lo que produce un efecto de teselado. Según su apariencia visual, se identifican tres tipos de *blocking*: *mosaicing effect* (efecto mosaico), *false edge* (borde falso) y *staircase effect* (efecto escalera). La distorsión *mosaicing effect* se produce en transiciones de luminancia en zonas de baja energía, la distorsión *false edge* se produce cerca de bordes reales de la imagen, mientras que la distorsión *staircase effect* suele aparecer en líneas diagonales o curvas de la imagen.
- ***Ringing***. Los contornos y bordes bien definidos en una imagen producen múltiples coeficientes en el dominio de la frecuencia. Con el proceso de cuantificación y la pérdida parcial de coeficientes, se originan estructuras artificiales en áreas cercanas a los contornos y a los bordes. Este tipo de artefactos es más notable cuando se producen en regiones uniformes con alto contraste.
- ***Basic pattern effect***. Esta distorsión se asemeja a la distorsión de *ringing*, aunque no se limita únicamente a los bordes de la imagen. También puede manifestarse en regiones de la imagen con texturas y con niveles moderados de energía.
- ***Color bleeding***. Este tipo de distorsión se produce por una representación desigual de la imagen entre los canales de luminancia y crominancia del vídeo, y está relacionado con el submuestreo de la señal de color (típicamente, submuestreo de la señal YCbCr de 4:2:2 para contribución de señal de vídeo y submuestreo 4:2:0 para distribución final). Una menor resolución en los canales de crominancia roja y azul implica necesariamente realizar operaciones de interpolación a la hora de reconstruir la imagen final, provocando en ocasiones problemas relacionados con la dispersión del color.

2.2.1.2 Distorsiones temporales

Las distorsiones temporales se manifiestan durante la reproducción del vídeo y suelen ser más complejas a la hora de cuantificar este tipo de artefactos mediante métricas objetivas [78]. Las principales distorsiones temporales son: *flickering* (parpadeo), *jerkiness* (sacudidas) y *floating* (flotante).

- ***Flickering***. Este tipo de distorsión se refiere a variaciones en las señales de luminancia y crominancia que dan sensación de parpadeo al alternar valores altos y bajos de nivel de brillo en una secuencia de vídeo. Suele aparecer en áreas de fondo de la imagen como en los bordes de los objetos. Se distinguen tres tipos de *flickering* según las características de la distorsión: *mosquito noise* (ruido mosquito), *fine-granularity flickering* (parpadeo de granularidad fina) y *coarse-granularity flickering* (parpadeo de granularidad gruesa).

- ***Jerkiness***. Esta distorsión se produce cuando la resolución temporal del vídeo que se utiliza para codificar el contenido no corresponde adecuadamente con la velocidad de movimiento de los objetos, generando una percepción visual de movimiento no continuo y entrecortado. La distorsión es especialmente notable cuando afecta directamente a movimientos enfocados o en el área de interés de la imagen. Un movimiento irregular de un objeto causa una pérdida considerable de calidad.
- ***Floating***. Este tipo de distorsión temporal se relaciona con un fenómeno visual en el que ciertas partes del vídeo parecen moverse o cambiar de forma en comparación con el fondo o con áreas adyacentes, provocando un movimiento percibido de manera incorrecta. Visualmente, se asemeja a zonas específicas u objetos que parecen flotar sobre el fondo de una imagen. Según las características de esta distorsión, se pueden identificar dos tipos: *texture floating* (textura flotante) y *edge neighborhood floating* (vecindad del borde flotante).

2.2.2 Factores de influencia en la Calidad de Experiencia

La evaluación de QoE requiere la identificación y el análisis de los diversos factores que impactan significativamente en la percepción del usuario final. En los últimos años, múltiples estudios han explorado la relación entre los parámetros técnicos propios de la QoS y la percepción subjetiva de la calidad. No obstante, la principal limitación de estas investigaciones radica en la omisión de elementos clave como las experiencias previas del usuario con un servicio específico, sus expectativas, estado emocional o relación precio-beneficio del servicio.

La QoE se evalúa mediante un conjunto de indicadores y variables que trascienden la supervisión tradicional de la QoS, la cual se enfoca predominantemente en aspectos relacionados con las condiciones de la red de transmisión. Mientras que la QoS se basa en métricas cuantitativas y objetivas, la evaluación de QoE requiere de un enfoque multidisciplinar que integra diferentes factores subjetivos y contextuales.

Para modelar adecuadamente la QoE, es fundamental considerar todos los elementos que inciden en la percepción del usuario. La integración de estos elementos en un modelo único representa un desafío importante en la actualidad [79]. La QoE resulta de una combinación compleja e interrelacionada de factores, que pueden agruparse en tres categorías principales: factores de influencia humanos, factores de influencia del sistema y factores de influencia del contexto [80]. La interrelación entre estos factores impide su estudio de forma aislada. Bajo condiciones específicas en cada una de estas categorías, la QoE se evalúa mediante diferentes criterios de calidad, los cuales son ponderados en un modelo integral. La Figura 2.3 ilustra los factores que influyen en la QoE.

2.2.2.1 Factores de influencia humanos

La evolución del concepto de QoE ha enfatizado la influencia del usuario en la evaluación de la calidad. En este sentido, su estudio se vincula con disciplinas como las ciencias sociales, la psicología cognitiva, la psicología social y la antropología, siendo todas ellas esenciales para comprender el impacto de los factores humanos en la percepción de la calidad.

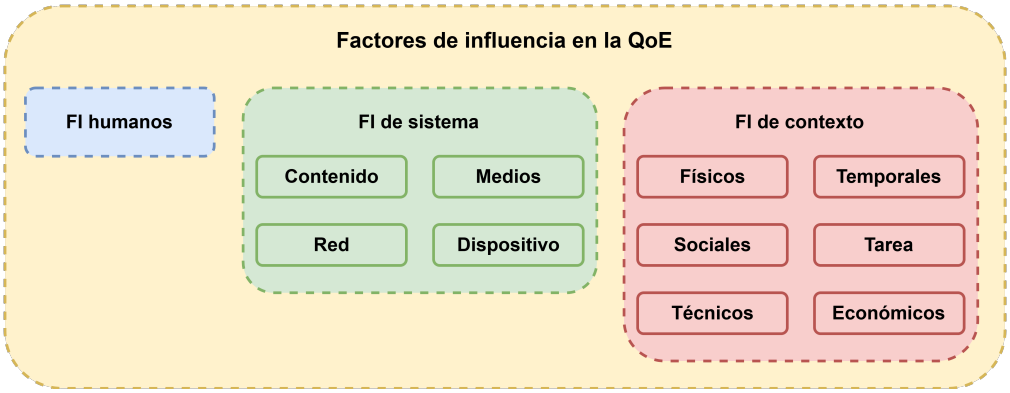

Figura 2.3: Factores de influencia en la QoE.

Los factores humanos engloban atributos individuales del usuario, tales como aspectos demográficos y socioeconómicos, estado emocional, expectativas, necesidades y condiciones físicas y mentales. Asimismo, incluyen características como el género, la edad, la agudeza visual y auditiva. Algunos de estos factores presentan correlaciones significativas, como el daltonismo, que es considerablemente más frecuente en hombres que en mujeres.

Algunos factores, como la edad, el género y la agudeza visual, son relativamente fáciles de cuantificar y suelen emplearse para caracterizar grupos de estudio en evaluaciones subjetivas. Sin embargo, no siempre se considera este tipo de información como variables independientes a la hora de modelar la estimación de la QoE. Por el contrario, factores psicológicos y sociales, como el nivel educativo, el contexto sociocultural o la situación económica del usuario, son de difícil medición debido a su naturaleza abstracta e intangible. Aspectos como emociones, valores, actitudes y expectativas influyen en la QoE de manera dinámica, variando en el tiempo y entre individuos. La motivación, la personalidad y las preferencias también se incluyen en este grupo.

A pesar del creciente interés por estos factores, su aplicación en modelos cuantitativos de QoE sigue siendo limitada. Las investigaciones futuras deberán considerar métodos más avanzados para capturar la diversidad de experiencias y emociones de los usuarios, así como la influencia de la satisfacción o la frustración en la percepción de la calidad por un determinado servicio.

2.2.2.2 Factores de influencia de sistema

Los factores del sistema abarcan aquellos elementos técnicos que impactan directamente en la calidad percibida del contenido audiovisual. Dentro de este tipo de factores se encuentran las imperfecciones del sistema óptico de captura de la señal de vídeo, las capacidades del sensor de la cámara, las particularidades del tipo de contenido, los parámetros de codificación del vídeo, la tasa binaria o las especificaciones de los dispositivos de visualización. Por todo ello, estos factores pueden agruparse en cuatro categorías principales: contenido, medios, red y dispositivo.

- **Contenido**. Incluye las propiedades intrínsecas del material audiovisual, como la información espacial, la información temporal, la colorimetría, las texturas y las características tridimensionales.

- **Medios**. Comprende parámetros asociados al contenido, tales como la resolución de la imagen, la frecuencia de refresco del vídeo, la tasa de muestreo del audio y la sincronización entre la señal de audio y la señal de vídeo.
- **Red**. Describe las condiciones de la transmisión, incluyendo la pérdida de paquetes, la latencia, las interrupciones en la reproducción del contenido y las fluctuaciones de la tasa binaria.
- **Dispositivo**. Se refiere a las especificaciones y configuraciones de los equipos utilizados para la reproducción del contenido, así como el tamaño y la resolución de pantalla, la distancia y ángulo de visualización. Las características del dispositivo como los niveles de brillo, contraste, la nitidez, el color o la naturalidad de la escena, también afecta considerablemente en la calidad final percibida por el usuario.

2.2.2.3 Factores de influencia de contexto

Los factores de contexto engloban aquellas condiciones externas al usuario y al sistema, y que también influyen en la QoE. Estos factores pueden clasificarse en:

- **Físicos**. Incluye el entorno físico en el que se encuentra el usuario, como la ubicación geográfica y las condiciones de iluminación.
- **Temporales**. Se refieren a aspectos cronológicos, como la hora y momento del día, y la duración de la interacción con el contenido.
- **Sociales**. Consideran la dinámica social del usuario al consumir el contenido, si está solo o acompañado.
- **Tarea**. Evalúan el propósito de uso del servicio, diferenciando entre entretenimiento, aprendizaje o trabajo, entre otros.
- **Técnicos**. Analizan el grado de interacción del usuario con el sistema, así como el uso de dispositivos adicionales.
- **Económicos**. Consideran la relación entre el precio y el beneficio del servicio, factor determinante en la percepción de calidad.

2.2.3 Percepción del Sistema Visual Humano

Durante décadas, numerosos estudios han investigado el impacto del HVS en la evaluación de la calidad de imágenes y vídeos. La comprensión completa de la visión humana sigue representando un desafío crucial para la neurociencia. De hecho, entre los cinco sentidos humanos, la vista es el sentido que emplea la mayor cantidad de neuronas en el cerebro y constituye la modalidad sensorial más avanzada del ser humano [81].

El HVS es un sistema complejo que integra todo el proceso relacionado con la visión del ser humano, y que abarca desde la recepción de estímulos visuales en los ojos hasta su procesamiento en las áreas corticales del cerebro. La luz ingresa al ojo humano a través de la córnea, una capa transparente situada en la parte frontal del ojo. Posteriormente, la luz pasa por la pupila, una abertura regulada por el iris encargada de controlar la cantidad

de luz entrante, y el cristalino, que enfoca la luz sobre la retina. La retina contiene células fotosensibles que transmiten señales a la corteza visual para su interpretación. Existen dos tipos principales de células fotosensibles: los conos y los bastones. Los conos son responsables de la percepción del color en condiciones de buena iluminación, mientras que los bastones permiten la visión en condiciones de baja luminosidad. La distribución de los conos en la retina es desigual, concentrándose una mayor cantidad de conos en la fóvea, zona central del campo visual. Existen tres tipos de conos, cada uno de ellos sensible a una determinada longitud de onda, que dividen la imagen proyectada en la retina en tres corrientes visuales, relacionadas con los colores rojo, verde y azul. Por su parte, los bastones predominan en la periferia de la retina, son más sensibles a una amplia gama de colores y forman la mayor parte de las células de la retina.

2.2.3.1 Características psicofísicas del Sistema Visual Humano

El desarrollo de modelos de HVS requiere un conocimiento detallado de cada fase biológica involucrada en el procesamiento de la información visual. Las características psicofísicas del HVS describen cómo los seres humanos perciben y procesan la información visual, incluyendo aspectos fisiológicos y psicológicos. A continuación, se detallan algunas de las características más relevantes del HVS:

- **Visión foveal y periférica**. La visión foveal se centra en la parte central de la retina llamada fóvea, la región con mayor densidad de conos, responsables de la agudeza visual y la percepción de detalles precisos. En contraste, la visión periférica abarca las zonas retinianas adyacentes a la fóvea, con menor densidad de conos. Esto implica que la resolución espacial es máxima en el punto de fijación del ojo humano y disminuye a medida que se aleja del punto de enfoque. La mayoría de los modelos de calidad visual suelen priorizar la visión foveal, al estar estrechamente vinculada con la atención visual.
- **Visión fotópica y escotópica**. El HVS opera bajo dos modos de funcionamiento según las condiciones de iluminación. La visión fotópica se da en entornos bien iluminados, con predominio de los conos y la percepción de los colores. En cambio, la visión escotópica ocurre en condiciones de baja luminosidad, utilizando los bastones, aunque sin discriminación cromática. La adaptación a la luz es el proceso que modula la cantidad de luz que entra al ojo humano a través de la pupila. Este fenómeno se refiere a la capacidad del ojo humano para ajustarse a los diferentes niveles de iluminación presentes en el entorno, funcionando adecuadamente en entornos con mucha luz (visión fotópica) y en situaciones de baja luminosidad (visión escotópica).

  La adaptación del HVS a las condiciones de la luz provoca que no haya una relación directa entre la intensidad objetiva de la luz (luminancia) y la forma en que se percibe su brillo. La manera en que se percibe el brillo depende tanto de la intensidad del entorno como de las variaciones recientes de esa intensidad que se hayan experimentado. El HVS opera en un rango muy amplio de intensidades luminosas, abarcando muchos órdenes de magnitud. La percepción del brillo no es lineal con la intensidad luminosa, conforme a la ley de Weber-Fechner [82].
- **Sensibilidad al contraste**. El contraste mide la variación de luminancia en una

región específica en comparación con las áreas adyacentes. La función de sensibilidad al contraste (CSF, *Contrast Sensitivity Function*) describe la capacidad del HVS para detectar diferencias en luminancia, y por tanto diferenciar un objeto del fondo. Esta sensibilidad se mide por el nivel de contraste más bajo que un observador puede percibir para un estímulo particular. El HVS responde de manera más sensible a cambios locales en la luminancia en comparación con la luminancia global o ambiental. De esta manera, conforme a la ley Weber-Fechner, el ojo humano percibe con mayor claridad las diferencias de brillo entre áreas cercanas que los niveles absolutos de iluminación. La percepción de contraste se produce debido a las variaciones espaciales en la intensidad, permitiendo diferenciar líneas, formas y objetos [83].

Otro concepto relacionado con la percepción del contraste es el umbral de visibilidad y el término JND (*Just Noticeable Difference*). El umbral de visibilidad indica el nivel mínimo de contraste necesario para poder percibir variaciones en la intensidad luminosa. Este umbral determina por tanto el punto en el que una diferencia se vuelve lo suficientemente perceptible para ser detectada. Por otro lado, el concepto JND define la menor variación en un estímulo visual (como el color, el brillo o el contraste) que puede ser detectado por el ojo humano. El concepto de JND es muy utilizado en modelos de percepción visual, ya que cambios de intensidad por debajo de este valor no son perceptibles por el usuario.

- **Enmascaramiento espacial**. Este fenómeno está relacionado con el umbral de visibilidad y describe cómo la presencia de una señal cercana (señal de enmascaramiento), puede alterar la percepción de otra señal visual. La magnitud del enmascaramiento depende de la similitud entre las señales involucradas, pudiendo aumentar o disminuir la visibilidad de ciertos elementos dentro de una imagen [84]. La mayoría de los métodos de evaluación de calidad también incluyen modelos de enmascaramiento para evaluar la visibilidad de determinados patrones específicos de la imagen.

2.2.3.2 Modelos basados en la percepción del Sistema Visual Humano

El HVS no percibe de manera uniforme todos los cambios que se producen en una imagen o vídeo. Por ello, numerosos modelos de percepción visual se centran en la extracción de características visuales mediante pruebas del umbral de visibilidad, de diferencias perceptibles (JND) por el ojo humano y de la atención visual.

Los modelos basados en JND determinan umbrales por debajo de los cuales las alteraciones en una imagen resultan imperceptibles. Factores como la edad, la experiencia visual previa y las condiciones de visualización pueden influir en la habilidad de una persona para detectar cambios en la calidad visual. El umbral JND también puede variar en función de otros aspectos como la resolución de imagen, la tasa binaria o la profundidad de color [85], [86], [87], [88].

Por otro lado, los modelos de atención visual analizan cómo los observadores dirigen su atención a distintas zonas de una escena. La atención visual es un mecanismo cognitivo crucial, estudiado en fisiología, psicología, neurociencia y visión artificial [89]. Este tipo de modelos sugieren que los seres humanos focalizan su atención en áreas específicas de la imagen, ignorando parcialmente otras zonas. En las zonas alejadas del área de interés de la imagen,

la sensibilidad visual disminuye significativamente, por lo que no todas las distorsiones se pueden considerar por igual.

En este contexto de estudio, la saliencia visual es un concepto fundamental que permite identificar las áreas más relevantes dentro de una imagen. En procesamiento de imágenes y vídeo, la saliencia se refiere al grado en que ciertos elementos de una escena visual destacan con respecto a su entorno [90]. La saliencia se determina según atributos como el color, el brillo, el contraste, la orientación, la textura y el movimiento de los objetos, generando mapas de saliencia (imagen binaria con valor blanco o negro) que destacan las áreas de interés visual [91]. Los modelos clásicos de detección de saliencia, basados en contraste y movimiento, han evolucionado con redes neuronales convolucionales [92]. En la literatura, destacan los modelos de detección de saliencia basados en la combinación de saliencia espacial y saliencia temporal [93], [94], [95], [96], [97], [98]. La saliencia espacial utiliza parámetros como el color, el brillo y las texturas, mientras que la saliencia temporal se basa en la estimación del movimiento a través del flujo óptico y la diferencia entre imágenes consecutivas en una secuencia de vídeo.

Uno de los propósitos clave en la detección de saliencia es disminuir el coste computacional. Algoritmos como la detección de rostros de personas [99] o la detección del movimiento [100], [101], implican un elevado coste computacional. Además, factores como el atractivo visual de una imagen o el sesgo del contenido también pueden afectar notablemente a la atención visual. La nitidez, el contraste y el color son características esenciales que determinan la preferencia o no por una imagen [102], [103]. Las imágenes suelen ser consideradas normalmente atractivas si son coloridas, están bien iluminadas, son nítidas o tienen altos contrastes. Por el contrario, las imágenes oscuras, borrosas o con bajo contraste son típicamente percibidas como de baja calidad o menos atractivas. El sesgo presente en el contenido puede afectar tanto a la percepción como a la opinión de los usuarios. Las preferencias individuales tienen la capacidad para impactar significativamente en la percepción de calidad. La estética, vinculada al sesgo del contenido, desempeña un papel crucial en la evaluación de imágenes como más o menos atractivas [104], [105], [106].

2.3 Evaluación de la Calidad de Experiencia

En los últimos años, la investigación en la evaluación de QoE ha adquirido una relevancia creciente. Este interés se evidencia en la aparición de estudios recientes, el incremento de iniciativas de estandarización impulsadas por la industria y la expansión de diversas actividades en el ámbito académico y tecnológico. Todos estos esfuerzos han sido motivados, en gran medida, por el aumento exponencial en la emisión de contenidos audiovisuales a nivel global.

La medición de QoE puede llevarse a cabo a través de dos enfoques: evaluación subjetiva de calidad y la evaluación objetiva de calidad. Tradicionalmente, debido a su naturaleza fundamentalmente subjetiva, la QoE ha sido evaluada mediante la participación directa de observadores reales, considerándose este método como el más fiable. No obstante, las múltiples limitaciones asociadas a la evaluación subjetiva han impulsado el desarrollo y la implementación de metodologías automáticas y objetivas para la estimación de QoE.

2.3.1 Evaluación subjetiva de calidad

La evaluación subjetiva constituye el método más fiable para evaluar la calidad percibida, al estar basada directamente en la percepción humana de un conjunto de observadores. Este enfoque no solo examina la señal audiovisual en si misma, sino que también incorpora otras factores determinantes, como las expectativas, las emociones y los sentimientos de los usuarios. En consecuencia, se considera la metodología ideal para determinar la calidad percibida en entornos audiovisuales.

El procedimiento de evaluación subjetiva implica solicitar a un conjunto de observadores reales que califiquen la calidad de una imagen o vídeo, generalmente utilizando una escala numérica estandarizada, como la escala MOS. La escala MOS, definida en la recomendación UIT-R BT.500 [107], asigna cinco posibles valores en función del juicio subjetivo de cada observador: 1 (malo), 2 (mediocre), 3 (aceptable), 4 (bueno) y 5 (excelente). Para obtener una estimación global de la calidad percibida, los valores individuales se promedian, complementándose con medidas de dispersión (como la desviación estándar, la varianza, los valores extremos o los intervalos de confianza) que reflejan la variabilidad de los datos y el carácter estadístico de la evaluación.

A pesar de su fiabilidad, la evaluación subjetiva presenta importantes limitaciones. La preparación, ejecución y análisis de estas pruebas requieren una inversión significativa de tiempo y recursos, lo que dificulta su aplicación en contextos que demandan evaluaciones en tiempo real. Además, para garantizar la aceptación de los resultados en la comunidad científica, es fundamental adherirse a metodologías establecidas por normas internacionales, como la UIT-R BT.500 [107] y la UIT-T P.910 [108]. Estas recomendaciones especifican diversos aspectos metodológicos, incluyendo las condiciones del entorno de prueba (iluminación, distancia de visualización y configuración de los dispositivos), la duración de las sesiones, el número mínimo de participantes, los materiales audiovisuales utilizados y la forma en que deben presentarse los contenidos. Asimismo, establecen los métodos y las escalas de evaluación, así como las métricas necesarias para el análisis de los resultados.

En lo que respecta a la presentación del contenido audiovisual durante una prueba subjetiva, se distinguen tres tipos de métodos diferentes: estímulo simple, estímulo doble y estímulo triple. En el método de estímulo simple, el usuario evalúa exclusivamente el contenido de prueba. En el estímulo doble, tanto el contenido de prueba como un contenido de referencia son mostrados al observador para su comparación. Finalmente, en el estímulo triple, el contenido de referencia se presenta en primer lugar, seguido de manera aleatoria por el contenido de prueba y el contenido de referencia, con el objetivo de reducir posibles sesgos en la evaluación.

Para la medición de la calidad percibida, se emplean diversas escalas subjetivas estandarizadas, entre las que destacan: escala de Calificación de Categoría Absoluta (ACR, *Absolute Category Rating*), escala de Calificación de Categoría de Degradación (DCR, *Degradation Category Rating*) y escala de Calificación de Categoría de Comparación (CCR, *Comparison Category Rating*). En la escala ACR, los observadores califican directamente la calidad del contenido utilizando una escala estándar, como la escala MOS de cinco valores. En la escala DCR, el contenido de prueba se compara con el contenido de referencia y se califica el nivel de degradación en una escala que va desde *muy molesto* hasta *imperceptible*. Finalmente, en la

escala CCR, los participantes comparan directamente dos contenidos de prueba y seleccionan cuál es superior, eligiendo entre siete categorías disponibles: *mucho peor*, *peor*, *ligeramente peor*, *igual*, *ligeramente mejor*, *mejor* y *mucho mejor*.

2.3.2 Evaluación objetiva de calidad

La evaluación objetiva de calidad tiene como propósito predecir de manera automática la calidad de un contenido audiovisual mediante modelos matemáticos. Durante años, el Grupo de Expertos en Calidad de Vídeo (VQEG, *Video Quality Experts Group*) ha investigado el desempeño de las evaluaciones objetivas de calidad en imágenes y vídeos. El objetivo principal de estos métodos es obtener una estimación cuantitativa de la calidad percibida que actúe como sustituto de la evaluación subjetiva realizada por observadores humanos. La mayoría de estos enfoques se basan en modelar las características estadísticas asociadas con la señal audiovisual, las imágenes naturales y las distorsiones, aplicando principios fundamentados en la percepción humana, como el umbral de visibilidad, las diferencias perceptibles por el ojo humano o la atención visual.

El rendimiento de cualquier modelo objetivo se mide en función de su correlación con evaluaciones subjetivas obtenidas mediante pruebas con observadores reales. Dado que estos modelos intentan imitar la percepción humana, su eficacia se encuentra estrechamente ligada a la capacidad de replicar fielmente las calificaciones otorgadas por los usuarios. Para este fin, se han desarrollado extensas bases de datos que incluyen diversos tipos de contenido y distorsiones visuales, junto con los estudios experimentales detallados con observadores reales. Las calificaciones derivadas de estas pruebas subjetivas actúan como referencia principal y resultan esenciales para evaluar la precisión de los algoritmos. Además, en los últimos años, los avances en aprendizaje automático e IA han permitido utilizar las calificaciones subjetivas obtenidas de observadores reales como datos de entrenamiento para los modelos, con el objetivo de mejorar su capacidad predictiva durante la fase de prueba. Algunas de las bases de datos más utilizadas en la evaluación de calidad de imágenes y vídeos son:

- *LIVE Image Quality Assessment Database* [109].
- *IRCCyN/IVC Image Quality Database* [110].
- *Categorical Subjective Image Quality (CSIQ) Database* [111].
- *KADID-10k IQA Database* [112].
- *ESPL-LIVE HDR Image Quality Database* [113].
- *LIVE Video Quality Database* [114].
- *LIVE Mobile VQA Database* [115].
- *IRCCyN/IVC HD Video Database* [116].
- *YouTube-UGC Database* [117].
- *MCL-V Database* [47].
- *Ultra Video Group Database* [118].

- *SJTU 4K Video Sequences* [119].
- *KoNVID-1k VQA Database* [120].
- *AVT-VQDB-UHD-1 Database* [51].
- *BVI-HD Database* [121].
- *MCML 4K UHD* [20].
- *ITS4S Database* [122].
- *TUM 1080p Database* [123].

En la evaluación objetiva de calidad, es posible establecer diversas clasificaciones. Una de las más relevantes se basa en la disponibilidad o no de la señal original, también denominada señal de referencia o señal sin distorsión. La señal de referencia permite evaluar el grado de alteración presente en la señal de prueba o señal distorsionada. Según la disponibilidad de la referencia, los métodos de evaluación objetiva se pueden agrupar en tres categorías diferentes: Referencia Completa (FR, *Full Reference*), Referencia Reducida (RR, *Reduced Reference*) o Sin Referencia (NR, *No Reference*), también conocida como Referencia Ciega (BVQA o BIQA, *Blind Video Quality Assessment* o *Blind Image Quality Assessment*).

- **Referencia completa**. En la evaluación FR, la señal de prueba se compara directamente con la señal de referencia original. Este enfoque ha sido ampliamente utilizado en la evaluación de calidad de imágenes y vídeos debido a su efectividad y a su simplicidad dentro del procesamiento de señales. Aunque estas métricas pueden presentar ciertas limitaciones en términos de percepción visual humana, siguen siendo populares gracias a su sencillez y fácil implementación.

  El objetivo principal de estas métricas es proporcionar una medida cuantitativa que describa el nivel de error o diferencia entre la señal original y la señal degradada. Sin embargo, su principal desventaja es la necesidad de contar con la señal de referencia completa en su estado original, lo que limita su aplicabilidad en escenarios donde la referencia no está disponible, como en la transmisión en tiempo real o en entornos de *streaming*.

- **Referencia reducida**. Las métricas RR evalúan la calidad del contenido utilizando únicamente una parte de la información de la señal de referencia, sin necesidad de disponer de ella en su totalidad. Estas métricas se basan en la extracción de características clave del contenido, como la información espacial, la información temporal y los parámetros de codificación, lo que permite realizar estimaciones precisas de calidad percibida.

  En general, a mayor cantidad de información disponible sobre la señal de referencia, más precisa será la métrica RR. Si la información proporcionada fuera suficiente para reconstruir completamente la señal original, las métricas RR serían equivalentes a las métricas FR.

- **Sin referencia**. Las métricas NR evalúan el contenido degradado sin necesidad de comparar la señal con una señal de referencia. Estas métricas son más sofisticadas que las métricas FR y RR, ya que deben estimar la calidad basándose únicamente en la

señal distorsionada.

Para ello, las métricas NR examinan varias características del contenido a través de diferentes dominios, como el espacial, el de Fourier, el de la transformada de la DCT y transformaciones polinómicas. Estas métricas se enfocan principalmente en aspectos visuales, tales como la nitidez, el color y la textura. Asimismo, para una evaluación completa del vídeo, es esencial considerar también características temporales, como el movimiento de los objetos en una secuencia de vídeo.

Además de la clasificación basada en la disponibilidad de la señal de referencia, existe otra categorización ampliamente utilizada que divide los métodos objetivos en dos grandes grupos: métricas de fidelidad de la señal y métricas de calidad perceptual [124].

- **Métricas de fidelidad de la señal**. Este tipo de métricas permiten evaluar la calidad de una señal distorsionada mediante su comparación con una señal de referencia, sin considerar la naturaleza del contenido. Las métricas cuantifican el grado de similitud entre la imagen de referencia y la imagen degradada, enfocándose en mediciones objetivas de las características del vídeo, sin incorporar factores relacionados con la percepción visual humana. Su uso está ampliamente aceptado en la comunidad científica debido a diversas razones: poseen definiciones claras, presentan una estructura matemática bien fundamentada y sus fórmulas son relativamente simples y de fácil implementación.

  El principio fundamental de estas métricas radica en la comparación de píxeles de manera individual entre la imagen de referencia y la imagen degradada, sin considerar las relaciones espaciales o temporales del contenido. Entre las métricas objetivas de fidelidad más utilizadas se encuentran la métrica PSNR (*Peak Signal to Noise Ratio*) y MSE (*Mean Squared Error*) [125], [126], y la métrica SSIM (*Structural Similarity Index Measure*) [127]. Aunque los valores obtenidos a través de estas métricas no siempre reflejan con precisión la percepción subjetiva de los observadores, resultan útiles para detectar variaciones en la calidad de la imagen, aunque no necesariamente cuantifiquen su impacto perceptual. Esta discrepancia entre la medición objetiva y la percepción humana se debe a varios factores:

  1. No todos los cambios en una imagen son perceptibles.
  2. La percepción de los píxeles no es uniforme en toda la imagen.
  3. No todos los cambios implican necesariamente una degradación visual.
  4. Cambios de igual magnitud pueden generar efectos perceptivos distintos, debido a los fenómenos de enmascaramiento espacial, temporal o cromático de las señales.

- **Métricas de calidad perceptual**. A diferencia de las métricas de fidelidad de la señal, las métricas de calidad perceptual consideran tanto las propiedades intrínsecas de la señal como los principios fundamentales de la percepción humana. Este tipo de métricas, fundamentadas en modelos psicofísicos y en el conocimiento de la percepción visual humana, incorporan características esenciales de la percepción. Su implementación, a pesar de ser compleja, se basa en la extracción de características de la señal audiovisual y en el uso de modelos de regresión, junto con herramientas avanzadas de aprendizaje

automático, incluidas las redes neuronales.

La mayoría de estas métricas descomponen la señal en imágenes y se sustentan en la detección de características visuales relevantes, como el contraste y el movimiento, así como la identificación de artefactos y distorsiones típicas, tales como la distorsión de *blocking*, de *ringing* o de *blurring*. Para optimizar su eficiencia, comúnmente se emplea la componente de luminancia (señal Y en el espacio de color YCbCr) en la caracterización de la señal, dado que la luminancia juega un papel más relevante en la percepción visual humana en comparación con la información cromática.

Basadas en las propiedades del HVS, tanto en el dominio espacial como en el dominio de la frecuencia, estas métricas cuantifican el impacto de las distorsiones mediante la integración de aspectos clave de la visión, como la percepción del color, la sensibilidad al contraste (umbral de visibilidad y concepto JND), los efectos de enmascaramiento y la atención visual. La información visual más relevante suele transmitirse a través del contraste, ya sea en términos de luminancia, color, textura o movimiento.

Entre las métricas objetivas de calidad perceptual más reconocidas se encuentran las métricas VMAF (*Video Multi-Method Assessment Fusion*) [128], MOVIE (*Motion-based Video Integrity Evaluation*) [129] y ST-RRED (*Spatio-Temporal Reduced Reference Entropic Differences*) [130]. Todas estas métricas están diseñadas para aproximarse con mayor precisión a la percepción subjetiva de los observadores y abordar las limitaciones identificadas en las métricas de fidelidad de la señal.

La evaluación objetiva de QoE permite superar muchas de las limitaciones propias a las evaluaciones subjetivas de calidad de vídeo. No obstante, el principal desafío de los métodos objetivos radica en lograr un nivel de precisión comparable al de los métodos subjetivos, al tiempo que se garantiza su viabilidad para aplicaciones en tiempo real. Esta tarea se ve dificultada por el alto coste computacional asociado al procesamiento de vídeo, lo que representa un obstáculo significativo para su implementación eficiente.

2.4 Evolución de las métricas objetivas de calidad

Con el tiempo, los estudios sobre calidad de imagen y vídeo han evolucionado para comprender cómo las distorsiones introducidas a lo largo de la cadena audiovisual afectan la señal multimedia y su impacto en la percepción del usuario. Asimismo, se ha observado una transición progresiva desde modelos tradicionales hacia enfoques basados en aprendizaje automático, lo que ha permitido avances significativos en la estimación de la calidad. La Figura 2.4 sintetiza y organiza la literatura revisada en este campo, destacando las principales tendencias y contribuciones. La evolución constante de los modelos objetivos ha conducido a mejoras sustanciales en la capacidad de predicción de la calidad audiovisual, optimizando la correlación entre las evaluaciones objetivas y la percepción subjetiva de los usuarios.

El desarrollo de modelos VQA basados en aprendizaje automático sigue siendo un campo de investigación activo que enfrenta múltiples desafíos. Una de las principales limitaciones en la implementación práctica de estos modelos es la escasez de bases de datos adecuadas para el entrenamiento. En general, las bases de datos disponibles no son lo suficientemente extensas

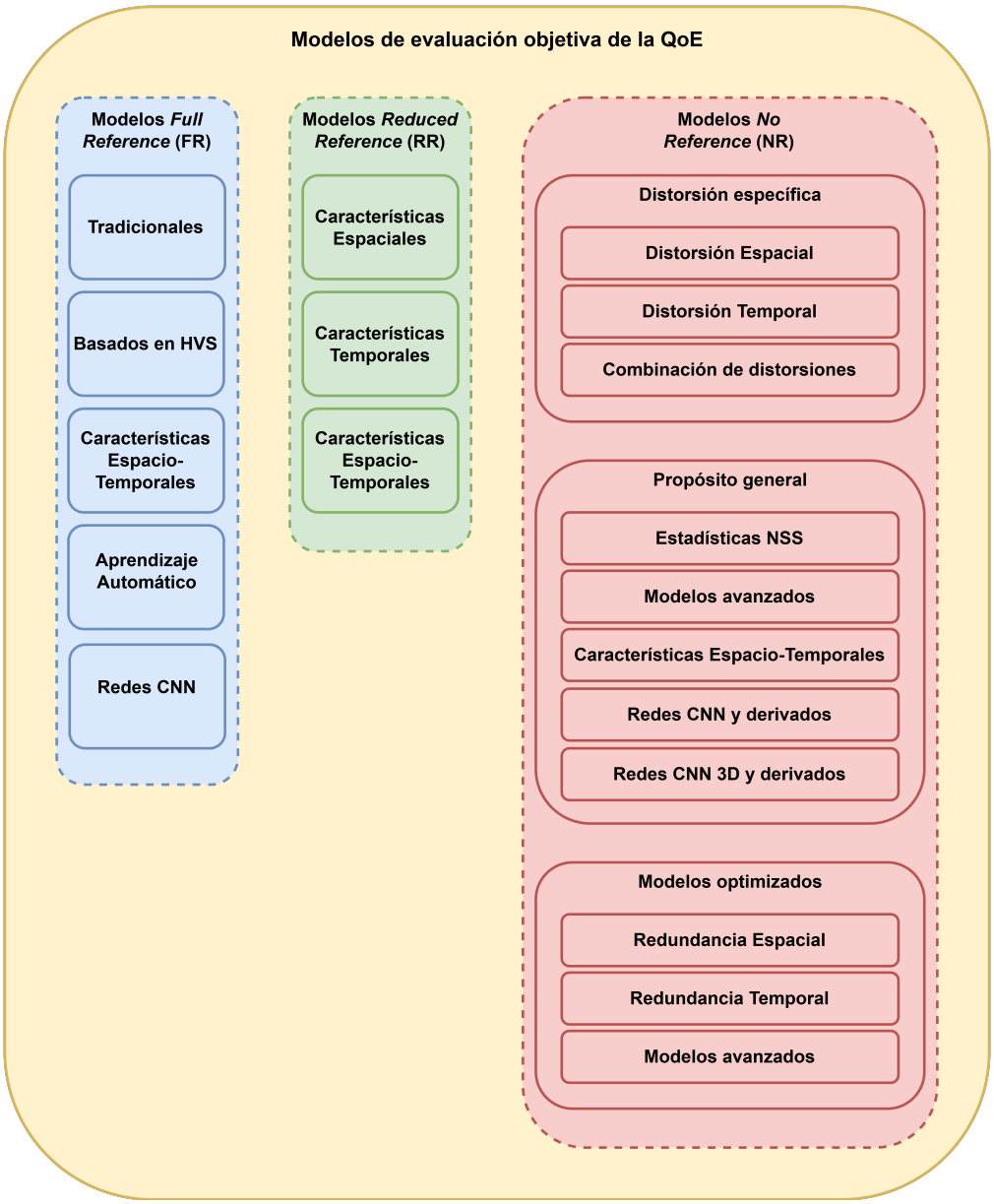

Figura 2.4: Modelos de evaluación objetiva de la QoE.

ni diversas, lo que impide la recopilación de información representativa para entrenar modelos capaces de identificar simultáneamente distintas distorsiones en los vídeos. Esta carencia limita la eficacia de los métodos de aprendizaje y reduce su capacidad de generalización [131]. Además, los modelos de entrenamiento tienden a experimentar problemas de sobreajuste (o *overfitting*), lo que restringe su desempeño óptimo a conjunto de datos específicos o a condiciones particulares. Esta limitación dificulta su aplicación en escenarios más amplios y heterogéneos, afectando su robustez y fiabilidad en entornos del mundo real.

Los modelos FR se basan en la cuantificación de las diferencias entre una señal de prueba y su correspondiente señal de referencia. El enfoque más básico para esta evaluación consiste en comparar directamente ambas señales, donde la calidad se determina a partir de la desviación

medida. Los modelos FR tradicionales, como MSE y PSNR [125], [126], analizan las diferencias entre imágenes de manera individual. Aunque estas métricas son ampliamente utilizadas por su simplicidad y facilidad de implementación, presentan una baja correlación con la percepción subjetiva de la calidad visual. Por otro lado, modelos que incorporan principios del HVS han demostrado una mejor correlación con las evaluaciones subjetivas. Ejemplos representativos de este enfoque incluyen las métricas de PSNR basado en HVS [132], SSIM [127], MS-SSIM (*Multi-Scale Structural Similarity Index Measure*) [133], VSNR (*Visual Signal to Noise Ratio*) [134], MAD (*Most Apparent Distortion*) [111], VIF (*Visual Information Fidelity*) [135], FSIM (*Feature Similarity Index Measure*) [136] y FMSE (*Foveated Mean Squared Error*) [137]. Estos modelos avanzados mejoran la evaluación objetiva de la calidad al integrar aspectos clave de la percepción visual, como la sensibilidad al contraste, la estructura espacial y la atención visual.

El ojo humano es especialmente sensible a las distorsiones inducidas por el movimiento de los objetos dentro de una escena. Dado el papel fundamental que el movimiento desempeña en la percepción visual y su impacto en la calidad percibida, se ha impulsado el desarrollo de modelos de evaluación espacio-temporal que consideran estos efectos de manera más precisa. Entre estos modelos avanzados se encuentran MOVIE [129], VIS3 (*Video Quality Assessment Via Analysis of Spatial and Spatiotemporal Slices*) [138], ST-MAD (*Spatio-Temporal Most Apparent Distortion*) [139], PVM (*Perceptual Video Metric*) [140] y FLOSIM-FR (*Flow-based Similarity Index Measure - Full Reference*) [141], los cuales emplean información del flujo óptico para modelar con mayor precisión las distorsiones asociadas al movimiento. Adicionalmente, el modelo FAST (*Full-reference Assessor Along Salient Trajectories*) [142] incorpora un análisis detallado de las trayectorias de los objetos, permitiendo una evaluación más realista de la calidad visual en escenarios dinámicos.

Las métricas FR más avanzadas incorporan técnicas de aprendizaje automático para mejorar la precisión en la estimación de la calidad del vídeo. Un ejemplo destacado es VMAF [128], una herramienta desarrollada por Netflix [3] [4], que combina múltiples características de VQA mediante un modelo de regresión basado en SVM (*Support Vector Machine*). Además del modelo VMAF, destacan otros modelos que emplean enfoques similares basados en aprendizaje automático, como el modelo ST-GREED (*Space-Time Generalized Entropic Difference*) [143] que utiliza un método de regresión basado en SVR (*Support Vector Regression*), y el modelo propuesto en [144], el cual implementa un algoritmo de regresión mediante *random forest*. Este último modelo correlaciona características clave como la textura, la saliencia, la información espacial y la información temporal con puntuaciones obtenidas en evaluaciones subjetivas, logrando una mejor aproximación a la percepción humana de la calidad visual.

Finalmente, los modelos FR más recientes incorporan el uso de redes neuronales convolucionales (CNN, *Convolutional Neural Networks*), para mejorar la evaluación de la calidad de vídeo. Ejemplos destacados con este enfoque incluyen los modelos DeepVQUE (*Deep Video Quality Evaluator*) [145], DeepVQA (*Deep Video Quality Assessor*) [146], C3DVQA (*Convolutional Neural Network with 3D kernels Video Quality Assessment*) [147], [148], [149], DISTS (*Deep Image Structure and Texture Similarity*) [150], DeepQA (*Deep Image Quality Assessment*) [151]

[3] https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12
[4] https://github.com/Netflix/vmaf

y CONTRIQUE-FR (*Contrastive Image Quality Evaluator - Full Reference*) [152]. Si bien estos modelos han demostrado un rendimiento superior en comparación con las métricas FR tradicionales, su eficacia sigue siendo limitada en la evaluación de distintos tipos de contenido con distorsiones específicas. Esta restricción, como ya se ha comentado anteriormente, se debe principalmente a la escasez de bases de datos subjetivas de entrenamiento, que dificulta la generalización de este tipo de modelos a una amplia variedad de escenarios y degradaciones visuales.

Los principios aplicados en los modelos FR también son relevantes para los modelos RR. Sin embargo, en los modelos RR, basta con disponer de información parcial de la señal de referencia para evaluar la calidad de la señal de prueba. Estos modelos pueden utilizar distintas fuentes de información, como información espacial, información temporal o información espacio-temporal, dependiendo del dominio en el que operen (dominio espacial, dominio temporal o una combinación de ambos). Algunos de los modelos RR más destacados incluyen: RRED (*Reduced Reference Entropic Differences*), TRRED (*Temporal Reduced Reference Entropic Differences*) y ST-RRED (*Spatio-Temporal Reduced Reference Entropic Differences*) [130], y el modelo SPEED QA (*Spatial Efficient Entropic Differencing for Quality Assessment*) [153].

Por otro lado, los modelos NR muestran un potencial significativo, ya que pueden estimar la calidad sin necesidad de una señal de referencia. Los modelos NR actuales se pueden clasificarse en dos grandes grupos: modelos basados en distorsiones específicas y modelos de propósito general.

- **Modelos NR basados en distorsiones específicas**. Estos modelos están diseñados para modelar artefactos visuales concretos que afectan a la calidad percibida de una señal audiovisual. Estos modelos se fundamentan en el análisis de información espacial, como la distorsión de píxeles o la detección de bordes [154]. Entre las principales distorsiones específicas estudiadas en la literatura se encuentran:
  - **Distorsión de *blocking*** [155], [156], [157], [158], [159]. Evalúa la presencia de discontinuidades entre bloques de píxeles generadas por el proceso de codificación del vídeo.
  - **Distorsión de *blurring*** [160], [161], [162], [163]. Mide la degradación en la nitidez de los contornos.
  - **Distorsión de *ringing*** [164], [165]. Identifica artefactos que aparecen alrededor de los bordes de la imagen derivados de la cuantización de coeficientes de alta frecuencia en la transformación DCT.
  - **Distorsión de *banding*** (bandas o anillamiento) [166] Analiza la aparición de transiciones abruptas en gradientes de color debido a una cuantificación insuficiente.
  - **Distorsión en la señal de luminancia en vídeos HDR** [167]. Se produce por limitaciones en los procesos de mapeo de tonos o por la compresión de señales con elevado rango dinámico.

  Por otro lado, diversos modelos se centran en la detección de distorsiones específicas que afectan la coherencia temporal en señales audiovisuales. En este ámbito, destacan

aquellos que analizan la pérdida de imágenes mediante la extracción de marcas de tiempo [168] o a través del estudio de discontinuidades entre imágenes consecutivas en una secuencia de vídeo [169]. Asimismo, se han desarrollado modelos enfocados a la detección de la distorsión de *jerkiness* [170], [171], [172], un fenómeno que altera la fluidez del movimiento en una secuencia de vídeo. Además, otros enfoques han abordado la detección de irregularidades en el movimiento mediante el análisis de diferencias en los coeficientes transformados de la DCT entre imágenes consecutivas de vídeo [173].

La principal limitación de los modelos NR basados en distorsiones específicas radica en su dependencia de los artefactos para los que fueron diseñados, lo que restringe mucho su aplicabilidad en escenarios donde múltiples distorsiones coexisten simultáneamente. Como consecuencia, diversas investigaciones han explorado el impacto de la combinación de distorsiones en la evaluación de la calidad visual. Entre estos estudios, se encuentran aquellos que analizan la interacción entre la distorsión de *blocking* y distorsión de *blurring* combinado con la distorsión de *ringing* [174], el impacto de las distorsiones de codificación en presencia de ruido de imagen [175], o la influencia combinada de artefactos de movimiento y la distorsión de *blocking* [176]. Además, otros trabajos han investigado la relación entre la pérdida de información durante la transmisión del contenido y la distorsión de *blocking* [177], [178].

- **Modelos NR de propósito general**. Estos modelos se fundamentan en la extracción de características visuales combinadas con técnicas de aprendizaje automático. Integran principios de percepción humana al incorporar atributos relacionados con el HVS, tales como la percepción del color, la sensibilidad estructural de la imagen y el enmascaramiento de texturas, entre otros. Además de los elementos perceptuales, estos modelos analizan características intrínsecas del contenido visual, incluyendo la nitidez, el brillo, el contraste, la composición cromática y la textura de las imágenes.

  La capacidad de estos modeles para extraer atributos relevantes mediante algoritmos de IA los hace significativamente más versátiles y generalizables en comparación con los enfoques basados en distorsiones específicas. Sin embargo, su aplicabilidad puede verse limitada por la naturaleza de los datos empleados durante su fase de entrenamiento, ya sea debido a la tipología del contenido, las características seleccionadas o la diversidad de valores de calidad presentes en los datos utilizados. Dado que ningún conjunto de datos supervisados puede abarcar la totalidad de las variaciones presentes en los vídeos del mundo real, los modelos NR de propósito general presentan restricciones inherentes a su capacidad de generalización [179].

  Entre los modelos más utilizados, destacan aquellos que emplean características de bajo nivel con relevancia perceptual, particularmente aquellas derivadas de los modelos de Estadísticas de la Escena Natural (NSS, *Natural Scene Statistics*) [180]. La premisa fundamental de los modelos NSS es que la degradación de una imagen natural altera sus propiedades estadísticas, lo que genera una percepción de artificialidad en la imagen resultante. Se han desarrollado con éxito modelos NSS de propósito general que analizan estas propiedades estadísticas en distintos dominios:

  – **Dominio espacial**. En este dominio, la representación de las imágenes se basa

en los valores de los píxeles según su posición en la matriz espacial. Este dominio permite analizar las características estructurales y la textura de una imagen sin necesidad de transformaciones adicionales. Ejemplos destacados de modelos que operan en este dominio incluyen: NIQE (*Natural Image Quality Evaluator*) [181] y BRISQUE (*Blind/Referenceless Image Spatial Quality Evaluator*) [182]. Ambos modelos emplean estadísticas NSS para evaluar la calidad de la imagen sin referencia.

- **Dominio del gradiente**. En este caso, las imágenes son representadas en función de los cambios de intensidad entre píxeles adyacentes, lo que permite detectar transiciones abruptas y bordes en la imagen. Modelos como GM-LOG (*Gradient Magnitude and Laplacian Of Gaussian*) [183], [184], HIGRADE (*HDR Image Gradient based Evaluator*) [185] exploran estas variaciones para evaluar la calidad visual.
- **Dominio de la DCT**. En este dominio, las imágenes son representadas en términos de las frecuencias de la señal, lo que permite capturar patrones de textura y compresión. Modelos como BLIINDS (*Blind Image Integrity Notator using DCT Statistics*) [186] y BLIINDS-II [187] aprovechan las propiedades estadísticas de la DCT para estimar la calidad visual de manera eficiente.
- **Dominio *wavelet***. Este enfoque representa las imágenes en diferentes niveles de resolución mediante una descomposición en componentes de alta y baja frecuencia. La representación en el dominio *wavelet* permite capturar estructuras multi-escala y detectar artefactos en distintos niveles de detalle. Modelos como BIQI (*Blind Image Quality Index*) [188] y DIIVINE (*Distortion Identification-based Image Verity and Integrity Evaluation*) [189] aplican este principio para evaluar la calidad de las imágenes.

Además de los modelos NR de propósito general basados en estadísticas NSS, existen otras metodologías que han demostrado un desempeño óptimo en la evaluación objetiva de calidad. Entre estos, se destacan los modelos FRIQUEE (*Feature Maps–based Referenceless Image Quality Evaluation Engine*) [190], CORNIA (*Codebook Representation for No-reference Image Assessment*) [191] y VIDEVAL (*Video Quality Evaluator*) [192]. El modelo FRIQUEE ha sido reconocido por su capacidad para evaluar con alta precisión imágenes afectadas por múltiples distorsiones auténticas, logrando un desempeño robusto en diferentes tipos de escenarios. Por su parte, el modelo CORNIA destaca por su eficiencia computacional, combinando efectividad y rapidez en la evaluación de calidad. En el ámbito del vídeo, el modelo VIDEVAL emplea una combinación de características espaciales de bajo nivel junto con propiedades estadísticas derivadas de modelos NSS, permitiendo la detección precisa de distorsiones en secuencias de vídeo.

Entre los modelos pioneros en la evaluación de calidad de vídeo sin referencia se encuentra el modelo VBLIINDS (*Video Blind Image Integrity Notator using DCT Statistics*) [193], el cual explora patrones estadísticos espacio-temporales en el dominio temporal para evaluar la coherencia estructural y el movimiento global mediante sofisticadas técnicas de estimación de movimiento. Adicionalmente, los modelos VIIDEO (*Video Intrinsic*

Integrity and Distortion Evaluation Oracle) [194] y 3D-DCT NR-VQA (*3D Discrete Cosine Transform No-Reference Video Quality Assessment*) [195] amplían el análisis mediante la explotación de regularidades estadísticas espacio-temporales, permitiendo una evaluación más precisa de vídeos con diversas distorsiones. El modelo STFC (*SpatioTemporal Feature Combination model*) [196], que logra un alto rendimiento en la evaluación de vídeos degradados con distorsiones reales. Finalmente, el modelo ChipQA (*Chip Quality Assessment*) [197] se distingue por emplear características espacio-temporales guiadas por el flujo de movimiento local, mejorando la estimación de calidad en secuencias de vídeo. No obstante, la extracción de características espacio-temporales puede implicar un elevado coste computacional. Para abordar esta limitación, el modelo TLVQM (*Two Level Video Quality Model*) [198] introduce un enfoque optimizado basado en un análisis jerárquico: calcula características de baja complejidad en todas las imágenes del vídeo, mientras que las características de mayor complejidad solo se extraen en un subconjunto específico de imágenes. Este enfoque permite una evaluación más eficiente sin comprometer excesivamente la precisión en la estimación de calidad.

En los últimos años, se han desarrollado diversos modelos NR de propósito general basados en CNN. Entre ellos, destacan los modelos PATCH VQ (*Patch Video Quality*) [199], MLSP VQA (*Multi-Level Spatially Pooled Deep-features Video Quality Assessment*) [200], GSTVQA (*Generalized Spatial-Temporal Deep Feature Representation for No-reference Video Quality Assessment*) [201], RankDVQA (*Ranking-Inspired Hybrid Training Deep VQA*) [131] y DEEPSTQ (*Deep SpatioTemporal Video Quality Assessor*) [202]. Asimismo, el modelo CNN-TLVQM (*Convolutional Neural Network - Two Level Video Quality Model*) [203] representa una evolución del modelo TLVQM, sustituyendo las características espacio-temporales de alta complejidad por aquellas extraídas mediante redes CNN, lo que permite una mejora en la precisión de la evaluación de calidad. Por otro lado, el modelo RAPIQUE (*Rapid and Accurate Video Quality Evaluator*) [204] logra un rendimiento significativo al combinar de manera eficiente las características espacio-temporales de la escena con aquellas obtenidas a través de redes CNN. En este contexto, los modelos VSFA [205] y MDTVSFA [206], este último como una versión optimizada del modelo VSFA, proponen la incorporación del efecto de dependencia del contenido y del efecto de memoria temporal en redes CNN. El primero de estos conceptos permite que la red neuronal adapte su evaluación en función de las características específicas del contenido analizado, mientras que el segundo concepto posibilita que la red neuronal retenga información sobre imágenes previas dentro de una secuencia de vídeo, mejorando así la coherencia en la estimación de la calidad percibida.

Además, algunos modelos recientes han empleado redes CNN tridimensionales con el fin de extraer y evaluar características espacio-temporales con mayor precisión. Ejemplos destacados de este enfoque son los modelos V-MEON (*Video Multi-task End-to-end Optimized Neural Network*) [207], STFEE (*SpatioTemporal Feature Extraction and Evaluation*) [208] y SACONVA (*Shearlet- and CNN-based NR VQA*) [209]. En una línea de investigación complementaria, el modelo DisCoVQA (*Distortion-Content Transformers for Video Quality Assessment*) [210] introduce el uso de arquitecturas basadas en transformadores para la evaluación de la calidad de vídeo, lo que permite un análisis más detallado de la evolución de las distorsiones a lo largo del tiempo en el

vídeo. Asimismo, el modelo COINVQ (*Comprehensive Interpretation Network for Video Quality*) [211] adopta un enfoque basado en redes CNN para examinar en profundidad la influencia del contenido, la calidad técnica y el nivel de codificación en la percepción de calidad de los vídeos UGC. Finalmente, el modelo propuesto en [212] se fundamenta en un enfoque de aprendizaje por transferencia, estrategia del aprendizaje automático que permite adaptar conocimientos adquiridos en un modelo preentrenado a nuevas tareas, optimizando así la extracción de información espacial y temporal en las escenas naturales.

A pesar de todos los avances logrados con estos modelos basados en redes neuronales profundas, aún persisten ciertas limitaciones inherentes a su implementación. En particular, la eficacia de estos modelos depende en gran medida del tamaño y diversidad de los conjuntos de datos utilizados en el entrenamiento, los cuales suelen ser insuficientes para abarcar la amplia variabilidad de degradaciones presentes en vídeos reales. Además, la extracción de características mediante redes CNN conlleva un elevado coste computacional, lo que dificulta (y prácticamente descarta) su aplicación en sistemas de evaluación de calidad en tiempo real.

Las investigaciones más recientes se han centrado en modelar de manera eficiente las características espacio-temporales en secuencias de vídeo, con el propósito de optimizar el rendimiento de los modelos NR. En este contexto, se ha explorado mucho la reducción del número de imágenes procesadas mediante técnicas de muestreo selectivo, bajo la premisa de que las imágenes consecutivas en una secuencia de vídeo contienen una gran cantidad de información redundante. En los estudios [213] y [214] se han desarrollado algoritmos destinados a minimizar la redundancia temporal mediante la selección de un subconjunto representativo de imágenes dentro de un vídeo. Esta estrategia permite reducir significativamente el coste computacional, en función del número de imágenes seleccionadas para el procesamiento y la extracción de características. Los resultados de estas investigaciones sugieren que un muestreo selectivo de imágenes puede ofrecer un rendimiento comparable al análisis de la totalidad de las imágenes de la secuencia, lo que evidencia el potencial de este enfoque en la estimación de calidad de vídeo.

Además de la reducción de la redundancia temporal mediante la selección de imágenes específicas en la secuencia, otras estrategias han sido propuestas en la literatura para disminuir los recursos computacionales de los modelos NR-VQA. Entre ellas, se encuentran aquellas que aprovechan la redundancia espacial en imágenes individuales, a través de la extracción de características en regiones específicas de la imagen o mediante el uso de versiones de la misma imagen a una menor resolución.

En este sentido, el modelo introducido en [215] implementa un mecanismo de selección de características espacio-temporales altamente relevantes para la evaluación de calidad. Por su parte, el modelo descrito en [216] combina la técnica de muestreo de imágenes con una estrategia de división de la imagen en regiones más pequeñas, asegurando así la conservación de la información global de toda la escena. De manera complementaria, el estudio presentado en [217] propone un enfoque que integra simultáneamente características locales y globales de la imagen, permitiendo la fusión de atributos específicos de áreas particulares con información global de la escena completa.

Tabla 2.1: Modelos de evaluación objetiva de la QoE.

Modelo	Clasificación	Características	Ejemplos
FR		Tradicional	MSE, PSNR ([125], [126])
		Basados en HVS	PSNR-HVS [132], SSIM [127], MS-SSIM [133], VSNR [134], MAD [111], VIF [135], FSIM [136], FMSE [137]
		Espacio-Temporal	MOVIE [129], VIS3 [138], ST-MAD [139], PVM [140], FLOSIM-FR [141], FAST [142]
		Aprendizaje automático	VMAF [128], ST-GREED [143], modelo *random forest* [144]
		CNN	DeepVQUE [145], DeepVQA [146], C3DVQA [147], modelo [148], modelo [149], DISTS [150], DeepQA [151], CONTRIQUE-FR [152]
RR		Espacial	RRED [130], SPEED QA [153]
		Temporal	TRRED [130]
		Espacio-Temporal	ST-RRED [130]
NR	Distorsión específica	Espacial	*Blocking* ([155], [156], [157], [158], [159]), *Blurring* ([160], [161], [162], [163]), *Ringing* ([164], [165]), *Banding* [166], luminancia HDR [167]
		Temporal	*Jerkiness*([170], [171], [172]), modelo de marcas de tiempo [168], modelo de discontinuidad temporal [169], modelo de movimiento irregular [173]
		Combinación	*Blocking*, *Blurring* y *Ringing* [174], *Blocking* y ruido de imagen [175], *Blocking* y artefactos de movimiento [176], *Blocking* y pérdida de información ([177], [178])
	Propósito general	Estadísticas NSS	Dominio espacial (NIQE [181], BRISQUE [182]), dominio gradiente (GM-LOG [183], modelo [184], HIGRADE [185]), dominio DCT (BLIINDS [186], BLIINDS-II [187]), dominio *wavelet* (BIQI [188], DIIVINE [189])
		Modelos avanzados	FRIQUEE [190], CORNIA [191], VIDEVAL [192]
		Espacio-Temporal	VBLIINDS [193], VIIDEO [194], 3D-DCT NR-VQA [195], STFC [196], ChipQA [197], TLVQM [198]
		CNN y derivados	PATCH VQ [199], MLSP VQA [200], GSTVQA [201], RankDVQA [131], DEEPSTQ [202], CNN-TLVQM [203], RAPIQUE [204], VSFA [205], MDTVSFA [206]
		CNN 3D y derivados	V-MEON [207], STFEE [208], SACONVA [209], DisCoVQA [210], COINVQ [211], modelo [212]
	Optimización	Redundancia Espacial	Modelo de selección de características de imagen [215], modelo de selección de características locales y globales de la imagen [217], modelo de división de la imagen [216]
		Redundancia Temporal	Modelos de muestreo selectivo de imágenes([213], [214])
		Modelos avanzados	Zoom-VQA [218], FAST-VQA [219], DOVER [220]

En una línea similar, el modelo Zoom-VQA [218] introduce una arquitectura sofisticada diseñada para la detección de características espacio-temporales en múltiples niveles, optimizando el análisis de información tanto local como global en la imagen completa y en regiones de interés. De la misma manera, FAST-VQA (*Fragment Sample Transformer for VQA*) [219] incorpora una técnica de muestreo de vídeo orientada a preservar la calidad, evitando métodos convencionales como la reducción de resolución o el recorte de imagen, y en su lugar, opera

sobre regiones representativas de la escena. Finalmente, el modelo DOVER (*Disentangled Objective Video Quality Evaluator*) [220] propone un esquema basado en dos evaluaciones de calidad independientes, los cuales emplean submuestreo espacial junto con técnicas de muestreo de imágenes para extraer información semántica y contextual del vídeo, mejorando así la capacidad del modelo para comprender mejor los elementos visuales y sus interrelaciones.

Las estrategias de reducción de redundancia tanto espacial como temporal adoptadas en estos modelos NR-VQA descritos anteriormente han demostrado mejoras sustanciales en términos de eficiencia, rendimiento y coste computacional. En consecuencia, estas técnicas constituyen un punto de partida prometedor para optimizar la herramienta NR-VQA de Video-MOS, facilitando su implementación en entornos y escenarios donde la eficiencia computacional sea un factor crítico para poder funcionar en tiempo real.

Finalmente, la Tabla 2.1 presenta de forma resumida los principales modelos de evaluación objetiva de la QoE mencionados en la Sección 2.4.

Capítulo 3

Material y métodos

Una secuencia de vídeo está compuesta por una serie de imágenes consecutivas que, en su mayoría, presentan una alta similitud entre sí. De manera análoga, dentro de cada una de las imágenes individuales, cada píxel suele mantener cierta correlación con los píxeles adyacentes, estableciéndose patrones de continuidad espacial.

Al igual que los codificadores de vídeo emplean técnicas de reducción de redundancia espacial y temporal para comprimir la señal y minimizar la cantidad de datos requeridos para la transmisión o el almacenamiento, los enfoques de optimización propuestos en esta tesis doctoral se fundamentan en la explotación de dichas redundancias con el objetivo de reducir el coste computacional en la evaluación de la QoE.

3.1 Herramienta de prueba

En esta investigación, la medida de QoE se obtiene a partir del valor de MOS estimado por la herramienta NR-VQA de Video-MOS. La elección de esta solución para el desarrollo de este estudio se justifica (a parte de por el acuerdo de colaboración con la empresa Video-MOS por medio de la cátedra universitaria), por dos razones principales:

- **Análisis de características espaciales y temporales**. La solución Video-MOS evalúa la calidad visual percibida a través de un conjunto de características espacio-temporales presentes en una secuencia de vídeo. Esta capacidad de extracción de información permite la implementación de técnicas de optimización basadas en la reducción de redundancia espacial y temporal de los vídeos.
- **Acceso completo al código fuente de la herramienta**. Gracias a la colaboración establecida entre la empresa Video-MOS y la UPM, formalizada mediante una cátedra universitaria, se garantiza el acceso total al código fuente de la herramienta. Este acceso facilita la implementación y despliegue rápido de los distintos enfoques de optimización propuestos en la investigación.

Para la fase de pruebas, se emplea la versión de desarrollo de la solución Video-MOS. Esta versión incluye todas las funcionalidades de la herramienta comercial y proporciona resultados

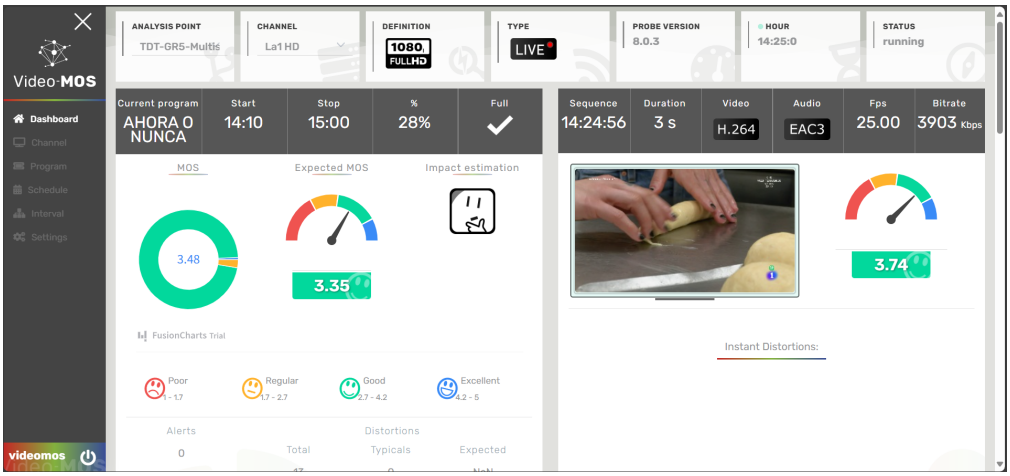

Figura 3.1: Captura de la solución comercial Video-MOS.

equivalentes en términos de extracción de características espacio-temporales, detección de distorsiones específicas y estimación del valor de MOS. No obstante, la principal diferencia entre ambas versiones radica en el lenguaje de implementación: mientras que la versión comercial está desarrollada en C++ [1] y se despliega en contenedores *software* con Docker [2], la versión de desarrollo utiliza lenguaje de alto nivel interpretado Python [3]. Aunque la herramienta de desarrollo presenta un menor rendimiento en términos de eficiencia computacional, su flexibilidad y facilidad para realizar modificaciones sin necesidad de un proceso de compilación la convierten en la opción idónea para el estudio planteado en la tesis doctoral. Cabe destacar que cualquier mejora obtenida mediante la herramienta de desarrollo podrá aplicarse posteriormente a la versión comercial de Video-MOS y, potencialmente, a otras herramientas de evaluación de QoE basadas en la extracción de características espaciales y temporales del vídeo. La Figura 3.1 muestra una captura de la herramienta comercial Video-MOS, analizando un canal de televisión HD emitido en la TDT en España.

La solución Video-MOS consiste en un modelo NR-VQA híbrido, que combina un enfoque basado tanto en distorsiones específicas de vídeo como en técnicas de propósito general. Su metodología se sustenta en la extracción de un conjunto de características del vídeo con el fin de estimar la calidad percibida por un usuario promedio. A partir de un conjunto de parámetros y métricas sin referencia de vídeo, la herramienta emplea un modelo de regresión no lineal basado en IA para procesar toda la información. Las técnicas de aprendizaje automático integradas en la herramienta permiten tanto la detección de distorsiones específicas como la estimación del valor numérico de calidad en escala MOS, siguiendo la recomendación UIT-R BT.500 [107]. La Tabla 3.1 presenta los principales parámetros y características utilizados por la herramienta Video-MOS en la evaluación de QoE.

La herramienta de desarrollo Video-MOS se basa en el uso de diferentes programas, paquetes de Python y bibliotecas de código abierto, fundamentales para el análisis del vídeo y el

[1] https://isocpp.org/
[2] https://www.docker.com/
[3] https://www.python.org/

Tabla 3.1: Extracción de parámetros y características de la solución Video-MOS.

Tipo	**Parámetros**
Metadatos de vídeo	Resolución de imagen Tasa de refresco de imagen Tipo de escaneo de imagen Codificador de vídeo Tasa binaria Profundidad de bit Submuestreo de color Espacio de color
Métricas NR de vídeo	*Spatial Information* (Información espacial) *Temporal Information* (Información temporal) *Brightness* (Información de brillo) *Contrast* (Información de contraste)
Métricas de distorsión	*Blurring* *Ringing* *Blockloss* (Pérdida de información) *Blocking*
Alarmas de distorsión de contenido	Pérdida de imágenes Pérdida del contenido Pérdida de señal Congelaciones Nivel de contraste Nivel de saturación Sobre-exposición Sub-exposición

procesamiento de imágenes. En particular, emplea el paquete de Python PyAV [4] para la captura del contenido audiovisual y la gestión del flujo de los paquetes de audio y de vídeo, la herramienta FFmpeg [5] para la extracción de los metadatos de audio y vídeo de la señal multimedia, y la biblioteca OpenCV para Python [6] para el procesamiento de las imágenes y la extracción de las características visuales. A continuación, se describe el funcionamiento detallado de la herramienta:

1. **Captura del contenido audiovisual y extracción de metadatos de vídeo**. La herramienta Video-MOS extrae parámetros fundamentales de la señal de vídeo, tales como la resolución, la tasa de refresco de imagen, el estándar de codificación del vídeo, la tasa binaria, la profundidad de bits, el submuestreo de crominancia y el espacio de color, entre otros.
2. **Decodificación y procesamiento de las imágenes del vídeo**. Durante este proceso se lleva a cabo la decodificación de las imágenes del vídeo, y se extraen los metadados a

[4] https://pyav.org/docs/stable/
[5] https://www.ffmpeg.org/
[6] https://opencv.org/

Tabla 3.2: Métricas de evaluación de calidad de Video-MOS.

Métrica de evaluación	**Características**	**Categoría**
Spatial Information	Información de altas frecuencias y nivel de detalle	A nivel de imagen
Temporal Information	Información de velocidad y movimiento	A nivel de píxel
Blurring	Información de la pérdida o ausencia de altas frecuencias	A nivel de imagen
Brightness	Información del nivel medio de las diferentes señales	A nivel de píxel
Contrast	Información del nivel de dispersión de las diferentes señales	A nivel de píxel
Ringing	Artefactos de imagen	A nivel de imagen
Blockloss	Artefactos de imagen	A nivel de imagen
Blocking	Artefactos de imagen	A nivel de imagen

nivel de imagen, incluyendo el tipo de cuadro de codificación (generalmente imágenes tipo I, tipo P o tipo B), el tipo de escaneo de imagen (progresivo o entrelazado), las marcas de tiempo de decodificación (DTS, *Decoding Time Stamp*) y de presentación (PTS, *Presentation Time Stamp*) de las imágenes. Además, utilizando funciones de la biblioteca OpenCV, la herramienta realiza las conversiones de colorimetría necesarias para obtener las componentes en diferentes espacios de color en cada imagen: RGB, YCbCr y HSV (*Hue, Saturation, Value*), permitiendo así una evaluación detallada de la calidad de imagen a través de la incorporación de parámetros de distintos espacios de color.

3. **Cálculo de métricas NR de vídeo y métricas de distorsión**. La herramienta implementa un conjunto de métricas NR para la extracción de características espacio-temporales de la secuencia de vídeo (*Spatial Information*, *Temporal Information*, *Brightness*, *Contrast*) y métricas de distorsión (*Blurring*, *Ringing*, *Blockloss*, *Blocking*). Este conjunto de ocho métricas forma parte del sistema de evaluación de calidad implementado en Video-MOS y será utilizado en el plan de pruebas desarrollado en el marco de esta tesis doctoral. La Tabla 3.2 resume las ocho métricas de evaluación de calidad de Video-MOS, y sus características principales para la evaluación de QoE.

4. **Detección de alarmas de distorsión de contenido**. La herramienta incorpora un módulo específico para la detección de anomalías en el vídeo. A partir de las métricas de evaluación de calidad, la solución de Video-MOS es capaz de identificar diversas distorsiones de contenido, tales como la pérdida de imágenes individuales, interrupciones de señal, congelación de imagen, niveles anómalos de contraste o saturación, así como situaciones de sobre-exposición o sub-exposición en la escena, entre otras.

5. **Estimación de la calidad mediante el valor de MOS**. La herramienta calcula la calidad visual percibida por un usuario medio a partir de un vector de características compuesto por los metadatos del vídeo y la métricas de evaluación de calidad de Video-MOS: *Spatial Information*, *Temporal Information*, *Blurring*, *Brightness*, *Contrast*, *Ringing*, *Blockloss* y *Blocking*. Para ello, la herramienta emplea un módulo de predicción basado en un regresor no lineal de IA, el cual estima un valor numérico en escala de

MOS de 1 a 5. La estimación de calidad se realiza en intervalos de medida definidos por el usuario de la herramienta, siendo el valor típico (o por defecto) intervalos de tiempo de tres segundos de duración.

El proceso de solicitud de patente en trámite bajo el número Nº202331007 en la Oficina Española de Patentes y Marcas, junto con la estrategia de comercialización de la herramienta Video-MOS, restringe la divulgación y la publicación de cualquier código asociado a la solución, tanto en su versión comercial como en su versión de desarrollo. Esta medida responde a la necesidad de proteger la propiedad intelectual y garantizar la exclusividad de la tecnología, evitando la difusión de información que pudiera comprometer su carácter innovador y su competitividad en el mercado audiovisual. Como se ha comentado anteriormente, la solución completa de Video-MOS está protegida mediante el registro de cuatro módulos *software* en el Registro Territorial de la Propiedad Intelectual de la Comunidad de Madrid (identificadores M-002018/2023, M-002033/2023, M-002037/2023 y M-002039/2023).

3.2 Secuencias de vídeo de prueba

El material utilizado para el plan de pruebas consta de 1123 secuencias de vídeo de tres segundos de duración. Además, se han empleado más de 84000 imágenes individuales extraídas a partir de las secuencias de vídeo de prueba.

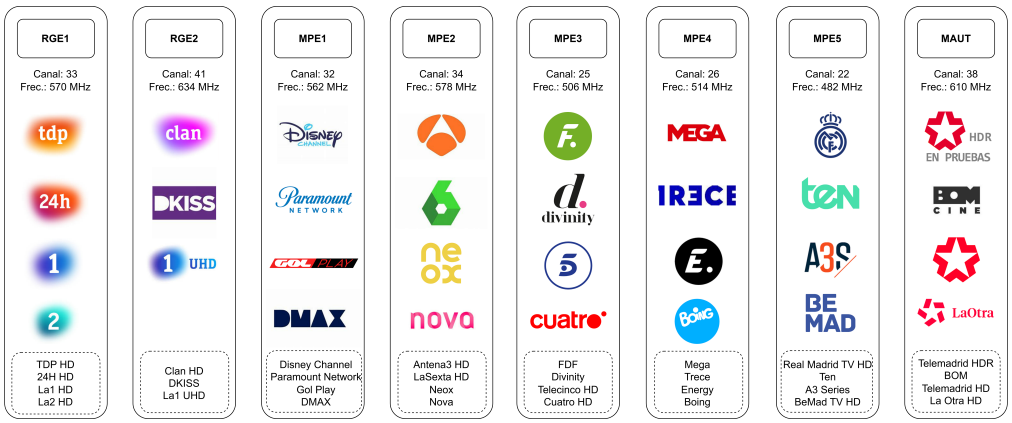

Figura 3.2: Canales de televisión de la TDT en España en la Comunidad de Madrid.

Las secuencias han sido obtenidas directamente de la TDT en España, utilizando equipamiento profesional para la sintonización y grabación de los múltiples de TDT RGE1 y RGE2. Estos dos múltiples son gestionados por el radiodifusor público español RTVE (Radiotelevisión Española) [7], que opera diversos canales de televisión en España. El uso de estos contenidos en la tesis doctoral cuenta con el permiso explícito de RTVE, gracias al convenio de colaboración establecido entre RTVE y la UPM en forma de cátedra universitaria desde el año 2015 [8]. La Figura 3.2 muestra los diferentes canales de televisión en cada uno de los múltiples de la TDT

[7] https://www.rtve.es/
[8] https://catedra.rtve.etsit.upm.es/

Tabla 3.3: Características técnicas de las secuencias de prueba.

Parámetro	**Valor**
Resolución	1920x1080
Relación de aspecto	16:9
Tasa de refresco	25 imágenes por segundo
Tipo de escaneo	Entrelazado
Submuestreo de color	YCbCr 4:2:0, 8 bits
Espacio de color	ITU-R BT.709 [35]
Codificación de vídeo	H.264/MPEG-4 AVC [71]

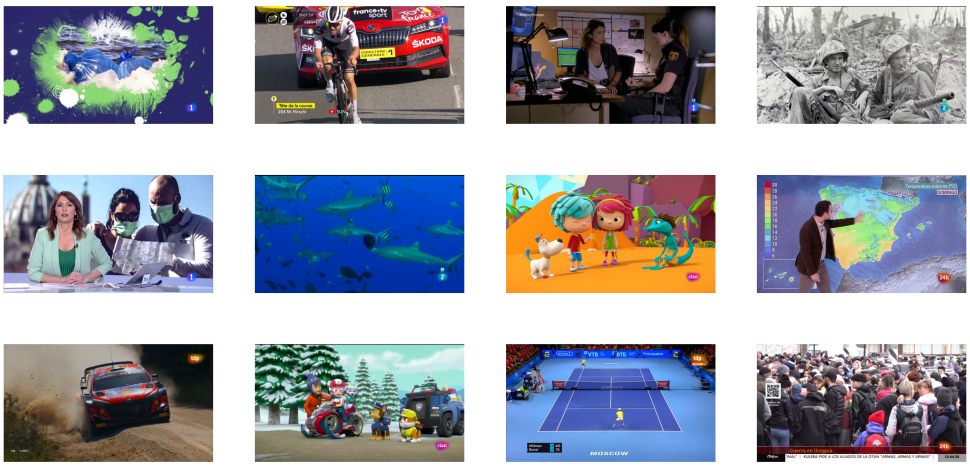

Figura 3.3: Capturas de las secuencias de prueba.

en España, para la comunidad de Madrid. Las secuencias de vídeo de prueba pertenecen a los canales TDP HD, 24H HD, La1 HD, La2 HD y Clan HD.

El formato de las secuencias de vídeo cumple con las especificaciones técnicas de la emisión TDT en HD en España, conforme al Plan Técnico Nacional de la Televisión Digital Terrestre, regulado por el Real Decreto 391/2019 (BOE nº 151, del 25 junio de 2019) [9]. Las principales características de las secuencias se resumen en la Tabla 3.3.

El material de prueba presenta una gran diversidad en cuanto al tipo de contenido. Incluye contenido sintético con gráficos, contenido antiguo en blanco y negro, documentales, programas informativos tanto en entornos interiores como en entornos exteriores, eventos deportivos, concursos televisivos, series y películas. La Figura 3.3 muestra algunas capturas de los vídeos de prueba utilizados.

[9] https://www.boe.es/diario_boe/txt.php?id=BOE-A-2019-9513

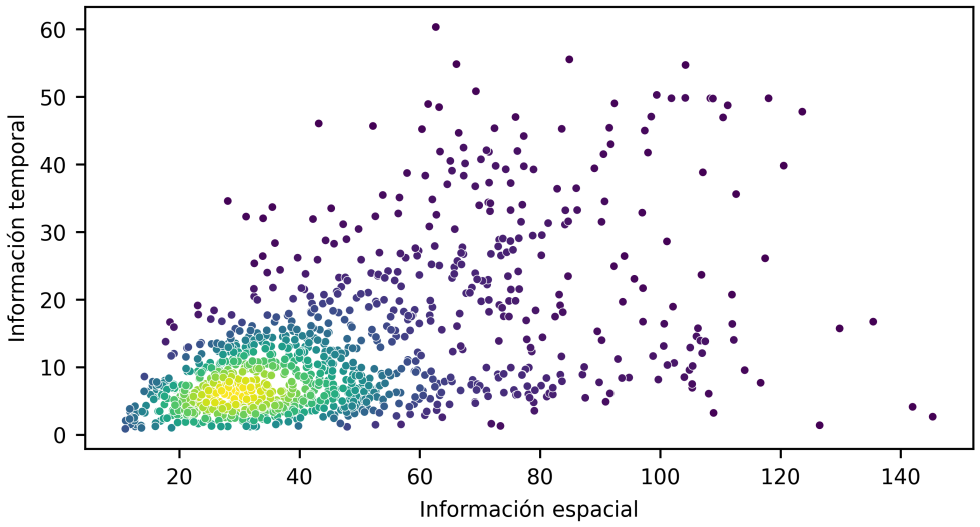

Figura 3.4: Diagrama SI-TI de las secuencias de prueba.

La variabilidad del conjunto de secuencias también se refleja en la diversidad de los valores de Información Espacial (SI, *Spatial Information*) e Información Temporal (TI, *Temporal Information*). En la Figura 3.4 se presenta el diagrama SI-TI correspondiente a los datos obtenidos con las 1123 secuencias de vídeo. Los valores de SI y TI se han calculado siguiendo las fórmulas especificadas en la recomendación UIT-T P.910 [108], edición 4.0 de noviembre de 2021.

Los valores de las métricas de evaluación de vídeo obtenidos con la herramienta Video-MOS sobre el conjunto de prueba también reflejan una gran variabilidad en el tipo de contenido. En la Figura 3.5 y Figura 3.6 se representan los valores de las distintas métricas de vídeo. La separación en dos figuras responde a la diferencia en los rangos y valores entre las métricas.

Para la representación de los datos en la Figura 3.5 y Figura 3.6 se ha utilizado un diagrama de caja (o *boxplot*) [10]. Este tipo de gráfico proporciona una visión clara de la distribución estadística y la dispersión de los datos, facilitando también la identificación de los valores atípicos (o *outliers*). En este tipo de representación, la *caja* representa el rango intercuartílico (IQR, *Interquartile Range*), definido como la diferencia entre el primer cuartil (Q1, percentil 25 %) y el tercer cuartil (Q3, 75 %). Además:

- El borde inferior del *boxplot* indica el primer cuartil (Q1), donde el 25 % de los datos presentan valores inferiores y el 75 % valores superiores.
- El borde superior representa el tercer cuartil (Q3), donde el 75 % de los datos tienen valores menores y el 25 % valores mayores.
- La línea dentro de la caja corresponde al valor de mediana (Q2, percentil 50 %), que divide la distribución en dos mitadas completamente iguales.
- Los *bigotes* del diagrama se extienden hasta 1.5 veces el rango IQR desde los bordes

[10] https://seaborn.pydata.org/generated/seaborn.boxplot.html

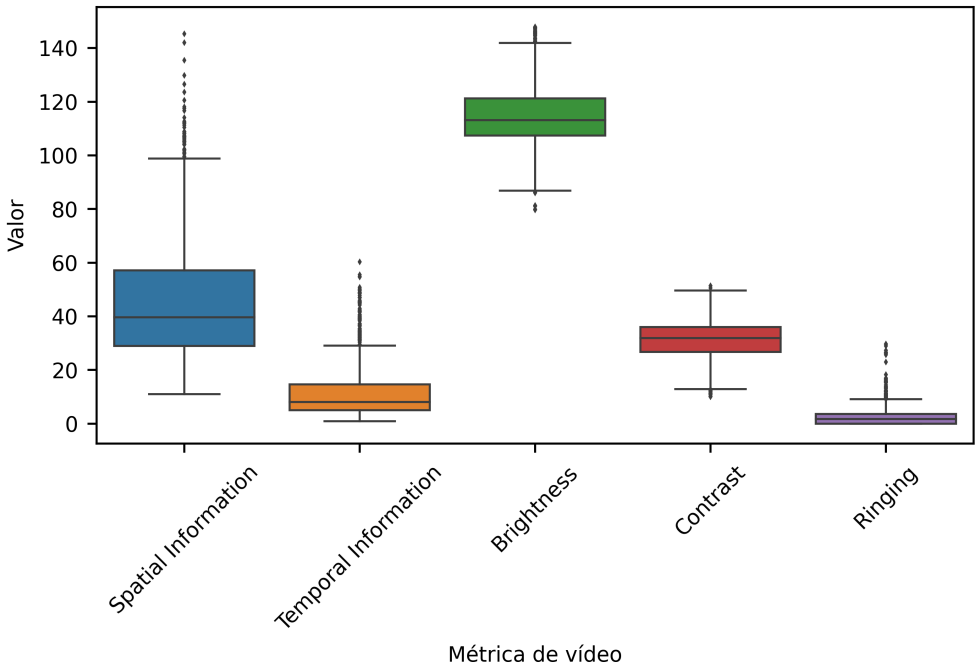

Figura 3.5: Valores de las métricas de vídeo de las secuencias de prueba (1).

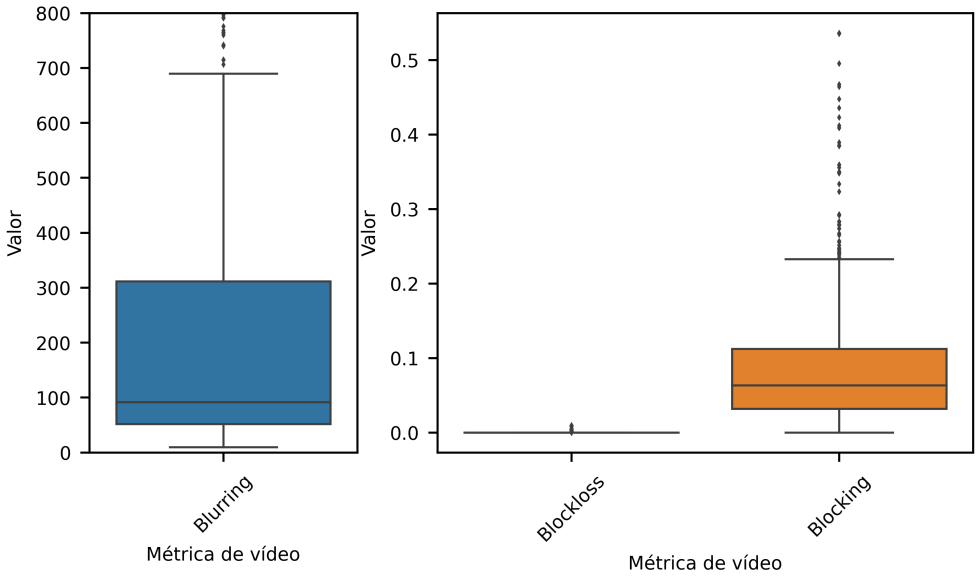

Figura 3.6: Valores de las métricas de vídeo de las secuencias de prueba (2).

de la caja. Los valores que exceden estos límites se consideran valores atípicos y se representan como puntos individuales en la representación gráfica.

El rango de cada métrica de vídeo se ha determinado calculando la diferencia entre el valor máximo (tercer cuartil, Q3) y mínimo (primer cuartil, Q1), excluyendo los valores atípicos. No obstante, en el caso de la métrica de vídeo de *Blockloss*, se ha considerado mantener los

Tabla 3.4: Rango de las métricas de vídeo de las secuencias de prueba.

Métrica de evaluación	Valor máximo (Q3)	Valor mínimo (Q1)	Valor medio	Rango
Spatial Information	98.82	11.03	42.93	87.79
Temporal Information	29.07	0.87	9.45	28.20
Blurring	689.28	9.58	128.47	679.70
Brightness	141.91	86.87	114.18	55.04
Contrast	49.67	12.93	31.51	36.74
Ringing	9.08	0.00	2.01	9.08
Blockloss	0.01	0.00	0.00	0.01
Blocking	0.23	0.00	0.07	0.23

valores atípicos para evitar un rango de valor cero. Esta decisión se debe a que la métrica de *Blockloss* mide características relacionadas con la pérdida de información y artefactos de imagen, poco frecuente en las emisiones de TDT. La Tabla 3.4 recoge el rango de valores de las distintas métricas de vídeo utilizadas en la evaluación de los vídeos de prueba.

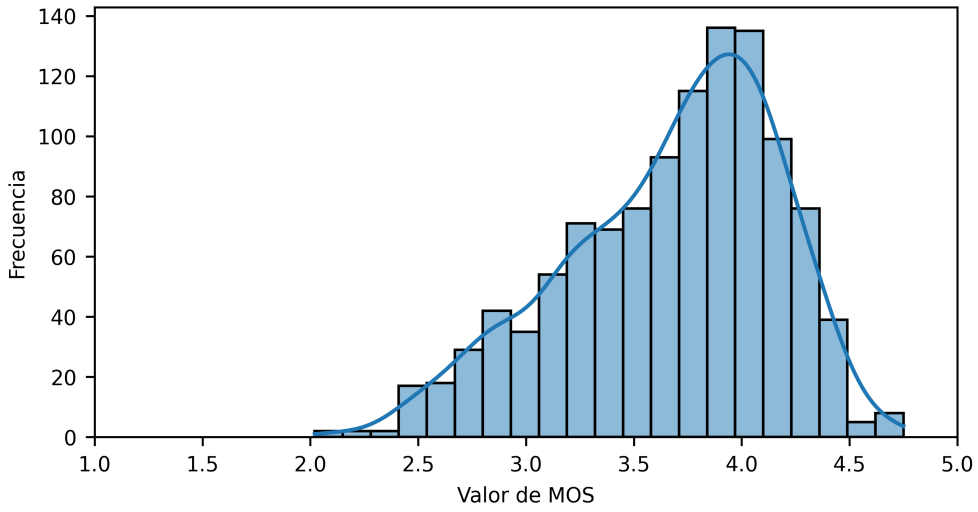

Figura 3.7: Histograma valor de MOS de las secuencias de prueba.

Finalmente, las secuencias de prueba también presentan variabilidad amplia en la calidad percibida, según los resultados obtenidos con la herramienta Video-MOS. En la Figura 3.7, se muestra el histograma de los valores de MOS correspondientes a todas las secuencias de prueba analizadas. El valor medio de MOS en el conjunto de prueba es de 3.67, con un valor máximo de 4.75 y un valor mínimo de 2.02. El valor medio de MOS se encuentra por encima del valor umbral de 3, que en la escala de MOS se considera como *Aceptable*, de acuerdo con la recomendación UIT-R BT.500 [107]. Además, este valor medio obtenido está por encima del

límite de calidad mínima para señal HD en televisión, según lo establecido por la organización internacional EBU (*European Broadcasting Union*) en la norma EBU R132 [221].

Los resultados muestran que la mayoría de las secuencias de vídeo cumplen con los estándares de calidad requeridos para emisiones de contenidos HD en la TDT. Solo el 11.67 % de las secuencias, equivalente a 131 de los 1123 vídeos, presentan un valor de MOS inferior a 3.

3.3 Equipo de prueba

El equipo utilizado para la realización de las pruebas en esta tesis doctoral presenta las especificaciones técnicas recogidas en la Tabla 3.5. Este equipo coincide con el sistema físico base que la empresa Video-MOS emplea tanto para la realización de pruebas como para la comercialización de la solución.

Tabla 3.5: Especificaciones técnicas del equipo de prueba.

Recurso	**Especificación**
Dispositivo	MSI
Procesador	12th Gen Intel(R) Core (TM) i7-12700H 2.70 GHz
Memoria RAM instalada	32.0 GB (31.7 GB útil)
Sistema	Sistema operativo 64-bit, procesador x64
Especificaciones de Windows	Windows 11 Pro

Utilizando este equipo y la herramienta de desarrollo de Video-MOS descrita en la Sección 3.1, se ha evaluado el rendimiento en la extracción de las características de las métricas de evaluación de vídeo, sobre el conjunto de pruebas detallado en la Sección 3.2 y compuesto por más de 84000 imágenes individuales.

Los resultados muestran que el tiempo medio requerido para extraer todas las características de una única imagen es de 0.5430 s. Por lo tanto, el tiempo total medio necesario para procesar una medida de tres segundos de vídeo asciende a 40.7250 s, un valor muy por encima del límite de procesamiento para funcionar en tiempo real. Para que la solución de Video-MOS pudiera funcionar en tiempo real en el equipo de prueba, sería necesario reducir el coste computacional en aproximadamente un 92.63 %, teniendo en cuenta únicamente el tiempo consumido en la extracción de características.

En la Tabla 3.6 se muestra el tiempo de procesamiento medio empleado por cada una de las métricas de vídeo en la extracción de las características, sobre el conjunto de imágenes de prueba. Se observa que, debido a la complejidad computacional de los algoritmos, las métricas de *Blockloss* y de *Blocking* representan más del 71.5 % del tiempo total de procesamiento entre el conjunto total de las métricas de vídeo.

Además, la Tabla 3.6 también incluye la medida de desviación estándar (STD, *Standard*

Tabla 3.6: Tiempo de procesamiento medio por imagen para las métricas de vídeo.

Métrica de evaluación	Tiempo medio (ms)	Peso (%) de Tiempo medio	STD de Tiempo medio (ms)
Spatial Information	46.36	8.54	9.81
Temporal Information	25.02	4.61	7.75
Blurring	20.15	3.71	6.29
Brightness	8.14	1.50	5.75
Contrast	44.65	8.22	10.35
Ringing	10.31	1.90	18.69
Blockloss	156.35	28.79	35.77
Blocking	232.04	42.73	158.39
Total	**543.03**	**100**	-

Deviation) del tiempo de procesamiento de cada una de las métricas. Se observa que las métricas que funcionan con valores de píxel (ver Tabla 3.2), como la métrica de *Brightness* y de *Contrast*, presentan un valor de STD menor. Por el contrario, métricas como *Ringing*, *Blockloss* y *Blocking* presentan mayores variaciones en el tiempo medio de procesamiento. Esto se debe a que estas métricas, basadas en la estructura de la imagen, dependen de características específicas de la escena (como bordes y texturas), lo que introduce una variabilidad en su procesamiento.

3.4 Plan de pruebas

Esta investigación establece un conjunto de pruebas orientadas a optimizar el coste computacional de la herramienta Video-MOS, garantizando la precisión en la extracción de características de vídeo y en la estimación del valor de MOS proporcionado por la solución.

La evaluación del coste computacional de un programa o proceso informático es una tarea compleja de medir, ya que diversos factores pueden influir en el rendimiento del sistema. Entre estos factores se encuentran la ejecución de procesos en segundo plano, el nivel de carga de la batería, las configuraciones de ahorro de energía, la disponibilidad de memoria, la temperatura del dispositivo y el uso de tarjeta gráfica integrada, entre otros. Aunque se reconoce la relevancia de considerar todos estos factores en la evaluación del coste computacional, la dificultad para medir con precisión algunos de ellos durante la realización de las pruebas justifica que sean excluidos del análisis final de los resultados.

En las distintas pruebas realizadas en esta investigación, la evaluación del coste computacional no se basa en el tiempo de procesamiento del vídeo, sino en datos relacionados (en la extracción de características) con la cantidad de píxeles analizados por imagen y el número de imágenes procesadas dentro de una medida de tres segundos de vídeo.

En lo que respecta al tamaño de la imagen, se considera un coste computacional del 100 % cuando la extracción de características se realiza sobre una imagen con resolución original

de 1920x1080 píxeles. De manera análoga, procesar la totalidad de las imágenes dentro de una medida de tres segundos (equivalente a procesar 75 imágenes en vídeos con una tasa de refresco de 25 imágenes por segundo), también supone un coste computacional del 100 %. Como referencia, la reducción de la resolución de una imagen a 960x540 píxeles implicaría un coste computacional del 25 % (ahorro de 75 %), mientras que el análisis de únicamente 15 imágenes a resolución 1920x1080 en tres segundos en un vídeo a 25 imágenes por segundo supondría un coste computacional del 20 % (ahorro computacional del 80 %).

Para optimizar la herramienta de desarrollo de Video-MOS y garantizar el funcionamiento en tiempo real en el equipo de prueba, es imprescindible alcanzar un coste computacional en la extracción de características inferior al 7.37 % (ahorro de coste computacional del 92.63 %, según las pruebas realizadas en la Sección 3.3). La herramienta de Video-MOS tiene un coste computacional del 100 % en su modo de funcionamiento normal, al procesar la totalidad de las imágenes de la secuencia de vídeo con su resolución original de 1920x1080 píxeles.

Además de la optimización del coste computacional, esta investigación también busca identificar el enfoque que minimice el error en la extracción de características y el error en la estimación de QoE con el valor de MOS. Para la evaluación del rendimiento de cada uno de los enfoques de optimización, se emplea la métrica de error MAE (*Mean Absolute Error*), una de las métricas más utilizadas en la medición de precisión de modelos de predicción, junto con la métrica RMSE (*Root Mean Square Error*). Ambas métricas son muy utilizadas en problemas de regresión, ya que permiten cuantificar el error entre los valores estimados y los valores reales. El MAE es el valor medio del valor absoluto de la diferencia entre valor estimado y valor real (el error). El RMSE es el valor medio del error cuadrático. Si los errores se distribuyen de manera Gaussiana, optimizar el RMSE es una estimación de máxima verosimilitud [222], pero esto supone asumir la distribución de errores. Al ser un valor cuadrático es muy sensible a valores anómalos o extremos, mientras que el MAE penaliza todos los errores de manera lineal. Esto hace que el MAE sea más directamente interpretable [223], [224], y es por ello por lo que se elige optimizar los modelos y evaluar los errores empleando esta métrica.

De manera análoga al objetivo de alcanzar un ahorro de coste computacional superior al 92.63 % para permitir la ejecución de la solución Video-MOS en tiempo real en el equipo de prueba, se ha definido el error máximo aceptable en la estimación de MOS para los enfoques de optimización propuestos. Concretamente, se considera que un valor de error de MOS (en términos MAE) inferior a 0.15 es adecuado, lo que equivale a un margen de error inferior a un 3 % en la escala MOS. Este requisito ha sido establecido por la propia empresa Video-MOS.

En las pruebas realizadas, además del análisis del error en la estimación de MOS, también se evalúa el error en la extracción de características (en términos MAE), con el fin de poder cuantificar la desviación de los enfoques de optimización propuestos respecto al modo de funcionamiento normal de la herramienta. Para obtener una medida relativa del error en la extracción de características que no dependa del rango específico de cada una de las métricas de evaluación de vídeo, el MAE se expresa en forma de porcentaje. Así, el error MAE de cada una de las métricas de vídeo se calcula como un valor porcentual en función de su rango (ver Tabla 3.4). La Ecuación 3.1 recoge el cálculo de MAE(%) en función del valor MAE y del valor de RANGO de la métricas de evaluación de vídeo.

$$\text{MAE}\ (\,\%) = \left(\frac{\text{MAE}}{\text{RANGO}}\right) \times 100 \tag{3.1}$$

Un valor de MAE(%) = 10 indica que el error absoluto medio representa un 10 % de la variabilidad total de la métrica. Este enfoque permite una interpretación más homogénea de la magnitud del error, independientemente del rango de cada una de las métricas de evaluación de vídeo. Los valores de los rangos o la variabilidad total de cada una de las métricas (sin valores atípicos, sobre el material de prueba) se presentaron en la Tabla 3.4.

Capítulo 4

Resultados

Este capítulo presenta en detalle los resultados de aplicar las distintas estrategias propuestas en esta investigación para la optimización del coste computacional en la evaluación de QoE, siguiendo la metodología y el plan de pruebas descrito en el Capítulo 3.

Los enfoques adoptados se fundamentan en la explotación de características inherentes al contenido de los vídeos, como la redundancia espacial y temporal, con el objetivo de reducir el tiempo de procesamiento requerido en la extracción de información para el cálculo de QoE, afectando lo mínimo posible la precisión (error MAE en el valor de MOS inferior a 0.15) en la estimación de calidad proporcionada por la solución Video-MOS.

Cada uno de los diferentes enfoques se describe en profundidad, detallando los mecanismos empleados para minimizar el coste computacional. Los resultados obtenidos permiten comparar el desempeño de cada uno de los enfoques de optimización propuestos frente al modo de funcionamiento normal de la herramienta. Para cada comparativa, se analiza el impacto tanto en la reducción de coste computacional como en la precisión de la estimación del valor de MOS y en la extracción de características de vídeo, evaluando asimismo las ventajas y los inconvenientes de cada una de las estrategias.

Al final del capítulo se recoge el enfoque final propuesto para garantizar el funcionamiento en tiempo real de la herramienta Video-MOS. El enfoque propuesto integra las estrategias más relevantes identificadas a partir de los análisis de redundancia espacial y temporal, logrando así el equilibrio buscado entre eficiencia computacional y precisión en la estimación de calidad. Con este enfoque se cumplen los objetivos establecidos en el plan de pruebas: ahorro de coste computacional superior al 92.63 % y error de MOS en términos MAE inferior a 0.15.

4.1 Enfoques basados en redundancia espacial

El primer conjunto de enfoques se centra en la explotación de la redundancia espacial presente en las imágenes de una secuencia de vídeo, con el propósito de emplear un tamaño de imagen reducido para la extracción características y su posterior estimación de calidad. Se espera que la disminución de la resolución de las imágenes permita una reducción significativa del coste computacional, al requerir el procesamiento de un menor número de píxeles por imagen.

Tabla 4.1: Coste computacional según la resolución.

Resolución	**Número de píxeles**	**Coste computacional (%)**
1920x1080	2073600	100
1280x720	921600	44.44
960x540	518400	25
640x360	230400	11.11
480x270	129600	6.25
320x180	57600	2.78

Las resoluciones de vídeo seleccionadas preservan la relación de aspecto de 16:9 de las secuencias de prueba, relación de aspecto estandarizada en señal de televisión en HD. Las resoluciones analizadas en este estudio incluyen: 1920x1080 (resolución original), 1280x720, 960x540, 640x360, 480x270 y 320x180 píxeles.

La Tabla 4.1 muestra el coste computacional proporcional basado en el número de píxeles de la imagen para cada una de las resoluciones mencionadas anteriormente. Dicho coste computacional, expresado en términos porcentuales, se ha calculado considerando la proporción de píxeles procesados con cada resolución en relación con el tamaño original de las imágenes de prueba.

Tabla 4.2: Tiempo de procesamiento medio por imagen según la resolución.

Métrica de evaluación	**Tiempo medio (ms)**					
	320x180	**480x270**	**640x360**	**960x540**	**1280x720**	**1920x1080**
Spatial Information	0.98	2.26	5.33	11.07	18.74	46.36
Temporal Information	0.26	0.57	2.73	5.99	9.93	25.02
Blurring	0.31	0.67	2.17	4.60	8.16	20.15
Brightness	0.26	0.55	0.99	2.07	3.73	8.14
Contrast	0.69	1.54	4.86	10.14	17.67	44.65
Ringing	0.73	1.00	1.17	2.50	3.33	10.31
Blockloss	4.79	10.99	21.24	42.60	79.56	156.35
Blocking	1.41	3.62	7.38	27.91	72.89	232.04
Total	**9.44**	**21.20**	**45.86**	**106.87**	**214.00**	**543.03**

Los resultados obtenidos en pruebas experimentales muestran una disminución importante del tiempo de procesamiento medio empleado por cada una de las métricas de vídeo en la extracción de características, al emplear resoluciones inferiores. En la Tabla 4.2 se recoge el

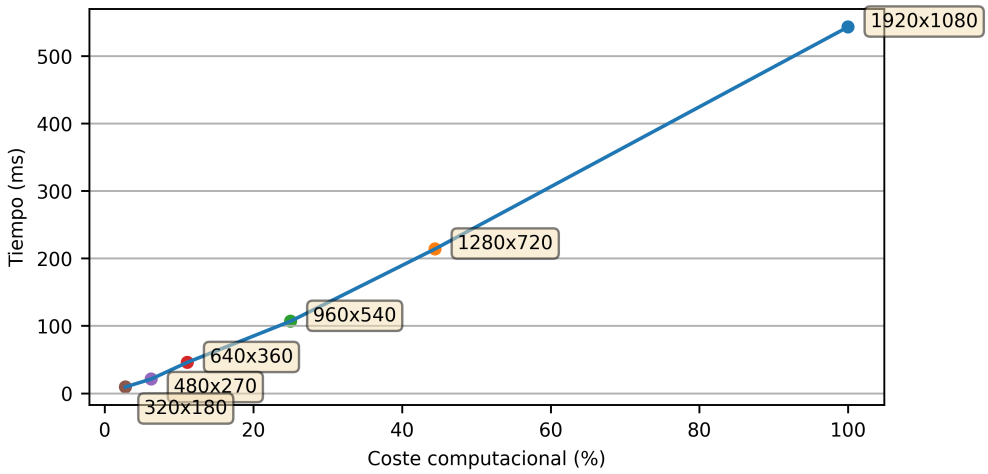

Figura 4.1: Tiempo de procesamiento medio por imagen vs. coste computacional.

tiempo medio de procesamiento por imagen, para cada una de las métricas de la solución Video-MOS.

El análisis de los datos revela una relación casi lineal entre el tiempo requerido para la extracción de características y el tamaño de la imagen procesada, como se muestra en la Figura 4.1.

El ahorro en el tiempo medio de procesamiento por imagen alcanza un 60.59 % para la resolución de 1280x720, un 80.31 % en 960x540, un 91.55 % en 640x360, un 96.08 % en 480x270 y un 98.25 % en 320x180. Si bien esta reducción en el coste computacional es significativa, es fundamental valorar el impacto que supone la extracción de características sobre imágenes de menor tamaño, y cómo afecta a la estimación del valor de MOS.

Dentro de los enfoques basados en la redundancia espacial de la imagen, se contemplan dos estrategias principales para reducir el tamaño de imagen: la selección de una región específica de la imagen o el cambio de resolución en toda la imagen.

4.1.1 Enfoque por selección de región específica de la imagen

En el ámbito de la percepción visual de los vídeos, el centro de la imagen es comúnmente identificado como el punto de mayor atracción para el observador. El principio de la jerarquía de la atención visual sostiene que los elementos ubicados en esta región poseen una mayor probabilidad de ser percibidos, debido a la predisposición natural del HVS a dirigir la mirada hacia el centro de la pantalla [225], [226].

Numerosos modelos de atención visual se basan en esta tendencia innata del ser humano a focalizarse en la zona central de una imagen [227]. Alternativamente, existen enfoques que emplean técnicas de detección de saliencia para identificar las áreas de mayor interés dentro de una imagen [228]. Sin embargo, el elevado coste computacional de estos métodos, junto con el uso de modelos de IA para la detección de objetos, la identificación de zonas con altos brillos y altos contrastes, la estimación del movimiento y el cálculo del flujo óptico,

representa una limitación significativa para su implementación en aplicaciones que requieren de procesamiento en tiempo real [229], [230].

4.1.1.1 Enfoque por región central

El primer enfoque propuesto se basa en la selección de la región central de la imagen, debido a su simplicidad y a la relevancia perceptual que esta zona tiene en la atención visual humana. La Tabla 4.3 presenta los resultados obtenidos al aplicar esta solución con las cinco resoluciones consideradas en el estudio.

Tabla 4.3: Error de MOS vs. coste computacional en el enfoque por región central.

Modo	Resolución	Coste computacional (%)	Error de MOS (MAE)
RC-720	1280x720	44.44	0.35
RC-540	960x540	25	0.39
RC-360	640x360	11.11	0.45
RC-270	480x270	6.25	0.47
RC-180	320x180	2.78	0.54

De igual manera, la Figura 4.2 ilustra gráficamente la relación entre el tamaño de la región central utilizada y la precisión en la estimación del valor de MOS (error de MOS, en términos MAE). Se observa que, para una región central de 1280x720 píxeles, el error de MOS asciende a 0.35, superando significativamente el umbral aceptable de 0.15, lo que compromete la precisión en la estimación de calidad. Con una región central de 320x180 píxeles, el error de valor de MOS es de 0.54.

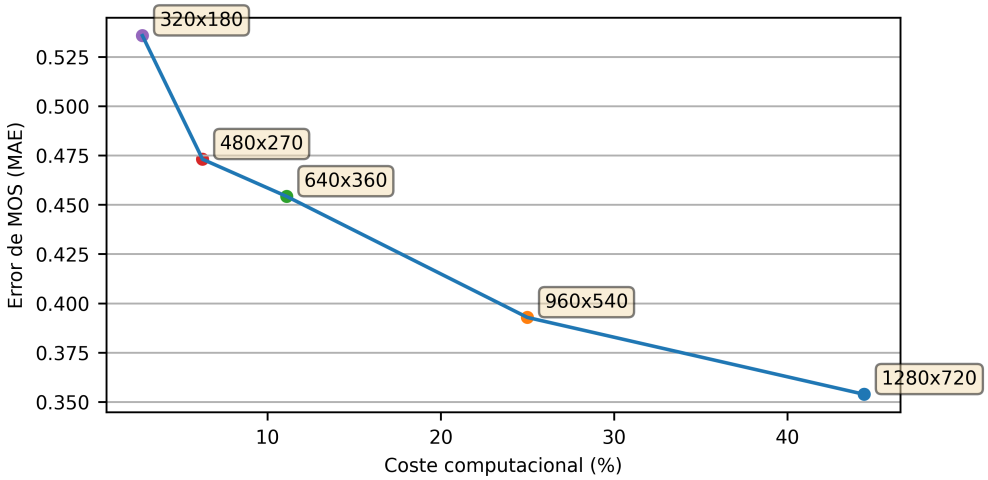

Figura 4.2: Error de MOS vs. coste computacional en el enfoque por región central.

Los resultados evidencian que, a medida que se reduce el tamaño de la región central analizada,

Tabla 4.4: Error en la extracción de características en el enfoque por región central.

Métrica de evaluación	**Error (MAE, en %)**				
	RC-180	**RC-270**	**RC-360**	**RC-540**	**RC-720**
Spatial Information	16.29	14.30	12.99	11.37	8.64
Temporal Information	11.02	9.67	8.67	6.80	4.64
Blurring	36.86	31.60	27.82	21.31	13.95
Brightness	26.07	22.34	19.41	14.54	9.67
Contrast	30.06	23.67	19.82	14.39	9.53
Ringing	26.15	22.07	19.59	15.54	12.79
Blockloss	6.25	2.08	3.13	1.04	2.08
Blocking	30.54	23.18	19.44	14.97	13.76
Vector de caract.	**22.90**	**18.62**	**16.36**	**12.50**	**9.38**

el error en la estimación de MOS aumenta. Esto se explica por el hecho de que el enfoque por región central restringe la extracción de características a una porción específica de la imagen, ignorando información contenida en las áreas adyacentes a la región central (áreas periféricas de la imagen). En caso de que las zonas no procesadas presenten características diferentes a las de la región central, el vector de características empleado en la estimación de calidad también se verá afectado, impactando directamente en el valor de MOS obtenido.

En la Tabla 4.4 se detallan los errores obtenidos en cada una de las métricas de evaluación de vídeo, expresados en términos MAE en valor porcentual. Para evaluar el impacto global del error en la extracción de información, se ha calculado el vector de características basado en el promedio de los errores de las ocho métricas de evaluación de vídeo. Si bien esta aproximación no pondera la contribución específica de cada métrica de la solución Video-MOS, debido a limitaciones derivadas del proceso abierto de solicitud de patente y de la privacidad industrial, se optó por realizar un promedio simple que proporciona una referencia útil para interpretar los resultados. Como era de esperar, el error en las métricas disminuye conforme se incrementa el tamaño de la región central. Un comportamiento similar se observa en el vector de características, donde, el error medio (MAE, en %) varía de 22.90 % con resolución de 320x180 píxeles a 9.38 % con resolución de 1280x720 píxeles.

El principal inconveniente del enfoque por región central radica en la omisión de información contenida en las áreas periféricas de la imagen. Esta limitación puede derivar en una evaluación incompleta de las características visuales del contenido, dado que estas zonas suelen incluir elementos relevantes que influyen en la percepción global de la calidad. En la producción televisiva, es habitual el empleo de elementos gráficos diseñados para captar la atención del espectador. Estos gráficos incluyen títulos, etiquetas, textos, logotipos y animaciones, los cuales no solo refuerzan la identidad del programa, sino que también guían la narrativa visual, siendo particularmente frecuente en informativos y programas de entretenimiento.

Figura 4.3: Elementos de grafismos en una secuencia de prueba de La1 HD.

Los gráficos televisivos suelen incorporar colores brillantes y altos contrastes con el fin de maximizar su impacto visual. La Figura 4.3 ilustra un ejemplo concreto de esta problemática, mediante una captura de un programa de entretenimiento emitido por el canal La1 HD de RTVE. En la imagen, se observa la presencia de múltiples elementos gráficos ubicados tanto en las esquinas como en la parte inferior de la imagen. La cuadrícula verde superpuesta delimita la región central de 960x540 píxeles utilizada en la extracción de características para el modo RC-540. Como se puede apreciar, todos los elementos gráficos quedan fuera del área de procesamiento del vídeo y, por tanto, no contribuyen a la estimación del valor de MOS cuando se utiliza este modo para este caso concreto.

Para evaluar el impacto de esta limitación, se han diseñado pruebas experimentales que comparan dos versiones de un mismo contenido: una versión con gráficos añadidos y otra sin gráficos. El contenido seleccionado para estas pruebas es la animación de código abierto *Big Buck Bunny* [1], desarrollada por la Fundación Blender. La inclusión de los elementos gráficos se ha realizado utilizando la herramienta profesional de edición y postproducción DaVinci Resolve [2], de *Blackmagic Design*. La Figura 4.4 presenta una captura de la versión con gráfico del contenido *Big Buck Bunny*.

Los resultados obtenidos tras el análisis de las dos versiones del contenido utilizando la solución Video-MOS se presentan en el Anexo A. La incorporación del elemento gráfico sobre el vídeo de prueba tiene un impacto significativo en la atracción visual del espectador. A pesar de ser un diseño relativamente simple (tres palabras en blanco sobre fondo negro), su presencia afecta de distinta manera a las métricas de evaluación de vídeo, lo que repercute directamente en la extracción final de características y, en última instancia, en la estimación de la calidad. Las Figuras A.1 a A.8, incluidas en el Anexo A, presentan una comparativa detallada de las dos versiones del contenido para cada una de las ocho métricas de evaluación de vídeo de

[1] https://peach.blender.org/
[2] https://www.blackmagicdesign.com/es/products/davinciresolve

Figura 4.4: Contenido *Big Buck Bunny* con grafismo.

Video-MOS. En particular, la métrica de vídeo *Temporal Information* (Figura A.2) no muestra diferencias significativas entre ambas versiones, ya que el gráfico permanece constante en el tiempo y no introduce variaciones de movimiento entre imágenes consecutivas. De manera similar, las métricas de *Brightness* y de *Constrast* (Figura A.4 y Figura A.5, respectivamente) presentan valores prácticamente similares, dado que el área que ocupa el gráfico es pequeña en comparación con el tamaño total de la imagen. Además, las métricas de *Ringing*, de *Blockloss* y de *Blocking* (Figura A.6, Figura A.7 y Figura A.8), no reflejan diferencias notables, ya que el gráfico no introduce distorsiones ni artefactos que afecten a la estructura de la imagen. Sin embargo, las métricas de *Spatial Information* y de *Blurring* (Figura A.1 y Figura A.3) presentan variaciones considerables. La presencia del gráfico afecta a la distribución de las altas frecuencias espaciales y al nivel de nitidez de las imágenes, lo que genera diferencias en la extracción de características para estas dos métricas de vídeo.

La Tabla 4.5 resume los errores obtenidos para esta prueba concreta en las métricas de evaluación de vídeo (MAE, en %), comparando ambos contenidos. Se observa que las métricas con mayores desviaciones (mayor error) son las métricas de vídeo *Spatial Information* y *Blurring*, con errores del 13.33 % y 8.67 %, respectivamente. El error en las métricas restantes es inferior al 3.6 %, siendo 3.92 % el error final obtenido para el vector de características.

La Figura 4.5 ilustra la estimación final de la calidad en términos de MOS para ambas versiones del contenido utilizando la herramienta Video-MOS. Se obtiene un error de MOS (MAE) de 0.29, error significativo dado el tipo de prueba realizada. Aunque la magnitud del error puede depender de las características específicas de cada vídeo, el valor obtenido pone de manifiesto la influencia de los gráficos en la extracción de determinadas características de imagen y, en consecuencia, en la estimación final de calidad.

Estos resultados cuestionan la validez del enfoque como solución de optimización. La eliminación de información presente en los bordes y en las esquinas de la imagen compromete la precisión de la herramienta de evaluación, especialmente en contenidos con gráficos.

Tabla 4.5: Error en la extracción de características en la prueba de influencia de grafismos.

Métrica de evaluación	**Error (MAE, en %)**
Spatial Information	13.33
Temporal Information	0.52
Blurring	8.67
Brightness	1.62
Contrast	2.31
Ringing	3.51
Blockloss	0.00
Blocking	1.38
Vector de caract.	**3.92**

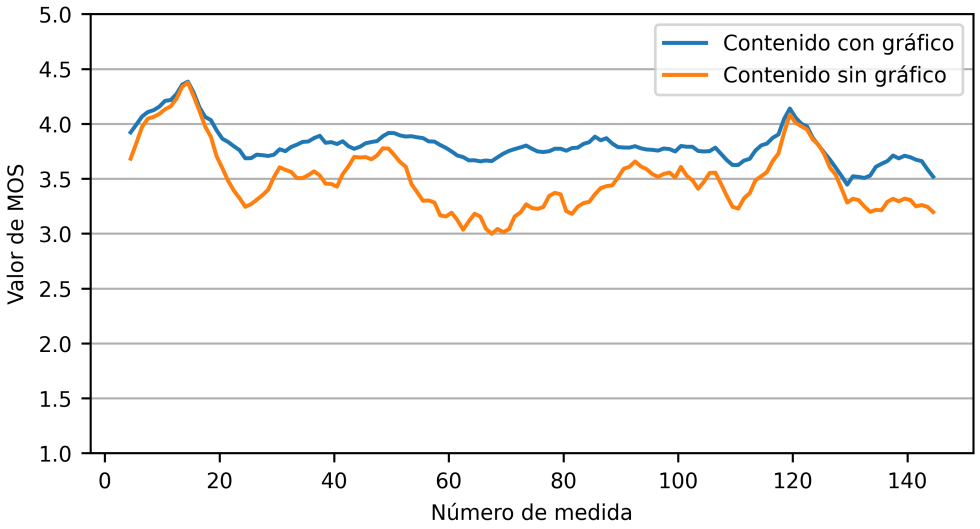

Figura 4.5: Influencia de grafismos en la estimación de MOS.

4.1.1.2 Enfoque por regiones

El enfoque basado exclusivamente en la región central resulta inadecuado para contenidos que incluyan gráficos, dado que estos elementos localizados en áreas periféricas de la imagen, como se ha evidenciado, ejercen un impacto significativo en algunas de las características de imagen.

Para abordar esta limitación, se propone un segundo enfoque basado en la selección específica de áreas de la imagen mediante un esquema de división por regiones. El método de división segmenta la imagen en nueve partes de igual tamaño, organizadas en una cuadrícula de 3x3. De este modo, cada región resultante tiene una resolución de 640x360 píxeles, manteniendo la

Figura 4.6: División por regiones en una secuencia de prueba de La1 HD.

relación de aspecto de 16:9. La elección de una segmentación de 3x3 y regiones de tamaño 640x360 píxeles también responde a consideraciones técnicas relacionadas con la codificación de vídeo. En particular, este tamaño de resolución permite una alineación precisa con bloques de píxeles de 8x8, ampliamente utilizados en diversas métricas de vídeo para la evaluación de artefactos en la estructura de la imagen, por la transformación de la DCT y por el proceso de codificación del vídeo. La Figura 4.6 ilustra la distribución de las nueve regiones dentro de la imagen, sobre un contenido perteneciente a un programa de entretenimiento del canal La1 HD de RTVE. Por ejemplo, la región RG-1 es la región localizada en la esquina superior izquierda, la región RG-5 es la región situada en el centro de la imagen, y la región RG-9 se encuentra en la esquina inferior derecha.

De manera análoga al análisis realizado por región central, la Tabla 4.6 presenta los errores de MOS obtenidos al evaluar cada una de las regiones de forma individual (regiones RG-1 a RG-9). Los resultados revelan que los errores son elevados en todas las regiones, con valores de error en la estimación de MOS que oscilan entre 0.40 y 0.56. Esta variabilidad sugiere que la estimación de calidad depende en gran medida de la información contenida en la totalidad de la imagen, lo que pone en cuestión la viabilidad de utilizar una única región de la imagen para la evaluación de QoE.

El análisis de los datos de la Tabla 4.6 pone de manifiesto varios aspectos clave. En primer lugar, se observa que no todas las regiones de la imagen tienen la misma importancia en la estimación del valor de MOS, lo que sugiere que ciertas áreas pueden contener información más representativa de la calidad percibida en función de las características globales de la escena. En particular, el modo RG-5, correspondiente al uso de la región central de la imagen (similar al modo RC-360 en enfoque por región central), muestra que el centro de la imagen no es necesariamente la región que ofrece la mejor estimación del valor de calidad global. Mientras que el error obtenido en el modo RG-5 es de 0.45, otros modos como el modo RG-3 (región superior derecha) o el modo RG-9 (región inferior derecha) presentan errores de MOS

Tabla 4.6: Error de MOS vs. coste computacional en el enfoque por regiones.

Modo	Nº región	Resolución	Coste computacional (%)	Error de MOS (MAE)
RG-9	9	640x360	11.11	0.40
RG-3	3	640x360	11.11	0.41
RG-2	2	640x360	11.11	0.44
RG-5	5	640x360	11.11	0.45
RG-8	8	640x360	11.11	0.47
RG-1	1	640x360	11.11	0.49
RG-7	7	640x360	11.11	0.54
RG-4	4	640x360	11.11	0.56
RG-6	6	640x360	11.11	0.56

más bajos (aunque elevados) de 0.41 y 0.40, respectivamente.

En lo que respecta a la extracción de características de las métricas de vídeo, el análisis del vector de características revela que, si bien el modo RG-5 no es el modo que proporciona el menor error en la estimación de MOS, sí que es el que ofrece el promedio (vector de características) de errores más bajo entre las ocho métricas de vídeo evaluadas, con un 16.36 % de error. En contraste, estos errores son de 21.80 % con el modo RG-3 y de 20.49 % con el modo RG-9. La Tabla 4.7 recoge, para cada uno de los modos del enfoque por regiones, los valores de error obtenidos para cada una de las métricas individuales de vídeo y para el vector de características.

Tabla 4.7: Error en la extracción de características en el enfoque por regiones.

Métrica de evaluación	Error (MAE, en %)								
	RG-1	**RG-2**	**RG-3**	**RG-4**	**RG-5**	**RG-6**	**RG-7**	**RG-8**	**RG-9**
Spatial Information	20.45	16.30	19.30	17.63	12.99	18.03	19.40	16.96	14.47
Temporal Information	12.84	10.49	13.41	10.02	8.67	10.71	14.04	10.90	12.84
Blurring	33.22	38.55	37.39	35.08	27.82	30.29	44.76	38.59	35.22
Brightness	24.04	21.62	24.20	17.88	19.41	18.60	24.38	23.72	24.02
Contrast	27.00	23.05	25.92	24.38	19.82	24.36	24.45	23.04	24.34
Ringing	30.01	22.00	31.47	21.48	19.59	21.57	20.31	19.31	31.19
Blockloss	3.78	2.11	2.32	2.32	3.13	2.28	3.11	3.49	2.48
Blocking	19.97	20.16	20.35	17.78	19.44	19.20	18.81	18.31	19.35
Vector de caract.	**21.41**	**19.28**	**21.80**	**18.32**	**16.36**	**18.13**	**21.16**	**19.29**	**20.49**

Los errores elevados obtenidos tanto en la estimación del valor de MOS como en la extracción de características sugieren que el enfoque por regiones no es una estrategia válida de optimización de la herramienta Video-MOS. La no validez de utilizar una única región para representar con precisión la calidad global de la imagen refuerza la necesidad de considerar estrategias alternativas que combinen diferentes regiones para preservar mejor la integridad de la información de toda la imagen.

4.1.1.3 Enfoque por combinación de regiones basado en promedios

Este enfoque propone un cambio significativo respecto a los enfoques anteriores, al no utilizar únicamente una sola región de la imagen para la extracción de la información. En su lugar, se busca combinar múltiples regiones y calcular valores promedios de las características de las métricas de vídeo, con el objetivo de preservar una mayor cantidad de información de la imagen. En el enfoque por combinación de regiones basado en promedios, todas las regiones de la rejilla 3x3 tienen el mismo peso en el proceso de combinación (ponderación por igual), sin importar la posición que ocupa la región dentro de la imagen y de la rejilla 3x3.

Para determinar la mejor combinación de regiones, se ha llevado a cabo un proceso de minimización con el conjunto de vídeos de prueba, en el que se han evaluado todas las posibles combinaciones con las nueve regiones diferentes. Se han explorado desde combinaciones que utilizan una única región (mismos resultados que los obtenidos en el enfoque por regiones) hasta aquella combinación que incorpora las nueve regiones. Para cada conjunto evaluado, se ha elegido la combinación de regiones que ofrece el menor error en la estimación de MOS. La combinación de regiones podría cambiar en caso de utilizar otro conjunto de datos, en función de las características de las imágenes. Los resultados obtenidos se presentan en la Tabla 4.8

Tabla 4.8: Error de MOS vs. coste computacional en el enfoque por combinación de regiones basado en promedios.

Modo	Nº regiones	Resolución	Coste computacional (%)	Error de MOS (MAE)
RG-N1-P	9	640x360 (x1)	11.11	0.40
RG-N2-P	3, 9	640x360 (x2)	22.22	0.36
RG-N3-P	3, 5, 9	640x360 (x3)	33.33	0.35
RG-N4-P	2, 3, 5, 9	640x360 (x4)	44.44	0.34
RG-N5-P	2, 3, 5, 8, 9	640x360 (x5)	55.55	0.35
RG-N6-P	1, 2, 3, 5, 8, 9	640x360 (x6)	66.66	0.35
RG-N7-P	1, 2, 3, 5, 7, 8, 9	640x360 (x7)	77.77	0.36
RG-N8-P	1, 2, 3, 5, 6, 7, 8, 9	640x360 (x8)	88.88	0.37
RG-N9-P	1, 2, 3, 4, 5, 6, 7, 8, 9	640x360 (x9)	100.00	0.38

El análisis de la Tabla 4.8 revela que el mejor resultado se obtiene con el modo RG-N4-P, con un coste computacional del 44.44 %, y utilizando únicamente cuatro regiones: región RG-2, región RG-3, región RG-5 y región RG-9. Con este modo, el error de MOS obtenido es de

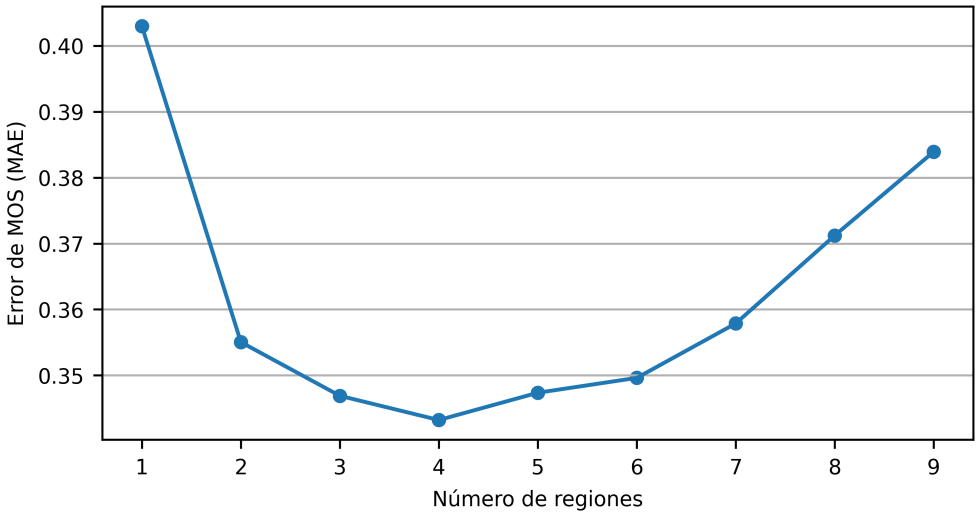

Figura 4.7: Error de MOS vs. número de regiones en el enfoque por combinación de regiones basado en promedios.

0.34, lo que representa una ligera mejoría con respecto a los enfoques anteriores. Con igualdad de coste computacional (44.44 %), con el modo RC-720 del enfoque por región central se obtuvo un valor de 0.35 de error de MOS. El peor resultado se obtiene cuando se emplea únicamente una región (similar al modo RG-9, en el enfoque por regiones), con un error de MOS de 0.40. La Figura 4.7 ilustra gráficamente el error de MOS en función del número de regiones utilizadas en el enfoque por combinación de regiones basado en promedios.

En cuanto a la extracción de características de las métricas de vídeo, los resultados reflejan una tendencia decreciente del error al aumentar el número de regiones utilizadas, tanto a nivel individual de cada métrica de vídeo como en el vector de características. Estos datos recogidos en la Tabla 4.9 confirman que la combinación de múltiples regiones contribuye a una estimación más precisa de la calidad del vídeo, al mantener mejor la integridad de toda la información. Mientras que el error en el vector de características en el modo RC-720 (enfoque por región central) era de 9.38 %, para un coste computacional del 44.44 %, en el modo RG-N4-P (enfoque por combinación de regiones basado en promedios) se consigue un error ligeramente inferior de 8.76 %.

Un aspecto llamativo del enfoque por combinación de regiones basado en promedios es que, incluso utilizando las nueve regiones (y por tanto el mismo coste computacional que si funcionara la herramienta en modo normal), los errores en la extracción de características de las métricas de vídeo no son nulos en todos los casos. En el modo RG-N9-P se obtiene un error en el vector de características de 4.97 %, lo que provoca un error en la estimación de MOS de 0.38. Estos datos sugieren que, aunque algunas métricas de vídeo presentan errores mínimos e inferiores al 1 % (métrica de *Blurring*, de *Blockloss* y de *Blocking*) o incluso nulos en la métrica de *Brightness* (media de las medias de todos los subconjuntos del mismo tamaño es igual a la media del conjunto completo), otras métricas experimentan desviaciones considerables (4.76 % en la métrica *Temporal Information*, 6.27 % en *Spatial Information*, 8.13 % en *Ringing* y 18.84 % en *Contrast*).

Tabla 4.9: Error en la extracción de características en el enfoque por combinación de regiones basado en promedios.

Métrica de evaluación	Error (MAE, en %)								
	RG-N1-P	RG-N2-P	RG-N3-P	RG-N4-P	RG-N5-P	RG-N6-P	RG-N7-P	RG-N8-P	RG-N9-P
Spatial Information	14.47	9.89	7.19	8.04	5.99	5.87	4.08	5.37	6.27
Temporal Information	12.84	9.94	5.59	5.40	4.54	4.71	5.03	4.92	4.76
Blurring	35.22	18.98	13.53	16.27	11.13	8.78	5.92	4.40	0.08
Brightness	9.08	5.65	3.39	3.05	2.61	1.87	1.25	0.97	0.00
Contrast	20.05	19.87	17.72	18.10	17.84	18.53	18.06	18.48	18.84
Ringing	31.19	18.62	11.78	9.82	8.36	8.36	8.03	8.12	8.13
Blockloss	2.48	2.07	2.14	1.83	1.57	1.24	1.00	0.91	0.82
Blocking	19.35	13.00	7.83	7.55	5.86	4.96	3.86	2.40	0.82
Vector de caract.	**18.08**	**12.25**	**8.65**	**8.76**	**7.24**	**6.79**	**5.90**	**5.70**	**4.97**

La principal causa de estos errores radica en el uso de estadísticos de varianza y desviación estándar en la implementación y en los algoritmos de las métricas de vídeo. La comparativa entre el cálculo de una métrica obtenida a partir de toda la imagen y una métrica obtenida a partir del promedio de valores introduce diferencias inevitables en la extracción de la información. En el Anexo B se proporciona un ejemplo concreto de una de las secuencia de prueba del canal La1 HD de RTVE. La Figura B.1 muestra los valores de MOS y de las métricas de vídeo *Spatial Information*, *Temporal Information* y *Blurring* para cada una de las nueve regiones de la imagen analizadas de forma independiente. El resto de valores de las métricas de vídeo se detallan en la Tabla B.1.

Para el caso concreto de la prueba del Anexo B, los valores de MOS obtenidos individualmente para cada región varían significativamente. Mientras que en el modo de funcionamiento normal de la solución Video-MOS el valor de MOS es de 4.28 para esta secuencia concreta de tres segundos de duración, el promedio de los valores de MOS calculados a partir de las nueve regiones individuales es de 3.82. Los valores de MOS obtenidos en cada una de las regiones son: 3.69 en la región RG-1, 3.86 en RG-2, 3.88 en RG-3, 3.91 en RG-4, 4.29 en RG-5, 3.97 en RG-6, 3.28 en RG-7, 3.98 en RG-8 y 3.52 en RG-9. Esta diferencia de valor representa un error MAE de 0.46, equivalente a un 11.5 % de error en la escala MOS.

Este comportamiento es consistente con los resultados obtenidos con el conjunto total de las secuencias de prueba. Mientras que el error en el vector de características es de 4.97 % para el conjunto total de datos prueba, para el caso concreto de la prueba del Anexo B se obtiene un error de 4.90 %. En la Tabla B.2 se puede observar cómo, para esta prueba concreta, algunas métricas de vídeo muestran errores bajos en la comparativa. Otras métricas presentan errores considerables: error de 15.38 % en la métrica de *Contrast* y error de 16.93 % en la métrica de *Ringing*.

Los errores elevados en la estimación de calidad y los problemas derivados de la combinación de regiones basada en promedios hacen que este enfoque no sea viable como técnica de optimización del coste computacional en la herramienta Video-MOS.

4.1.1.4 Enfoque por combinación de regiones basado en saliencia

Este nuevo enfoque propuesto, fundamentado en la división de la imagen en las nueve regiones y en la rejilla 3x3 vista anteriormente, introduce un mecanismo de ponderación basado en la saliencia del vídeo. En función de las zonas de interés detectadas en las imágenes, se asignan distintos pesos a cada región, determinando así su relevancia en el proceso de extracción de características y en la estimación de MOS.

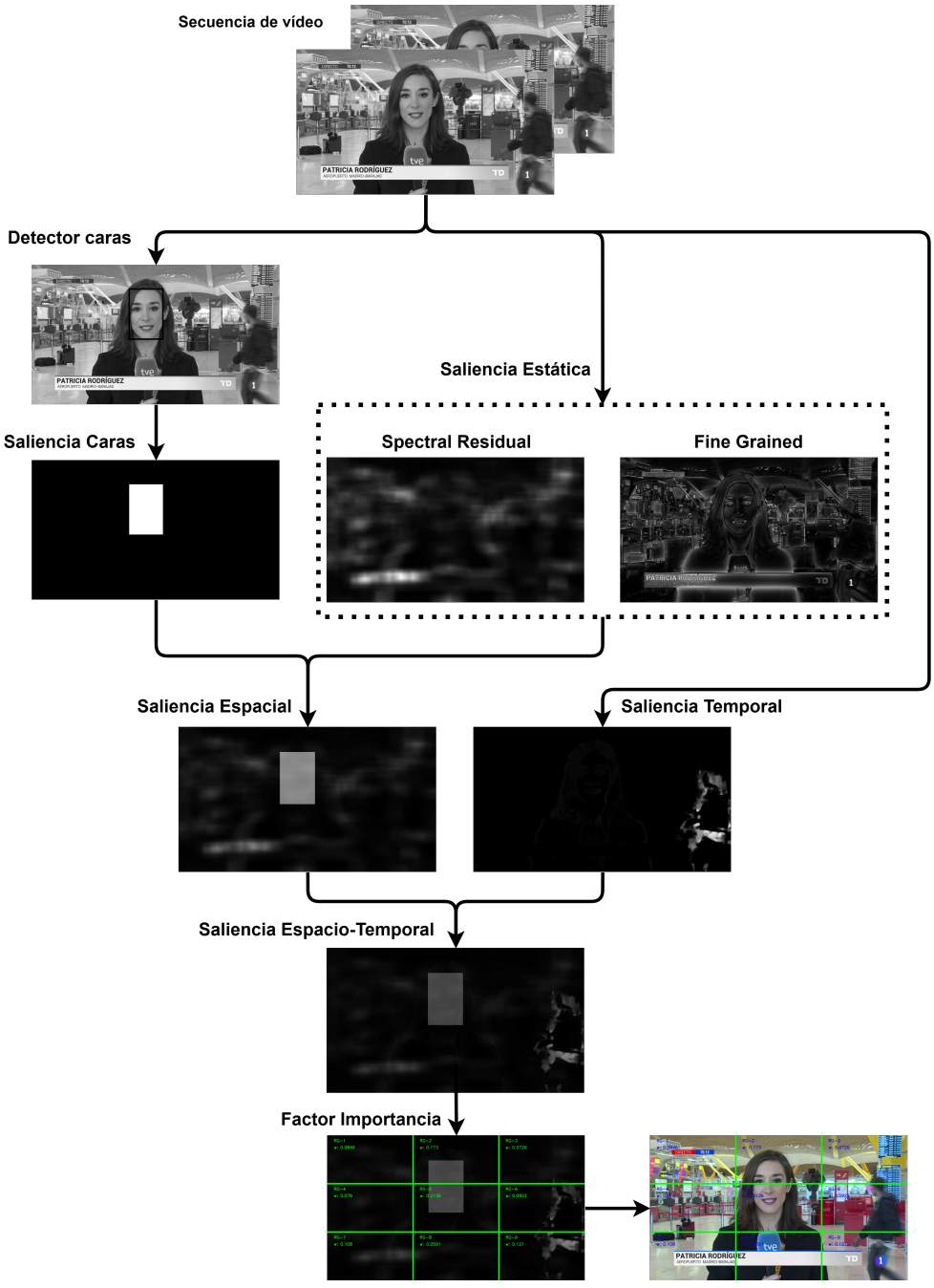

Figura 4.8: Método propuesto de detección de saliencia.

Con el objetivo de optimizar el coste computacional, en esta tesis doctoral se propone un método de detección de saliencia que utiliza exclusivamente las dos primeras imágenes de la secuencia de vídeo a una resolución reducida de 480x270 píxeles. Esta aproximación permite reducir significativamente la carga de procesamiento sin comprometer la eficacia en la identificación de las áreas de mayor interés del contenido visual.

La Figura 4.8 ilustra el procedimiento empleado para la detección de saliencia en cada secuencia de vídeo. A partir de las dos imágenes iniciales, se generan los mapas de saliencia espacial y saliencia temporal, que al combinarse, forman el mapa de saliencia espacio-temporal. Todos los mapas de saliencia están normalizados entre 0 y 255, garantizando así una escala correcta en la representación de los resultados. En función del número de regiones a considerar para la extracción de las características de la imagen, se identifican aquellas con mayor relevancia según el mapa de saliencia obtenido y se asignan los pesos (normalizados a 1) a cada una de las regiones para ponderar la extracción de la información.

La obtención del mapa de saliencia espacial se basa en la combinación de la saliencia estática y el mapa de saliencia de la detección de caras de personas humanas. Para la obtención de la saliencia estática se han empleado dos métodos implementados en la biblioteca de OpenCV: Spectral Residual (SR) [3] y Fine Grained (FG) [4]. Ambos métodos permiten detectar (aunque de manera diferente) los elementos visuales destacados en la escena, considerando características como el color, el contraste, el brillo, la textura, la forma y la orientación de los objetos. Adicionalmente, para la detección de rostros se ha empleado el modelo de IA YuNet [5], ampliamente reconocido por su eficiencia computacional, rapidez y precisión en la detección de caras humanas en imágenes.

Para el cálculo de la saliencia temporal se ha empleado el método de *Farneback*, una técnica basada en la estimación del flujo óptico que permite determinar el desplazamiento de los píxeles entre imágenes consecutivas. El método se fundamenta en el cálculo de gradientes espaciales y temporales, generando un campo de vectores que describe el movimiento horizontal y vertical presente en la secuencia de vídeo (en este caso concreto, el movimiento entre la primera y la segunda imagen de la secuencia de vídeo). Para el cálculo del flujo óptico mediante este método, se ha utilizado la función disponible en la biblioteca OpenCV [6].

El proceso de generación de los mapas de saliencia espacial y espacio-temporal se ha basado en la combinación de distintos mapas de saliencia previamente calculados. El mapa de saliencia espacial se obtiene mediante la fusión del mapa de saliencia estática y el mapa de saliencia con la detección de caras. El mapa de saliencia espacio-temporal resulta de la fusión del mapa de saliencia espacial con el mapa de saliencia temporal. Para llevar a cabo este proceso de combinación entre dos mapas de saliencia, se ha empleado la función de OpenCV *addWeighted* [7], la cual permite fusionar imágenes mediante una mezcla lineal ponderada,

[3] https://docs.opencv.org/4.11.0/df/d37/classcv_1_1saliency_1_1StaticSaliencySpectralResidual.html

[4] https://docs.opencv.org/4.11.0/da/dd0/classcv_1_1saliency_1_1StaticSaliencyFineGrained.html

[5] https://github.com/geaxgx/depthai_yunet

[6] https://docs.opencv.org/4.11.0/d4/dee/tutorial_optical_flow.html

[7] https://docs.opencv.org/4.11.0/d5/dc4/tutorial_adding_images.html

ajustando la contribución de cada mapa de saliencia en función de los coeficientes escogidos. El resultado de esta operación es una nueva imagen que integra (en función de los pesos asignados, de manera equilibrada o no) las características de ambas imágenes de entrada, proporcionando un nuevo mapa de saliencia final. La fórmula matemática de la función *addWeighted* se define en la Ecuación 4.1.

$$\text{Imagen Salida}(x, y) = \alpha \cdot \text{Imagen Entrada 1}(x, y) + \beta \cdot \text{Imagen Entrada 2}(x, y) \tag{4.1}$$

Donde:

- Imagen Entrada 1 (x, y) e Imagen Entrada 2 (x, y) representan las imágenes de entrada que se van a combinar.
- Imagen Salida (x, y) representa la imagen resultado de la combinación de las dos imágenes de entrada.
- α y β representan los pesos asignados y aplicados a cada una de las dos imágenes de entrada, respectivamente.

Para la generación del mapa de saliencia espacial, se han utilizado coeficientes de ponderación α y β de 0.5, con el propósito de asignar un peso equitativo del 50 % tanto a la saliencia estática como al mapa de saliencia derivado de la detección de caras. En el caso del mapa de saliencia espacio-temporal, se han aplicado distintas combinaciones de pesos para evaluar su impacto en la fusión de la información: (0.5, 0.5), (0.75, 0.25) y (1.0, 0.0), asignando estos valores α y β al mapa de saliencia espacial y al mapa de saliencia temporal, respectivamente.

El Anexo C presenta un conjunto de capturas del método de saliencia implementado, utilizando una de las secuencias de vídeo de prueba del canal La1 HD de RTVE. La Figura C.1 y Figura C.2 muestran la primera y la segunda imagen en escala de grises de la secuencia de prueba correspondiente a la medida de tres segundos. La Figura C.3 ilustra el resultado del proceso de detección de rostros mediante el modelo de IA YuNet, mientras que la Figura C.4 muestra el mapa de saliencia obtenido a partir de esta detección de caras.

La Figura C.5 y Figura C.6 representan, respectivamente, los mapas de saliencia estática generados mediante los métodos SR y FG. Posteriormente, la combinación ponderada del 50 % entre el mapa de saliencia de detección de caras y el mapa de saliencia estática SR se muestra en la Figura C.7, resultando en el mapa de saliencia espacial. De manera análoga, la Figura C.8 presenta el mapa de saliencia espacial generado utilizando el método FG para el cálculo de la saliencia estática.

El mapa de saliencia temporal se ilustra en la Figura C.9. La Figura C.10, Figura C.11 y Figura C.12 recogen los resultados de los mapas de saliencia espacio-temporal, aplicando ponderaciones de (50 %, 50 %), (75 %, 25 %) y (100 %, 0), respectivamente, utilizando el método SR para la saliencia estática.

Finalmente, desde la Figura C.13 hasta la Figura C.21, se presentan los mapas de saliencia espacio-temporal obtenidos con la metodología basada en el método SR de saliencia estática y con una ponderación 50 % - 50 % en la combinación entre la saliencia espacial y la saliencia temporal. La Figura C.13 muestra el mapa de saliencia espacio-temporal utilizando una