https://doi.org/10.37815/rte.v37nE1.1317
Artículos originales

Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video

Evaluation and Comparison of Objective Metrics PSNR, SSIM, and LPIPS for Video Quality Analysis

Carlos Flores Maza[1] https://orcid.org/0009-0001-6915-3158,
Santiago González Martínez[1] https://orcid.org/0000-0001-6604-889X

[1]*Universidad de Cuenca*, Cuenca, Ecuador
bladimir.flores@ucuenca.edu.ec,
santiago.gonzalezm@ucuenca.edu.ec

Enviado: 2025/05/08
Aceptado: 2025/08/20
Publicado: 2025/10/15

Resumen

Este artículo presenta una herramienta para la evaluación de la calidad de video, que permite seleccionar parámetros de escalabilidad de calidad (QP), temporal (FPS) y espacial (bitrate). La propuesta integra métricas tradicionales como *Peak Signal-to-Noise Ratio* (PSNR) y *Structural Similarity Index* (SSIM), junto con la métrica perceptual *Learned Perceptual Image Patch Similarity* (LPIPS), basada en redes neuronales profundas. Para validar su efectividad, se aplicó una metodología en dos fases de evaluación subjetiva. En la primera, los participantes evaluaron videos codificados con un mismo parámetro de escalabilidad, mostrando alta correspondencia entre la percepción visual y las métricas. En la segunda, se compararon diferentes configuraciones, evidenciando preferencia por alta calidad y escalabilidad espacial intermedia. Asimismo, en experimentos adicionales con distorsiones comunes (difuminado y ruido), LPIPS alcanzó una sensibilidad del 73.64 %, superando a PSNR y SSIM en su alineación con la percepción humana. La principal contribución de este trabajo es una herramienta que combina evaluaciones objetivas y subjetivas, facilitando un análisis más completo y cercano a la percepción visual humana.

Palabras clave: escalabilidad, PSNR, SSIM, LPIPS, calidad de video, objetivo, subjetivo.

Cómo citar: Flores, C. & González, S. (2025). Evaluación y Comparación de Métricas Objetivas PSNR, SSIM y LPIPS para el Análisis de Calidad de Video. *Revista Tecnológica - Espol, 37(E1), 56-76.* https://doi.org/10.37815/rte.v37nE1.1317

Abstract

This paper presents a tool for video quality assessment that allows the selection of quality (QP), temporal (FPS), and spatial (bitrate) scalability parameters. The proposal integrates traditional metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), together with the perceptual metric Perceptual Image Patch Similarity (LPIPS), which is based on deep neural networks. To validate its effectiveness, a two-phase subjective evaluation methodology was applied. In the first phase, participants assessed videos encoded with the same scalability parameter, showing a strong correspondence between visual perception and objective metrics. In the second phase, different configurations were compared, revealing a preference for high quality and intermediate spatial scalability. Additionally, in experiments with common distortions such as blurring and noise, LPIPS achieved a sensitivity of 73.64%, outperforming PSNR and SSIM in its alignment with human perception. The main contribution of this work is a tool that combines objective and subjective evaluations, enabling a more comprehensive analysis that closely reflects human visual perception.

Keywords: scalability, PSNR, SSIM, LPIPS, video quality, objetive, subjetive.

Introducción

La evaluación de la calidad visual en imágenes y videos es un desafío clave en campos como la transmisión de video, la realidad aumentada, la restauración de imágenes, la síntesis de superresolución y las aplicaciones médicas (Huynh-Thu y Ghanbari, 2012). Tradicionalmente, se han empleado métricas como Peak Signal-to-Noise Ratio (PSNR) (Izquierdo, 2017) y Structural Similarity Index (SSIM) (Huynh-Thu y Ghanbari, 2012). Aunque útiles por su simplicidad, estas métricas no capturan adecuadamente las complejidades de la percepción visual humana, pues suelen ser poco sensibles a distorsiones perceptuales relevantes para la experiencia del espectador (Hou et al., 2022; J. Wang et al., 2023).

Con el avance de la inteligencia artificial, han surgido métricas basadas en redes neuronales profundas que modelan de mejor manera la percepción visual. Entre ellas destacan Perceptual Image Patch Similarity (LPIPS) y Perceptual image-error Assessment through Pairwise Preference (PieAPP), que han mostrado un alineamiento más estrecho con la evaluación subjetiva humana (Prashnani et al., 2018; R. Zhang et al., 2018). Estudios recientes confirman su superioridad frente a PSNR y SSIM en distintas tareas de restauración y superresolución (K. Zhang et al., 2021; S. Zhang et al., 2023), teniendo en cuenta que en Gu et al., (2020) mostraron que LPIPS y PieAPP lograron una mejor correlación con la percepción humana, con valores de Spearman's Rank Correlation Coefficient (SRCC) de 0.488 y 0.534, en comparación con los 0.239 y 0.338 de PSNR y SSIM. Asimismo, se han propuesto variantes mejoradas como A-DISTS (Ding et al., 2020, 2023), que optimiza la métrica DISTS mediante una asignación más precisa de pesos locales y logra una correlación más alta con evaluaciones humanas. De igual forma, en áreas específicas como la calidad de imágenes médicas o en competiciones internacionales como NTIRE, métricas perceptuales como LPIPS, DISTS o sus variantes han mostrado correlaciones significativamente mayores con la percepción humana (Kastryulin et al., 2023; Gu et al., 2022).

En el caso de video, la evaluación es aún más compleja debido a la naturaleza dinámica del contenido. Métricas como FloLPIPS buscan integrar distorsiones espaciales y temporales, ofreciendo una valoración más completa en aplicaciones como la interpolación de fotogramas y la restauración de secuencias (Danier et al., 2022; Hou et al., 2022). No obstante, persisten limitaciones, sobre todo en entornos no controlados o en videos *in-the-wild*, donde los enfoques

sin referencia aún muestran correlaciones incompletas con la percepción humana (Li et al., 2019). Paralelamente, han surgido propuestas innovadoras como IQAGPT, que incorpora modelos de lenguaje para evaluar la calidad de imágenes médicas, marcando una tendencia hacia métricas más especializadas y cercanas a la valoración humana (Chen et al., 2023).

La principal contribución de este trabajo es el desarrollo de una herramienta novedosa que permite evaluar la calidad visual de videos de manera general, combinando métricas tradicionales (PSNR, SSIM) y métricas perceptuales basadas en redes neuronales profundas (LPIPS). A diferencia de estudios previos que analizan cada métrica de forma aislada, nuestra herramienta integra evaluaciones objetivas y subjetivas en un mismo entorno, permitiendo relacionar directamente los parámetros de escalabilidad del video con la percepción humana. Esto facilita un análisis más completo y preciso de la calidad visual, acercando los resultados de la evaluación automática a la experiencia real del espectador.

Materiales y Métodos

El presente estudio se enfoca en la evaluación de la calidad de video utilizando métricas tanto objetivas como subjetivas. Para ello, se desarrolló una herramienta que facilita la selección, codificación y análisis de videos de distintas categorías. La metodología diseñada se esquematiza en la Figura 1, y está compuesta por cuatro etapas: carga de video, codificación, selección de videos y generación de gráficas.

Cada una de estas etapas se detalla a continuación, describiendo los materiales utilizados, el proceso de muestreo, así como las técnicas de análisis aplicadas.

Muestreo

Para el estudio se definieron un total de cinco categorías de videos: videovigilancia, videoconferencia, entretenimiento, animaciones y tiempo real, con dos videos por cada categoría, obtenidos de la base de datos pública Xiph Media (Xiph Foundation, 2023). Los videos fueron elegidos en formato .yuv, debido a que este formato sin compresión es ideal para realizar comparaciones precisas entre el contenido original y el contenido procesado. Todos los videos tienen una resolución CIF (352 × 288) y una duración menor a 15 segundos, lo cual es crucial para el procesamiento eficiente de las evaluaciones objetivas y para evitar la fatiga en las evaluaciones subjetivas. Los detalles de cada video por categoría se presentan en la Tabla 1.

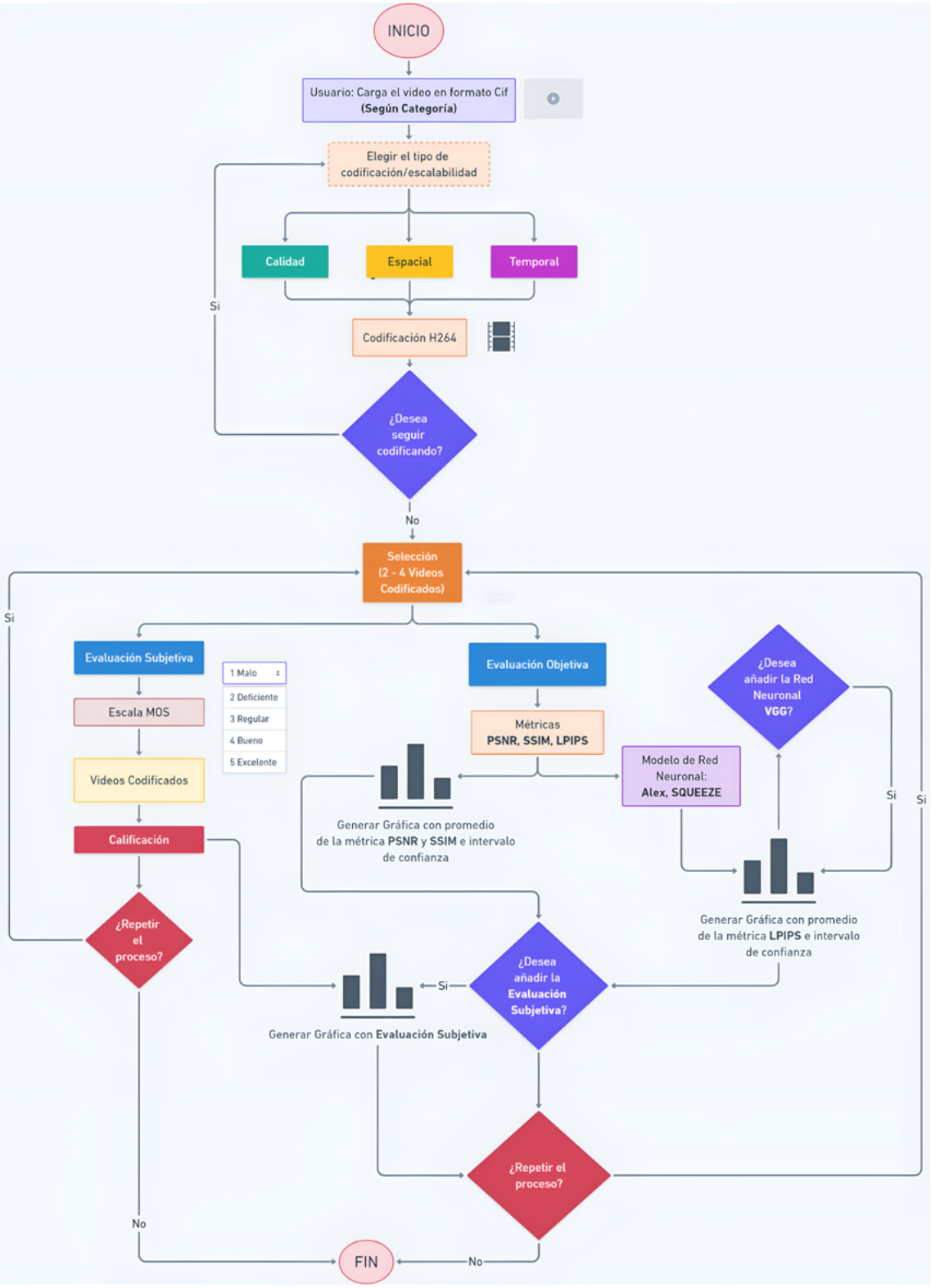
Tabla 1

Videos seleccionados por categoría

CATEGORÍA DE VIDEO	VIDEO SELECCIONADO
Videovigilancia	bus hall_monitor
Videoconferencia	akiyo deadline
Entretenimiento	football (b) soccer
Animaciones	big_buck_bunny elephants_dream
Tiempo real	crew football (a)

Figura 1

Metodología general para la evaluación objetiva y subjetiva de video



Procedimientos

Carga de video

Los videos seleccionados en formato .yuv son organizados por categorías, permitiendo al usuario elegir cuáles serán codificados en la siguiente etapa. Esta selección permite una variedad de escenarios de uso real, representando distintas demandas de calidad de imagen y fluidez.

Codificación

Una vez seleccionado el video, se procede a su procesamiento utilizando el códec H.264 y la herramienta FFmpeg (FFmpeg API, 2024). La codificación se ajusta según tres parámetros de escalabilidad: calidad, temporal y espacial, que permiten una adaptación eficiente del video según las necesidades específicas de cada aplicación, como se muestra en la Tabla 2 (Bowker, 2021; L. Wang, 2021; Watt, 2022).

Tabla 2

Relación entre categorías de video y escalabilidad

CATEGORÍA DE VIDEO	ESCALABILIDAD
Videovigilancia	Temporal, calidad
Videoconferencia	Temporal, espacial
Entretenimiento	Calidad, espacial
Animaciones	Calidad, temporal
Tiempo real	Temporal, espacial

- Calidad: Se controla mediante el parámetro de cuantización (QP), que varía en un rango de 0 a 60. Un valor de QP más bajo implica una mayor calidad visual y menor compresión, mientras que valores más altos incrementan la compresión a costa de la calidad. Este ajuste es crucial para equilibrar la relación entre calidad visual y tamaño del archivo.
- Temporal: Este aspecto está representado por los frames por segundo (FPS), que pueden ajustarse entre 2 y 30. Un mayor número de FPS proporciona una mejor fluidez en el video, lo que es importante para escenas con mucho movimiento, mientras que un valor menor puede ser suficiente para videos con menos dinamismo, reduciendo así el tamaño del archivo.
- Espacial: El bitrate define la cantidad de datos utilizados para representar cada segundo de video, comenzando desde 25 kbps. Al aumentar el bitrate, se incrementa la cantidad de información por cuadro, lo que resulta en una mejor calidad visual. Sin embargo, también se traduce en mayores requisitos de almacenamiento y ancho de banda.

Estos tres parámetros permiten una codificación flexible, adaptando la compresión y la calidad del video a las necesidades específicas. La Figura 2 ilustra el flujo de este proceso de codificación y cómo se aplican estos ajustes para optimizar el rendimiento y la calidad visual del video final.

Selección de videos

Después de la codificación, el usuario puede seleccionar entre dos y cuatro versiones de un mismo video con distintos niveles de escalabilidad para ser evaluados. Esto permite una comparación directa de las métricas de calidad entre las distintas versiones, facilitando el análisis objetivo y subjetivo de los videos.

Evaluación de la calidad

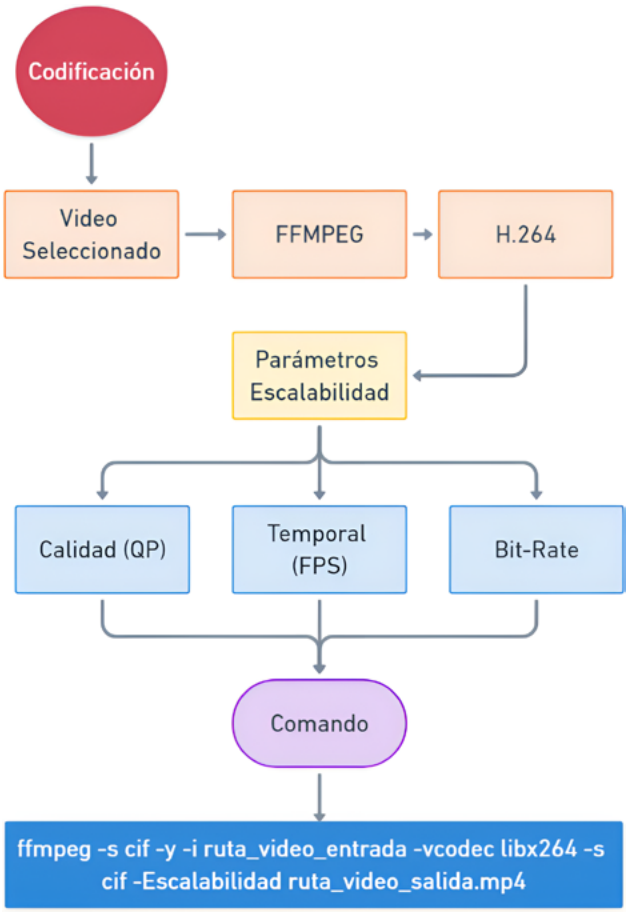
La calidad objetiva de los videos se evaluó mediante tres métricas ampliamente utilizadas en el campo de procesamiento de imágenes y video: PSNR, SSIM y LPIPS.

En primer lugar, el PSNR es una métrica clásica que destaca por su simplicidad y rápida implementación, ya que no demanda un alto costo computacional y, aun así, ofrece una buena aproximación a la calidad del video. Por otro lado, el SSIM analiza la similitud entre imágenes, considerando aspectos perceptuales como el brillo, el contraste y la estructura, lo que lo convierte en un indicador más representativo que el PSNR. Finalmente, el LPIPS introduce un enfoque basado en aprendizaje profundo: mediante redes convolucionales, evalúa la calidad desde una perspectiva perceptual más cercana al sistema visual humano, superando las limitaciones de las métricas tradicionales.

Estas métricas se aplicaron tras convertir los videos de su formato original .yuv a bgr, formato compatible con las bibliotecas de procesamiento de video como OpenCV y PyTorch; este proceso se ilustra en la Figura 3.

Figura 2

Metodología para codificar un video



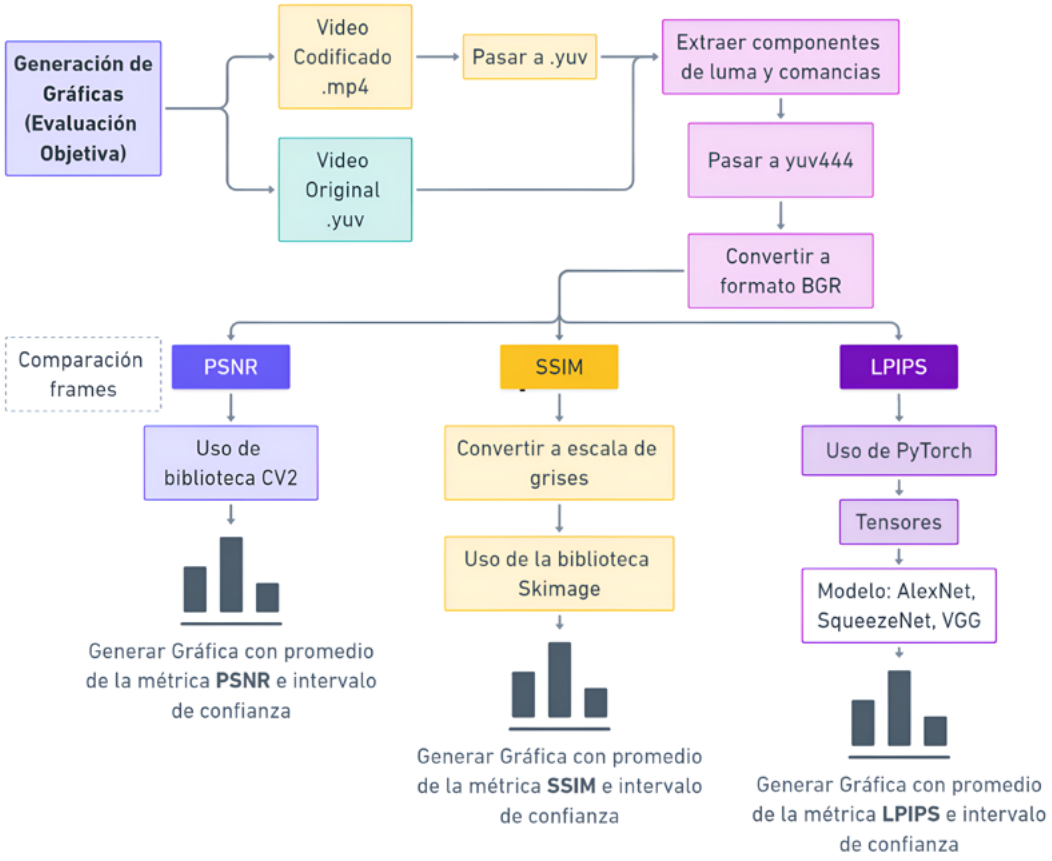
- PSNR: Evaluada con OpenCV, mide la diferencia entre el video original y el codificado en términos de decibelios (dB). Se calcula con la fórmula PSNR = 10log10(L2/MSE), donde L es el valor máximo del píxel y MSE es el error cuadrático medio entre las imágenes. El MSE se define como la Ecuación 1:

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[O(i,j) - D(i,j)]^2 \quad (1)$$

Donde M y N son las dimensiones de la imagen, *O(i,j)* es el píxel original y *D(i,j)* es el píxel distorsionado. Valores entre 30 y 50 dB indican buena calidad (Kotevski y Mitrevski, 2010).

Figura 3

Generación de gráfica evaluación objetiva



- SSIM: Calculada con Skimage, compara la similitud estructural entre dos imágenes. Se basa en tres componentes: luminancia (l), contraste (c) y estructura (s), definidos como las ecuaciones 2, 3, 4.

$$l(i,j) = \frac{2\mu_O\mu_D + C_1}{\mu_O^2 + \mu_D^2 + C_1} \quad (2)$$

$$c(i,j) = \frac{2\sigma_O\sigma_D + C_2}{\sigma_O^2 + \sigma_D^2 + C_2} \quad (3)$$

$$s(i,j) = \frac{\sigma_{OD} + C_3}{\sigma_O\sigma_D + C_3} \quad (4)$$

Donde μ_O y μ_D son las medias de los píxeles en las imágenes original y distorsionada, σ_O y σ_D son las desviaciones estándar, y σ_{OD} es la covarianza. Los valores se combinan usando la fórmula que se muestra en la Ecuación 5.

$$SSIM(i,j) = \frac{(2\mu_O\mu_D + C_1)(2\sigma_O\sigma_D + C_2)}{(\mu_O^2 + \mu_D^2 + C_1)(\sigma_O^2 + \sigma_D^2 + C_2)} \quad (5)$$

El valor resultante varía entre 0 y 1, donde 1 indica la mejor calidad. Este índice se calcula aplicando una ventana de 8x8 píxeles en toda la imagen. De este modo,

resultan tres mapas de índices SSIM, uno por plano de color. Los tres mapas se combinan linealmente en uno, habitualmente otorgando un 80 % del peso a la luminancia y dejando un 10 % a cada plano de croma (Richardson, 2010).

- LPIPS: Para esta métrica se emplean tensores y redes neuronales preentrenadas, como AlexNet, SqueezeNet y VGG para medir la similitud perceptual entre dos imágenes. Se basa en las diferencias entre características profundas extraídas de las imágenes, tales como texturas y bordes. La fórmula general para calcular la distancia perceptual entre dos patchs es la Ecuación 6. Un patch de imagen es un pequeño fragmento cuadrado de una imagen, típicamente de dimensiones 32x32 píxeles, aunque se pueden utilizar otras dimensiones como 16x16, 64x64 píxeles (Zhang et al., 2018).

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} |w^l \odot \left(\widehat{y^l_{hw}} - \widehat{y^l_{0hw}}\right)|_2^2 \quad (6)$$

Donde d(x, x0) es la distancia perceptual entre las imágenes x (original) y x0 (distorsionada), wl es el vector de ponderación, y (y^l)ˆrepresenta las activaciones de la red en la capa l. Un valor bajo de d indica que las imágenes son perceptualmente similares.

Todo el código fuente está liberado en GitHub accediendo al siguiente enlace (Flores, 2024).

Evaluación subjetiva

La evaluación subjetiva fue diseñada para complementar las métricas objetivas, capturando la percepción de la calidad visual por parte de los usuarios. Para ello, se empleó el método MOS (Mean Opinion Score), que mide la calidad percibida en una escala del 1 al 5, con 1 representando "malo" y 5 representando "excelente". La evaluación fue realizada en dos etapas: una donde los participantes evaluaron videos con un solo parámetro de escalabilidad y otra donde se evaluaron videos con múltiples parámetros.

Diseño de experimentos

Para demostrar el funcionamiento de la herramienta y su relación con la evaluación de los usuarios, se desarrollaron dos tipos de pruebas. En la primera, se mostraron los videos con un solo parámetro de escalabilidad, y en la segunda, se utilizaron varios parámetros de escalabilidad, según la Tabla 3, que presenta las categorías y los videos seleccionados. La escalabilidad con el respectivo parámetro se intercaló para mostrar las diferencias a los participantes en la evaluación subjetiva, utilizando los valores de QP (20, 30, 40, 50), FPS (2, 8, 16, 30) y bitrate (100 kbps, 200 kbps, 300 kbps, 400 kbps).

Tabla 3

Video seleccionado por categoría

CATEGORÍA DE VIDEO	ESCALABILIDAD	VIDEO SELECCIONADO
Videovigilancia	Temporal	Hall Monitor
Animaciones	Calidad	Buck Big Bunny
Tiempo real	Espacial	Crew

Calificaciones subjetivas

Las calificaciones subjetivas se recopilaron utilizando una escala de estrellas, como se describe en la Tabla 4. Los participantes evaluaron los videos con base en varios factores, como la pérdida de color, pixelación y fluidez del movimiento. Estas características fueron seleccionadas debido a su relevancia en la percepción de calidad visual en transmisiones de video comprimido.

Materiales de prueba

Siguiendo el diseño de prueba descrito, se seleccionaron 12 secuencias de video divididas en tres categorías: videovigilancia, animaciones y tiempo real, con una resolución CIF (352x288). Los videos fueron codificados usando el códec H.264 y presentados en formato .mp4 a los participantes mediante una pantalla de 86 pulgadas, definición 4K Ultra HD, optimizada para evitar reflejos y asegurar una evaluación consistente de la calidad.

Tabla 4

Video seleccionado por categoría

ESCALA	VALORACIÓN
1eEstrella	Malo
2 estrellas	Deficiente
3 estrellas	Regular
4 estrellas	Bueno
5 estrellas	Excelente

Metodología basada en la ITU-R BT.500-14

El diseño de la evaluación se basó en las recomendaciones de la ITU-R BT.500-14 y sigue la metodología adoptada por Elecard (Kruglov, 2022). Esta metodología asegura que los usuarios no estén al tanto de los parámetros de codificación usados para cada video, eliminando cualquier sesgo que pudiera influir en su percepción.

Participantes que evalúan un solo parámetro de escalabilidad

La muestra consistió en 37 estudiantes, con edades entre 20 y 22 años, de género masculino y femenino. Todos los participantes eran usuarios no expertos en compresión de video, lo que permitió obtener evaluaciones representativas de la experiencia de un espectador promedio. Cada secuencia de video se repitió entre 9 y 10 veces para permitir la identificación clara de diferencias.

Participantes que evalúan distintos parámetros de escalabilidad

Los participantes evaluaron videos con distintos niveles de escalabilidad en tres grupos:

- Grupo 1 (36 estudiantes): Evaluaron videos con los mejores parámetros de escalabilidad (QP bajo, FPS alto, mayor bitrate).
- Grupo 2 (20 estudiantes): Se les presentaron videos con parámetros intermedios.
- Grupo 3 (11 estudiantes): Evaluaron videos con parámetros de escalabilidad bajos (QP alto, FPS bajo, menor bitrate).

Todos los participantes tenían edades entre 19 y 24 años, de género masculino y femenino, y sin experiencia previa en evaluación de calidad de video, observando los videos

de manera natural, como cualquier espectador promedio. Cada grupo realizó la evaluación en sesiones independientes para evitar influencias cruzadas entre ellos.

Análisis comparativo

Los resultados subjetivos se compararon con los datos objetivos obtenidos de las métricas PSNR, SSIM y LPIPS, buscando relacionarse entre la percepción humana y las mediciones automáticas de calidad. Los datos de la evaluación subjetiva proporcionan información valiosa para validar cómo las métricas objetivas reflejan la percepción visual real.

Resultados y Discusión

Evaluación subjetiva

Los resultados de la evaluación subjetiva muestran diferencias significativas en la percepción de calidad según el tipo de escalabilidad: calidad, temporal y espacial. En la Figura 4, siendo escalabilidad de calidad, se observa que el video con un QP de 20 fue el mejor valorado, con un valor promedio de 4.32 en la escala MOS de 5 valores, mientras que la configuración con un QP de 50 obtuvo la peor puntuación (1.16). A medida que aumenta el QP, la calidad percibida disminuye, con un menor error estándar en configuraciones de alta compresión, lo que indica mayor consenso entre los participantes.

En la escalabilidad temporal (Figura 4), el video con 30 FPS alcanzó el mayor promedio de evaluación (4.89), mostrando una clara preferencia por configuraciones más fluidas. La disminución en los FPS redujo la percepción de calidad, siendo la configuración de 2 FPS la peor evaluada (1.19). La baja variabilidad en las evaluaciones de FPS altos refuerza el consenso sobre la superioridad visual en estas configuraciones.

Finalmente, en la escalabilidad espacial (Figura 4), los videos con bitrate de 400 kbps fueron los más valorados (4.7), mientras que los de 100 kbps recibieron las peores calificaciones (1.68). A mayor bitrate, mayor fue la calidad percibida, con menor variabilidad en las opiniones para configuraciones de mayor tasa.

Evaluación objetiva

Se realizaron evaluaciones con las métricas PSNR, SSIM y LPIPS, aplicando diferentes modelos de redes neuronales como AlexNet, SqueezeNet y VGG.

PSNR

La métrica PSNR se evaluó para tres tipos de escalabilidad: calidad, temporal y espacial. En general, los resultados muestran que:

- Escalabilidad de calidad (Figura 5): El PSNR disminuye progresivamente al aumentar el QP, mostrando que una mayor compresión reduce la calidad del video.
- Escalabilidad temporal (Figura 5): No se observan cambios significativos con diferentes tasas de FPS, aunque se evidencia una mayor consistencia a tasas altas.
- Escalabilidad espacial (Figura 5): Existe una relación directa entre el bitrate y la PSNR; videos con mayor bitrate presentan mejor calidad.

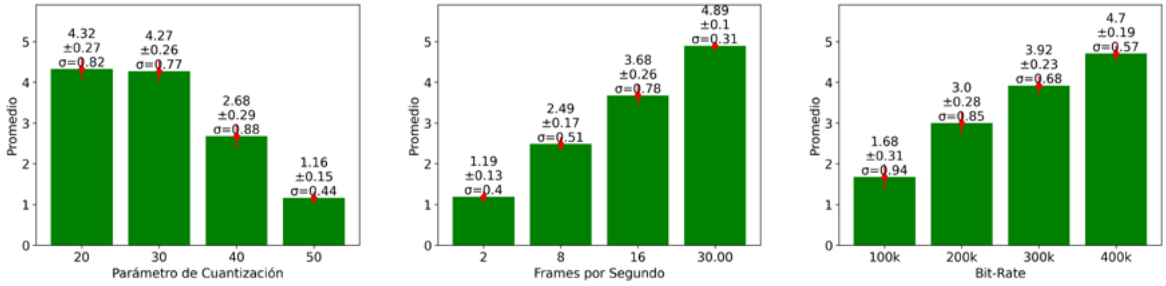
SSIM

El SSIM fue calculado de manera similar para los tres tipos de escalabilidad:

- Escalabilidad de calidad (Figura 6): La similitud estructural disminuye al aumentar el QP, reflejando pérdidas perceptibles en la estructura visual.
- Escalabilidad temporal (Figura 6): SSIM se mantiene relativamente estable, indicando consistencia en la percepción de calidad a distintas tasas de FPS.
- Escalabilidad espacial (Figura 6): Un mayor bitrate se traduce en mejoras en la calidad estructural del video.

Figura 4

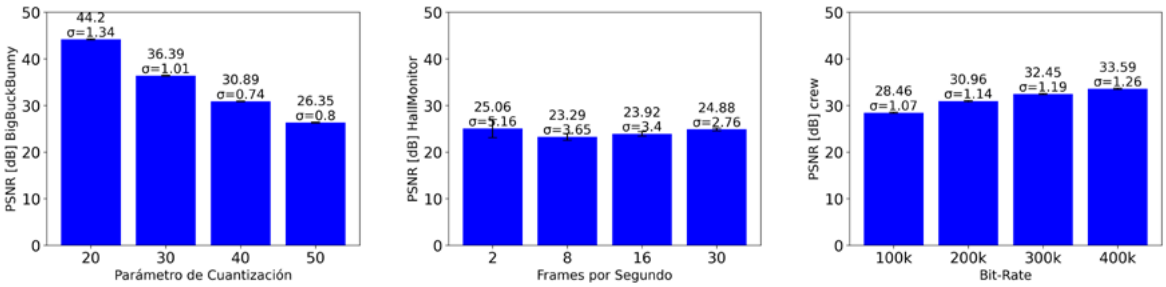
Descripción general de las gráficas de resultados para la evaluación subjetiva con diferentes tipos de escalabilidad: (a) calidad, (b) temporal, y (c) espacial



(a) (b) (c)

Figura 5

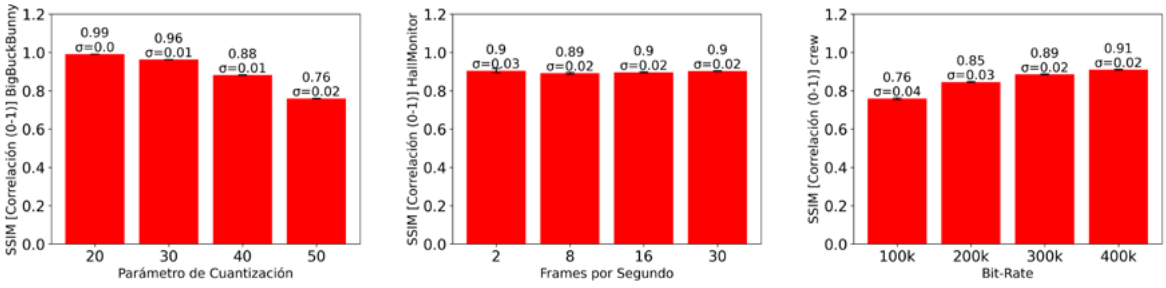
Resultados con la métrica PSNR para diferentes tipos de escalabilidad: (a) calidad, (b) temporal, y (c) espacial



(a) (b) (c)

Figura 6

Resultados con la métrica SSIM para diferentes tipos de escalabilidad: (a) calidad, (b) temporal, y (c) espacial



(a) (b) (c)

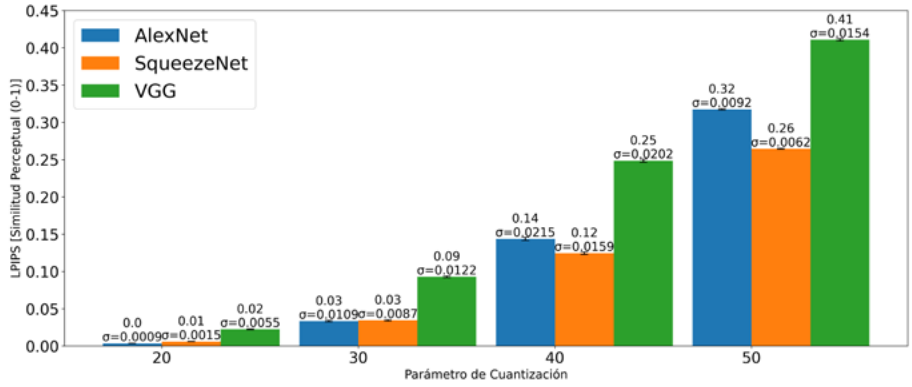
LPIPS

La métrica LPIPS, que se basa en redes neuronales profundas, también fue analizada para los diferentes modelos y parámetros:

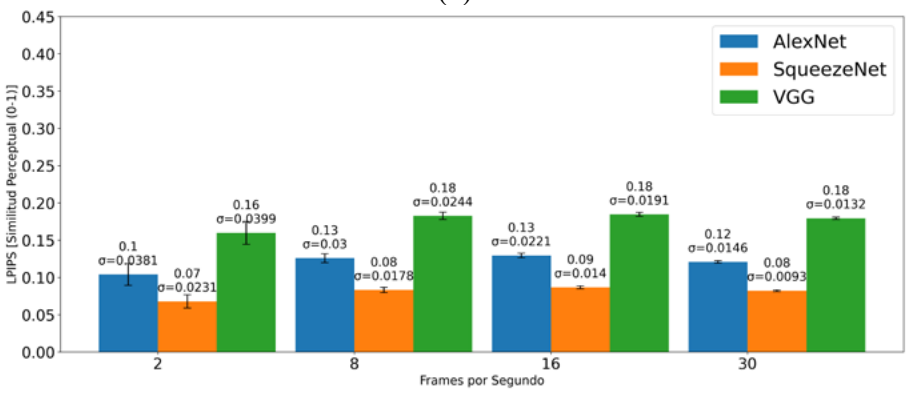
- Escalabilidad de calidad (Figura 7):
  - La métrica LPIPS aumenta con el QP, reflejando que la percepción de calidad disminuye a medida que se incrementa la compresión. VGG mostró la mayor sensibilidad, mientras que AlexNet y SqueezeNet capturan de manera consistente la degradación perceptual.
- Escalabilidad temporal (Figura 7):
  - Los valores de LPIPS se mantienen relativamente estables con distintas tasas de FPS, mostrando ligeras mejoras en la percepción a tasas altas.
- Escalabilidad espacial (Figura 7):
  - Un mayor bitrate se traduce en menor LPIPS, indicando una mejora en la calidad perceptual; la sensibilidad de los distintos modelos sigue la misma tendencia.

Figura 7

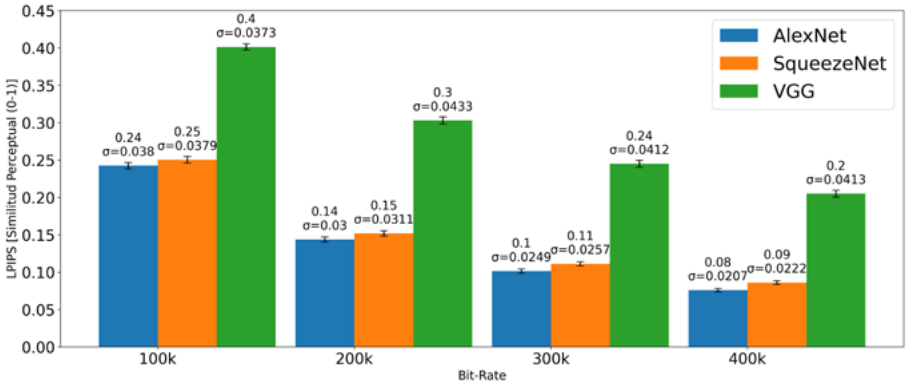
Resultados con la métrica LPIPS para diferentes tipos de escalabilidad: (a) calidad, (b) temporal y (c) espacial



(a)



(b)



(c)

En general, LPIPS confirma las tendencias observadas con PSNR y SSIM: la calidad percibida disminuye con mayor compresión (QP alto) y aumenta con mayores bitrates, mientras que los cambios en FPS tienen un efecto menor. Además, LPIPS refleja con mayor precisión las degradaciones perceptuales percibidas por los espectadores.

Comparación y análisis de las diferentes métricas

La comparación entre evaluaciones subjetivas y métricas objetivas revela una correspondencia notable en las escalabilidades de calidad y espacial, demostrando la efectividad de PSNR, SSIM y LPIPS para reflejar la percepción visual de los espectadores. Los resultados subjetivos fueron contrastados con los valores numéricos y la Figura 8, Figura 9 y Figura 10 ofrecen una referencia cualitativa detallada para cada métrica. Los valores objetivos se agruparon en cuatro niveles según su promedio, donde se seleccionaron las escalas inferiores en pobre-malo debido a los resultados obtenidos.

Relación en la escalabilidad de calidad

Las métricas PSNR y SSIM presentan una relación clara: a mayores valores, mejores calificaciones subjetivas, lo que indica una mayor retención de detalles. LPIPS, por su parte, también respalda esta tendencia, mostrando que los modelos de red más avanzados, como VGG, son más sensibles a las diferencias perceptuales.

Relación en la escalabilidad de calidad

La escalabilidad espacial sigue un patrón similar al de la calidad, con una correspondencia consistente entre valores objetivos (PSNR, SSIM) y percepciones subjetivas. LPIPS aporta un mayor nivel de discriminación, revelando variaciones en la percepción según el modelo de red neuronal utilizado.

Análisis de escalabilidad temporal

En cuanto a la escalabilidad temporal, en la Figura 5 y Figura 6 se observa poca variabilidad con valores más bajos en PSNR (25.06 dB a 23.29 dB) y SSIM (0.904 a 0.892), lo que refleja su limitación para evaluar la fluidez de movimiento. LPIPS, con valores cercanos a cero, también muestra baja variabilidad en la percepción de fluidez, lo que destaca su enfoque en la calidad estática. Sin embargo, sigue la tendencia de las métricas tradicionales, validando su uso como métrica complementaria.

Figura 8

Relación de las puntuaciones MOS con los valores de PSNR por escalabilidad

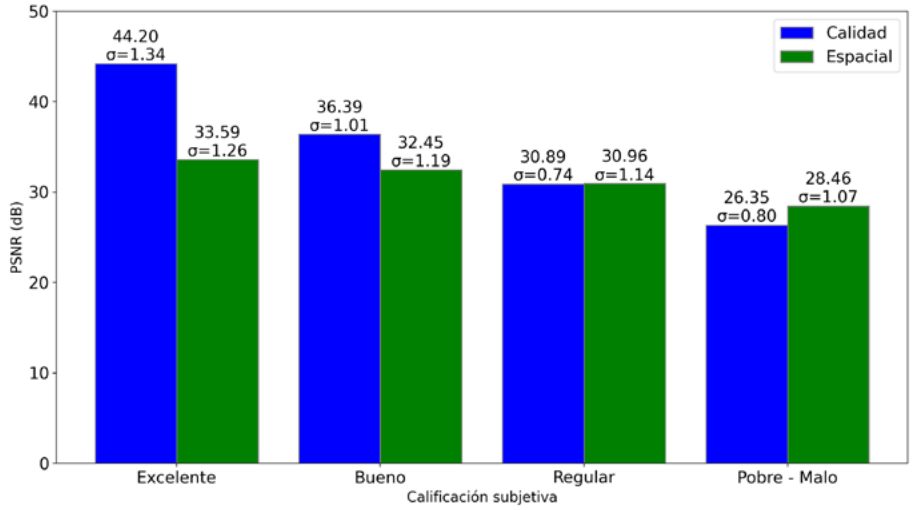


Figura 9

Relación de las puntuaciones MOS con los valores de SSIM por escalabilidad

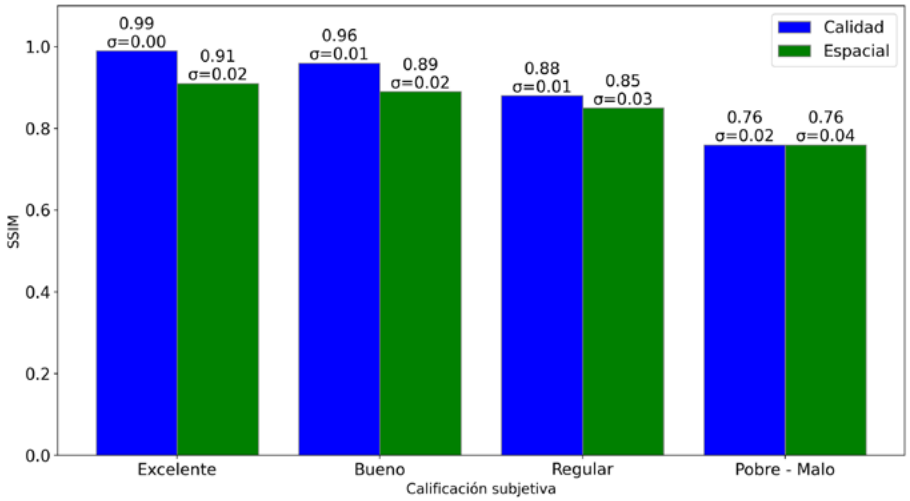
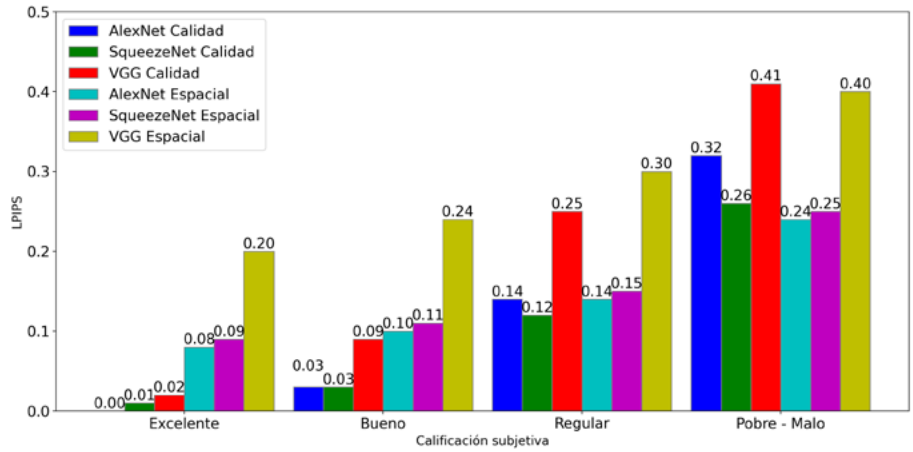


Figura 10

Relación de las puntuaciones MOS con los valores de LPIPS por escalabilidad y modelo



Preferencias de los usuarios

Después de los análisis previos, se llevó a cabo un estudio adicional para evaluar las preferencias de los participantes en tres categorías de contenido: animación, videovigilancia y aplicaciones en tiempo real. Se analizaron tres parámetros de escalabilidad: escalabilidad de calidad (qp), escalabilidad temporal (fps) y escalabilidad espacial (bitrate), utilizando valores fijos de *qp = 20, fps = 30 y bitrate = 400kbps* para el primer conjunto de pruebas.

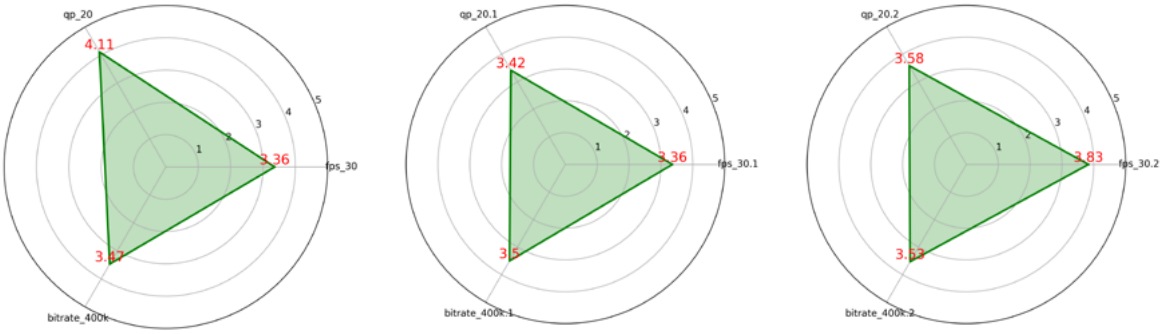
- Animación: Los participantes priorizaron la calidad visual, con *qp = 20* recibiendo la puntuación más alta (promedio 4.11). La escalabilidad espacial y temporal obtuvo valores menores, reflejando que la fidelidad visual es el factor determinante en este tipo de contenido, como se muestra en la Figura 11.
- Videovigilancia: Las preferencias fueron más equilibradas, sin embargo, la escalabilidad espacial (bitrate) fue ligeramente más valorada (promedio 3.50), indicando que una mayor tasa de bits contribuye a la percepción de detalles importantes en este contexto, como se observa en la Figura 11.
- Aplicaciones en tiempo real: Se observó una clara preferencia por la escalabilidad temporal, priorizando la fluidez del video (*fps = 30*, promedio 3.83). Los participantes valoraron menos la calidad y el bitrate, coherente con la importancia de la continuidad de movimiento en estas aplicaciones. (ver Figura 11).

Las preferencias de los usuarios varían según el tipo de contenido: la calidad visual es clave en animación, la escalabilidad espacial es relevante en video vigilancia y la fluidez temporal domina en aplicaciones en tiempo real.

Se realizó una segunda evaluación utilizando valores de *qp = 30, fps = 16 y bitrate = 300kbps*. Los resultados fueron los siguientes:

Figura 11

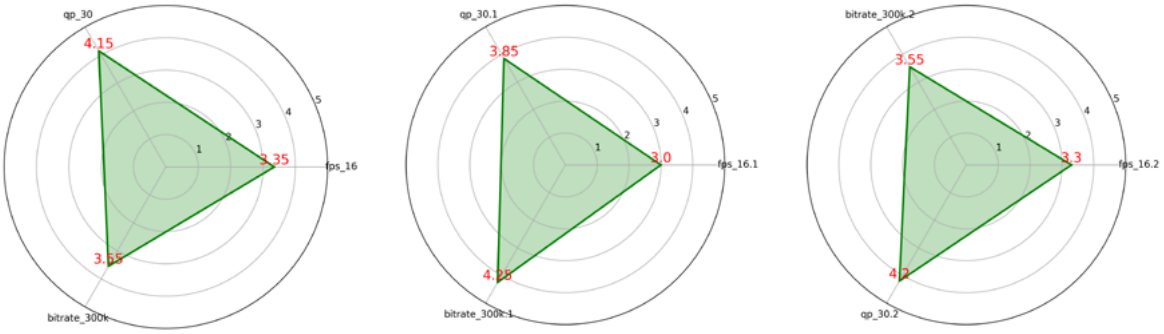
Resultados con la calificación subjetiva regular evaluación 1 para diferentes tipos de categorías: (a) animación, (b) video vigilancia y (c) tiempo real



(a) (b) (c)

Figura 12

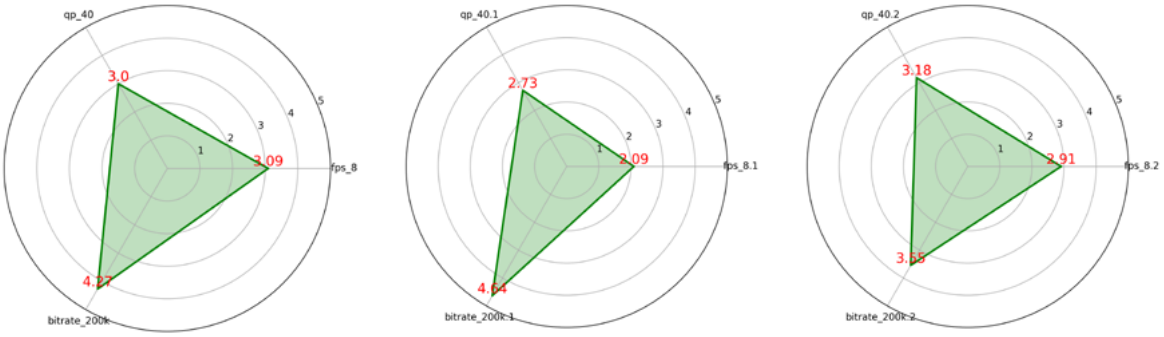
Resultados con la calificación subjetiva regular evaluación 2 para diferentes tipos de categorías: (a) animación, (b) video vigilancia, y (c) tiempo real



(a) (b) (c)

Figura 13

Resultados con la calificación subjetiva regular evaluación 3 para diferentes tipos de categorías: (a) animación, (b) video vigilancia y (c) tiempo real



(a) (b) (c)

- Animación: Los participantes continuaron priorizando la calidad visual; *qp = 30* recibió la puntuación más alta (promedio 4.15), mientras que bitrate y fps fueron menos valorados, lo que confirma la importancia de la fidelidad visual en esta categoría (ver Figura 12).
- Videovigilancia: La escalabilidad espacial (bitrate) fue claramente preferida (promedio 4.25), superando a qp y fps, lo que indica que la percepción de detalles visuales es más relevante en este tipo de contenido (ver Figura 12).
- Aplicaciones en tiempo real: En esta condición de menor bitrate, la escalabilidad espacial se convirtió en el parámetro más valorado (promedio 4.20), mientras que qp y fps recibieron puntuaciones más bajas, lo que mostró que los usuarios priorizan mantener calidad visual sobre fluidez cuando los recursos son limitados (ver Figura 12).

Con una reducción de fps y bitrate, las preferencias de los participantes se ajustan: la calidad visual sigue siendo clave en animación, la escalabilidad espacial domina en videovigilancia y en aplicaciones en tiempo real, los usuarios priorizan la calidad espacial cuando los recursos son restringidos.

Finalmente, la evaluación se realizó con los parámetros de escalabilidad de *qp = 40, fps = 8 y bitrate =200kbps*. Los resultados fueron los siguientes:

- Animación: Los participantes priorizaron la escalabilidad espacial (bitrate), con la puntuación más alta (promedio 4.27), lo que indica que la calidad espacial adquiere mayor relevancia cuando la calidad global disminuye (ver Figura 13).
- Videovigilancia: La escalabilidad espacial fue aún más destacada (promedio 4.64), superando ampliamente a qp y fps, lo que refleja la importancia crítica del bitrate para percibir detalles en este tipo de contenido (ver Figura 13).
- Aplicaciones en tiempo real: Aunque la fluidez sigue siendo importante, la escalabilidad espacial predominó (promedio 3.55), lo que mostró que cuando los recursos se reducen, mantener la calidad visual es más valorado que la tasa de cuadros o la calidad global (ver Figura 13).

En condiciones de menor calidad y recursos limitados, la escalabilidad espacial se convierte en el parámetro más relevante en todas las categorías de contenido, destacando su papel clave en la percepción de calidad cuando se prioriza la visibilidad de detalles sobre la fluidez o la calidad general.

Los resultados objetivamente muestran que PSNR y SSIM reflejan correctamente la degradación general de la calidad con cambios en qp y bitrate, lo que coincide con Huynh-Thu y Ghanbari (2012) e Izquierdo (2017). Sin embargo, su limitada sensibilidad a distorsiones perceptuales coincide con hallazgos de Hou et al. (2022) y J. Wang et al. (2023).

LPIPS demostró mayor alineación con la percepción humana, coherente con Prashnani et al. (2018), R. Zhang et al. (2018) y Gu et al. (2020). Las diferencias entre modelos de redes profundas reflejan observaciones similares en K. Zhang et al. (2021) y S. Zhang et al. (2023).

En las evaluaciones subjetivas, la preferencia por calidad visual en animación y por escalabilidad espacial en videovigilancia y aplicaciones en tiempo real coincide con estudios de Kastryulin et al. (2023) y Danier et al. (2022). Sin embargo, la prioridad de la escalabilidad espacial sobre fps en condiciones de bajo bitrate para aplicaciones en tiempo real muestra una

ligera discrepancia con trabajos previos que destacan la fluidez como más crítica, sugiriendo dependencia del tipo de contenido y del nivel absoluto de calidad (Danier et al., 2022; Li et al., 2019).

Validación de las métricas con la percepción humana

Antes de presentar los resultados, es importante definir cómo se calculan y se interpretan los valores porcentuales de sensibilidad utilizados en este experimento. Para evaluar qué métrica (LPIPS, PSNR, SSIM) es más sensible a la percepción humana ante diferentes tipos de distorsiones, la sensibilidad se calculó mediante la fórmula (7).

$$\text{Valor } \% = \left(\frac{\text{Métrica Distorsionada} - \text{Métrica Original}}{\text{Métrica Original}}\right) \times 100 \quad (7)$$

En esta ecuación, el valor porcentual representa el cambio relativo entre el valor de la métrica en una imagen distorsionada y su valor en la imagen original y se expresa en forma de porcentaje.

Valores porcentuales superiores al 100 % indican que la métrica detecta una distorsión mucho mayor que la esperada en comparación con la imagen original, lo que sugiere una alta sensibilidad a los cambios introducidos.

Valores negativos se interpretan como un resultado en el que la métrica percibe la imagen distorsionada como "mejor" o más similar a la original que la imagen base, lo que puede indicar que la métrica no es adecuada para detectar ese tipo de distorsiones.

Los resultados se presentan en la Figura 14, Figura 15 y Figura 16, donde se muestra un aumento gradual de la distorsión en el fotograma para los tres tipos de métricas evaluadas.

La Figura 14 muestra que LPIPS es considerablemente más sensible a distorsiones localizadas en elementos clave, como el rostro humano. En este caso, LPIPS detectó con mayor precisión las alteraciones en el rostro, presentando niveles de sensibilidad de 126.15 %, 19.83 %, 48.61 % y 73.64 % cuando solo se distorsionó el rostro. Para el fotograma completamente distorsionado, los valores fueron similares, mostrando 126.91 %, 17.54 %, 37.22 % y 48.21 %. Estos resultados sugieren que LPIPS es especialmente eficaz al identificar cambios en objetos reconocibles como rostros, lo que refleja una mayor sensibilidad de la métrica ante distorsiones que afectan áreas de importancia perceptual.

En contraste, las métricas tradicionales como PSNR y SSIM demostraron ser menos sensibles a este tipo de alteraciones. Para PSNR, los niveles de sensibilidad fueron menores, registrando valores de -9.88 %, -2.23 %, -4.78 % y -6.97 %, mientras que SSIM mostró valores de -1.33 %, -0.44 %, -1.10 % y -1.75 %. Esto indica que, aunque las distorsiones eran evidentes, PSNR y SSIM no lograron capturar con la misma precisión los cambios perceptuales relevantes, lo que sugiere, desde la perspectiva de estas métricas, que la imagen distorsionada sigue siendo similar a la original, en especial para elementos importantes como los rostros.

En la Figura 15 y Figura 16, que presentan las distorsiones de ruido Poisson y ruido de sal y pimienta, se observa una tendencia similar. LPIPS nuevamente mostró valores más altos, lo que indica mayor sensibilidad a estas distorsiones, tanto para el fotograma completamente distorsionado como para una parte específica de la imagen. En cambio, PSNR

y SSIM mantuvieron valores bajos y negativos en algunos casos, reflejando que estas métricas tradicionales no capturan adecuadamente las alteraciones perceptuales más finas.

Cabe mencionar que para estos experimentos se utilizó la red neuronal VGG en LPIPS, aunque también se realizaron pruebas con otras redes preentrenadas como AlexNet y SqueezeNet, obteniendo resultados similares. En todas las pruebas, LPIPS mostró mayor sensibilidad en comparación con las métricas tradicionales, lo que sugiere que las redes neuronales profundas capturan mejor las alteraciones perceptuales que afectan la percepción humana.

Figura 14

Resultado de la métricas para distintos tipos de distorsión: difuminado

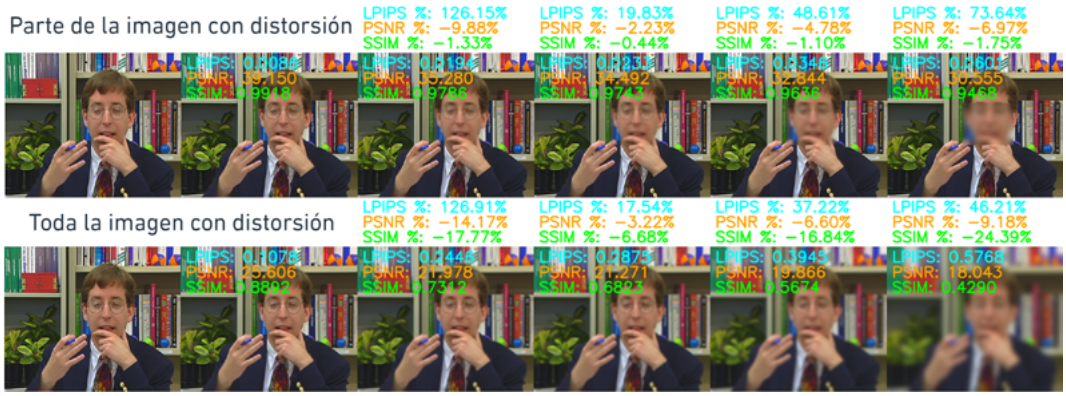
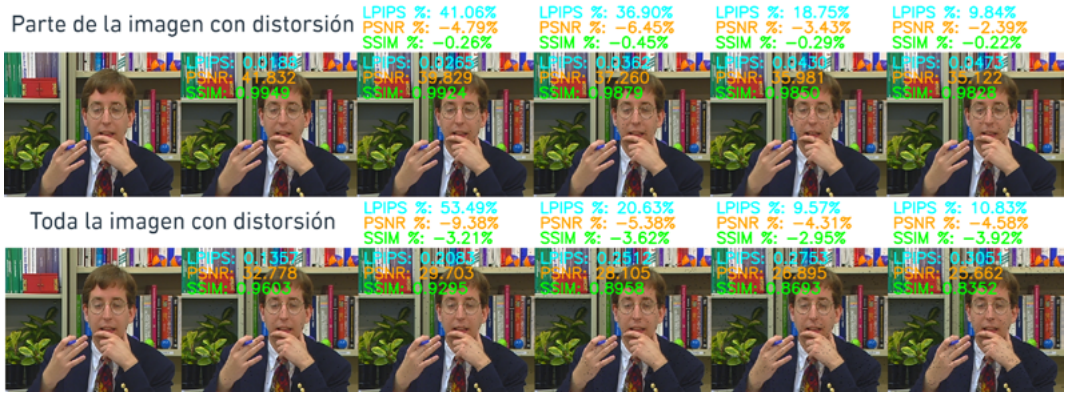


Figura 15

Resultado de la métricas para distintos tipos de distorsión: ruido de Poisson

Figura 16

Resultado de la métricas para distintos tipos de distorsión: ruido sal y pimienta



Conclusiones

Este estudio presentó una herramienta para la evaluación de la calidad de video, enfocada en configuraciones de escalabilidad en calidad (qp, Quantization Parameter), resolución espacial (bitrate) y frecuencia temporal (Frames per Second, FPS). Una de las principales contribuciones de la herramienta es su interfaz gráfica de usuario (GUI, Graphical User Interface), que simplifica la carga de videos en formato YUV y permite a los usuarios seleccionar métricas de evaluación como Peak Signal-to-Noise *Ratio* (PSNR), Structural Similarity Index (SSIM) y Learned Perceptual Image Patch Similarity (LPIPS). Esto la hace accesible a usuarios con diferentes niveles de experiencia técnica y especialmente útil en escenarios donde se requiere evaluar la calidad bajo distintos parámetros de codificación.

En cuanto a los resultados de la evaluación subjetiva, se confirmó que PSNR y SSIM reflejan adecuadamente tendencias generales en la degradación de calidad, mientras que LPIPS mostró una mayor coherencia con la percepción humana, coincidiendo con trabajos recientes en el área. Estos hallazgos refuerzan la pertinencia de integrar métricas perceptuales basadas en aprendizaje profundo en herramientas de análisis de video.

Desde una perspectiva práctica, la herramienta tiene potencial aplicación en entornos de transmisión adaptativa de video (Adaptive Streaming), plataformas de videoconferencia, sistemas de videovigilancia y telemedicina, donde es crucial equilibrar calidad visual, fluidez y eficiencia en el uso del ancho de banda. Al permitir comparar configuraciones de escalabilidad y vincular métricas objetivas con la percepción de los usuarios, esta herramienta puede apoyar la toma de decisiones en la selección de parámetros de codificación más adecuados para cada contexto.

Finalmente, se identificó como limitación que las métricas empleadas no capturan de forma adecuada la fluidez del movimiento, ya que se centran en la comparación de cuadros individuales y no en la continuidad temporal. Esto resalta la necesidad de integrar métricas orientadas a la percepción del movimiento, como FloLPIPS, para optimizar la evaluación en escenarios donde la escalabilidad temporal sea determinante.

Reconocimientos y Declaraciones

Los autores agradecen al Departamento de Eléctrica, Electrónica y Telecomunicaciones de la Facultad de Ingeniería de la Universidad de Cuenca por el apoyo y soporte brindado en el desarrollo de este trabajo de investigación.

Los autores declaran que, en la elaboración del presente artículo, se ha utilizado la herramienta de IA ChatGPT (versión GPT-4o, OpenAI) únicamente como apoyo en la redacción y mejora del estilo de redacción en la sección de Introducción. El diseño de la investigación, la obtención y análisis de los resultados, así como las conclusiones, son responsabilidad exclusiva de los autores.

Los autores declaran la contribución y participación equitativa de roles de autoría para esta publicación.

Referencias

Bowker, D. (2021). *Bitrate Defined: How It Impacts Video Quality*. https://artlist.io/blog/what-is-bitrate/

Chen, Z., Hu, B., Niu, C., Chen, T., Li, Y., Shan, H., & Wang, G. (2023). *IQAGPT: Image Quality Assessment with Vision-language and ChatGPT Models*. https://doi.org/10.48550/arXiv.2312.15663

Danier, D., Zhang, F., & Bull, D. (2022). Flo LPIPS: A bespoke video quality metric for frame interpolation. *2022 Picture Coding Symposium (PCS)*, 283–287. https://doi.org/10.48550/arXiv.2207.08119

Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(5), 2567–2581. https://doi.org/10.48550/arXiv.2004.07728

Ding, K., Zhong, R., Wang, Z., Yu, Y., & Fang, Y. (2023). Adaptive Structure and Texture Similarity Metric for Image Quality Assessment and Optimization. *IEEE Transactions on Multimedia*, 1–13. https://doi.org/10.1109/TMM.2023.3333208

FFmpeg Developers. (2024). *Introduction to FFmpeg*. FFmpeg. https://ffmpeg.org/about.html

Flores, B. (2024). Evaluación calidad video [Repositorio en GitHub]. GitHub. https://github.com/Akilescasteo/Evaluacion_calidad_video

Xiph Foundation. (2023). Xiph.org video test media [derf's collection]. https://media.xiph.org/video/derf/

Gu, J., Cai, H., Chen, H., Ye, X., Ren, J., & Dong, C. (2020). *PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration*. https://doi.org/10.48550/arXiv.2007.12142

Gu, J., Cai, H., Dong, C., Ren, J. S., Timofte, R., Gong, Y., Lao, S., Shi, S., Wang, J., Yang, S., Wu, T., Xia, W., Yang, Y., Cao, M., Heng, C., Fu, L., Zhang, R., Zhang, Y., Wang, H., … Tiwari, A. K. (2022). NTIRE 2022 Challenge on Perceptual Image Quality Assessment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 951–967. https://doi.org/https://doi.org/10.48550/arXiv.2206.11695

Hou, Q., Ghildyal, A., & Liu, F. (2022). A perceptual quality metric for video frame interpolation. *European Conference on Computer Vision*, 234–253. https://doi.org/10.48550/arXiv.2210.01879

Huynh-Thu, Q., & Ghanbari, M. (2012). The accuracy of PSNR in predicting video quality for different video scenes and frame rates. *Telecommunication Systems, (49)*, 35–48. https://doi.org/10.1007/s11235-010-9351-x

García Izquierdo, F. (2017). *Desarrollo de una herramienta para la medida de calidad de vídeo* [Tesis de grado, Universidad de Sevilla]. https://biblus.us.es/bibing/proyectos/abreproy/91129/fichero/Memoria+TFG+-+Desarrollo+de+una+herramienta+para+la+medida+de+calidad+de+video.pdf

Kastryulin, S., Zakirov, J., Pezzotti, N., & Dylov, D. V. (2023). Image Quality Assessment for Magnetic Resonance Imaging. *IEEE Access, 11*, 14154–14168. https://doi.org/10.1109/ACCESS.2023.3243466

Kotevski, Z., & Mitrevski, P. (2010). Experimental comparison of PSNR and SSIM metrics for video quality estimation. *International Conference on ICT Innovations*, 357–366. https://doi.org/10.1007/978-3-642-10781-8_37

Kruglov, A. (2022). *Interpretation of Objective Video Quality Metrics*. Elecard: Video Compression Guru. https://www.elecard.com/page/article_interpretation_of_metrics

Li, D., Jiang, T., & Jiang, M. (2019). Quality assessment of in-the-wild videos. *Proceedings of the 27th ACM International Conference on Multimedia*, 2351–2359. https://doi.org/10.48550/arXiv.1908.00375

Prashnani, E., Cai, H., Mostofi, Y., & Sen, P. (2018). PieAPP: Perceptual image-error assessment through pairwise preference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1808–1817. http://doi.org/10.1109/CVPR.2018.00194

Richardson, I. E. G. (2003). *H.264 and MPEG-4 video compression: Video coding for next-generation multimedia* (2. ed.). Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/0470869615

Wang, J., Chan, K. C. K., & Loy, C. C. (2023). Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(2), 2555–2563. https://doi.org/10.1609/aaai.v37i2.25353

Wang, L. (2021). *How Bitrate and Quantization Parameter (QP) Affect Video Quality*. https://lesliewongcv.github.io/posts/2021/11/blog-post-1/

Watt, J. (2022). *6 Factors Decide Video Quality: Resolution, Bitrate, Frame Rate, CRF, Bit Depth*. https://www.winxdvd.com/video-transcoder/6-factors-decide-video-quality-bitrate-resolution-framerate.htm

Zhang, K., Liang, J., Van Gool, L., & Timofte, R. (2021). Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4791–4800. https://doi.org/10.48550/arXiv.2103.14006

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595. https://doi.org/https://doi.org/10.48550/arXiv.1801.03924

Zhang, S., Lin, Z., & Zhou, Y. (2023). *Accelerate diffusion based human image generation via consistency models*. https://dx.doi.org/10.2139/ssrn.4719919