https://doi.org/10.24867/GRID-2024-p24 Original scientific paper

OBJECTIVE VIDEO QUALITY COMPARISON OF USER-GENERATED CONTENT PLATFORMS

Jaroslav Svoboda
Prague University of Economics and Business, Faculty of Informatics and Statistics, Prague, Czech

Abstract: *Uploading videos to platforms for sharing has been a core feature of the internet for nearly two decades. Advances in video encoding algorithms, their standardization, and increased computational power have significantly improved the visual quality of online videos. This paper evaluates the video quality across several popular platforms, including YouTube, Facebook, and Vimeo. Using three test sequences with varying content characteristics, the study employs full-reference objective video quality metrics-VMAF, PSNR-HVS-M, and MS-SSIM-to assess video quality at resolutions supported by all platforms. The results provide insights into the comparative performance of each platform in delivering high-quality video.*

Key words: PSNR-HVS-M, video, quality, VMAF, MS-SSIM, VOD, compression

1. INTRODUCTION

Video on Demand (VOD) has become the dominant mode of content consumption in recent years, with numerous platforms offering a wide range of media, including movies, series, television shows, and music videos. These platforms typically deliver professionally produced content created by film studios, television broadcasters, or the streaming services themselves, continuing a model that has been in place for over a century, though continually enhanced by technological advancements in production, distribution, and consumption.

In parallel, the internet has given rise to a new category of content-videos created by users for users. Platforms like YouTube, among others, have enabled this type of user-generated content to thrive, offering diverse content that ranges from entertainment to educational videos. This shift in content creation has reshaped how video is produced, distributed, and consumed globally.

A key challenge for video platforms, whether hosting professional or user-generated content, is to deliver the best possible audiovisual quality while minimizing data usage and ensuring compatibility across a wide range of devices and formats. Efficient video coding standards and encoding settings are critical to achieving these goals. Fortunately, there are multiple methods available for evaluating and optimizing the visual quality of streamed video. This paper examines the visual quality of selected video platforms that host user-generated content, focusing on the encoding strategies employed and their impact on user experience.

2. FULL-REFERENCE OBJECTIVE VIDEO QUALITY EVALUATION TECHNIQUES

Video quality assessment (VQA) is vital for ensuring a high-quality viewing experience in various applications such as streaming, broadcasting, and video conferencing. Full-reference (FR) objective metrics assess video quality by comparing a test video to its original, undistorted reference version. These methods are particularly valuable as they offer objective, reproducible measures of video quality. This chapter will explore several widely-used FR objective video quality metrics, including Peak Signal-to-Noise Ratio (PSNR), PSNR-Human Visual System (PSNR-HVS), PSNR-HVS Masking (PSNR-HVS-M), Structural Similarity Index (SSIM), Multiscale SSIM (MS-SSIM), and Video Multimethod Assessment Fusion (VMAF). The discussion will focus on how each method operates, its advantages and limitations, and how VMAF integrates various approaches to improve upon traditional techniques.

2.1 Peak Signal-to-Noise Ratio (PSNR)

PSNR is perhaps the most basic and widely used objective metric for video quality assessment. It is calculated by comparing the pixel-wise differences between the reference video and the distorted video,

using the mean squared error (MSE) as a measure of the distortion. The formula for PSNR is given by Equation 1:

$$PSNR = 10 \times log_{10}(\frac{MAX^2}{MSE}) \quad (1)$$

Where MAX is the maximum possible pixel value (for instance, 255 for 8-bit video), and MSE is the mean squared error between the pixel values of the reference and test video frames. PSNR is often favored because of its simplicity and ease of implementation. Despite its widespread use, PSNR has notable limitations. It does not take into account the human visual system (HVS), meaning that it often fails to correlate well with subjective human perception of video quality. Specifically, PSNR treats all pixel errors equally, even though the HVS is more sensitive to certain types of distortions (such as those occurring in high-contrast or edge areas). As a result, PSNR may report small numerical differences in pixel values as major quality losses, even when these differences are not perceptually significant (Gonzalez, 2009).

2.2 PSNR-Human Visual System (PSNR-HVS)

To address the shortcomings of PSNR, PSNR-HVS incorporates certain properties of the human visual system. One of the key improvements is the integration of the Contrast Sensitivity Function (CSF), which models the sensitivity of the human eye to different spatial frequencies. In essence, PSNR-HVS weighs the frequency components of the image based on the HVS's sensitivity to different frequencies. This allows the metric to prioritize distortions that are more noticeable to humans, improving its correlation with subjective quality assessments. While PSNR-HVS offers better alignment with human perception than traditional PSNR, it still only accounts for basic aspects of the HVS. Specifically, it models contrast sensitivity but does not include other important perceptual phenomena, such as visual masking. Thus, while PSNR-HVS is an improvement over PSNR, it remains limited in its ability to accurately reflect subjective visual quality in all situations (Egiazarian et al, 2006).

2.3 PSNR-Human Visual System Masking (PSNR-HVS-M)

PSNR-HVS-M builds upon PSNR-HVS by incorporating a more comprehensive model of the HVS, including visual masking effects. Visual masking is the phenomenon where certain visual details become less noticeable in the presence of other visual stimuli, particularly in textured or high-contrast regions of a video. By modeling this effect, PSNR-HVS-M can better account for areas of the video where distortions might be less visible to human observers, leading to a more accurate reflection of perceived video quality. Compared to both PSNR and PSNR-HVS, PSNR-HVS-M demonstrates improved correlation with human perception, particularly for content with complex textures or motion. However, this comes at the cost of increased computational complexity. Additionally, PSNR-HVS-M may not always perform optimally across different types of video content and compression schemes, as the degree of visual masking can vary depending on the scene (Ponomarenko et al., 2007).

2.4 Structural Similarity Index (SSIM)

SSIM represents a shift from traditional pixel-wise comparison methods like PSNR and its variants. Instead of focusing on absolute differences between pixel values, SSIM measures the structural similarity between the reference and distorted videos. It assesses three components: luminance, contrast, and structural information (the local spatial pattern of pixel intensities). SSIM is computed over local windows, and the final SSIM score is the mean of all local comparisons. The mathematical formulation for SSIM is based on comparisons of luminance, contrast, and structure between corresponding windows in the reference and distorted images. This approach allows SSIM to focus on preserving structural information, which is more important for human perception than mere pixel-level accuracy. As a result, SSIM tends to correlate better with subjective quality assessments than PSNR. However, SSIM is not without limitations. While it performs well for certain types of distortions, it can still fail to capture quality degradations in videos with more complex content. Additionally, because it operates on local windows, SSIM may not capture quality variations that occur over larger regions of the video (Wang et al., 2004).

2.5 Multiscale Structural Similarity Index (MS-SSIM)

MS-SSIM extends SSIM by introducing a multiscale analysis, which better reflects the multiresolution characteristics of the HVS. In MS-SSIM, the video is analyzed at multiple scales by progressively downsampling the reference and test videos. SSIM is computed at each scale, with more weight assigned to coarser scales where human perception is more sensitive to structural changes. By incorporating information across multiple scales, MS-SSIM provides a more robust assessment of video quality, especially for videos with varying levels of detail. MS-SSIM typically achieves higher correlations with subjective quality scores than single-scale SSIM and can better capture complex distortions such as those introduced by compression. While MS-SSIM offers significant improvements over SSIM and PSNR-based methods, it is computationally more intensive due to its multiscale nature. Additionally, its performance can still be limited by specific types of visual artifacts that are not well captured by structural similarity metrics (Wang et al., 2003).

2.6 Video Multimethod Assessment Fusion (VMAF)

VMAF is a more recent development in FR video quality assessment and differs significantly from the previously discussed methods. Developed by Netflix, VMAF is a machine-learning-based approach that combines multiple video quality metrics (including SSIM, PSNR, and motion information) into a single, unified score. VMAF uses subjective quality scores from human viewers to train its model, allowing it to better predict perceived video quality. One of the key strengths of VMAF is its ability to integrate information from multiple metrics, making it more robust across a wide variety of video content and distortions. VMAF not only includes traditional pixel-based metrics like PSNR but also incorporates features related to motion and temporal artifacts, which are important for video quality. Additionally, VMAF allows for customization, enabling its model to be fine-tuned for specific use cases or content types. Compared to metrics like PSNR, SSIM, and even MS-SSIM, VMAF consistently achieves higher correlations with subjective quality assessments. This makes it one of the most accurate and reliable FR metrics for modern video quality assessment tasks. However, VMAF is computationally demanding due to its use of multiple metrics and machine learning models. Moreover, VMAF's performance is highly dependent on the quality of the training data used to develop the model, meaning that it may need to be retrained for new content types or viewing conditions (Li et al., 2016).

3. STREAMING SERVICES

Table 1 summarizes the features of the selected platforms, focusing on key factors such as the maximum allowed file size and support for various frame rates, including high frame rate content, which was critical in the context of this paper.

Table 1: Multimedia features of selected video platforms

	Maximum upload filesize	HFR	HDR support	Surround sound	3D video	360° video	25/30 FPS
YouTube	256 GB (verified account)	Yes	Yes	5.1	Yes	Yes	Both
Facebook	4GB (regular user)	No	Tonemapping	No	Yes	Yes	Both
Instagram	Unknown	No	Tonemapping	No	No	No	Both
Vimeo	1GB Free	Yes	Tonemapping	7.1	Yes	Yes	Both
Dailymotion	4GB Free 64GB Premium	Yes	No	No	No	No	Both

3.1 Video Platform Selection

For this study, five video platforms offering video upload and sharing services were selected for evaluation:

- YouTube: A video-sharing platform owned by Google, boasting 2.5 billion users worldwide. It supports a wide range of video formats and resolutions, making it one of the most popular platforms for content creators.
- Facebook: A social network owned by Meta with over 3 billion users. Facebook allows users to share videos on personal profiles, pages, and groups, and supports a variety of video resolutions.
- Instagram: Also owned by Meta, Instagram has 2 billion users and focuses primarily on short-form video content through its Reels and Stories features, although longer videos can be uploaded to IGTV.
- Vimeo: A SaaS-based video hosting and sharing service with 300 million users. Vimeo is widely used by creative professionals and offers high-quality video uploads with extensive customization and privacy settings.
- Dailymotion: An online video-sharing platform owned by Vivendi, which supports a broad range of video content types and has a user base of millions globally.

Some platforms were excluded from this evaluation due to limitations that would have impacted the study's objectives. For instance, X (formerly known as Twitter) was not included because it restricts video resolution to 720 p for free users, with 1080 p available only to Premium subscribers. Additionally, at the time of writing, the platform experienced issues with video encoding, as illustrated in Figure 1. Twitch, a streaming platform focused on gaming, was also excluded. While Twitch allows video uploads, this feature is restricted to Affiliate or Partner accounts, which were not available for this study. Table 2 provides an overview of the available resolutions and video coding standards used by the selected video platforms. It is important to note that these platforms do not apply all supported formats to every video; certain formats are reserved for specific resolutions, popular content, or premium users. For the purposes of this quality evaluation, only the resolutions supported by all platforms were considered.

Table 2: Supported resolutions and available video coding standards of selected video platforms for SDR content (resolutions tested in this article are highlighted)

Resolution and framerate	YouTube	Facebook	Instagram	Vimeo	Dailymotion
7680 × 4320@60	AV1				
7680 × 4320@30	AV1				
3840 × 2160@60	AV1, VP9			H.264	H.264
3840 × 2160@30	AV1, VP9			H.264	H.264
2560 × 1440@60	AV1, VP9			H.264	H.264
2560 × 1440@30	AV1, VP9			H.264	H.264
1920 × 1080@60	AV1, VP9, H.264			H.264	H.264
1920 × 1080@30	AV1, VP9, H.264	AV1, H.264	VP9	H.264	H.264
1280 × 720@60	AV1, VP9, H.264			H.264	H.264
1280 × 720@30	AV1, VP9, H.264	AV1, H.264	VP9	H.264	H.264
960 × 540@30		AV1, H.264		H.264	
848 × 480@30					H.264
640 × 360@30	AV1, VP9, H.264	AV1, H.264	VP9	H.264	H.264
512 × 288@30					H.264
426 × 240@30	AV1, VP9, H.264			H.264	
256 × 144@30	AV1, VP9, H.264				
256 × 144@15	VP9				

3.2 Source Files Preparation, Upload and Download

Three video sequences were selected as source material for testing. Each sequence was converted to three different resolutions and frame rates, as outlined in Table 1, using FFmpeg 7.0.2 and x265 3.6. The following encoding parameters were applied: "-x265-params lossless=1 -an -map_metadata -1 -pix_fmt yuv420p", ensuring the highest possible quality, with 8-bit channel depth, 4:2:0 chroma subsampling, and no

metadata. This approach was chosen to create high-quality reference sequences suitable for accurate visual quality evaluation.

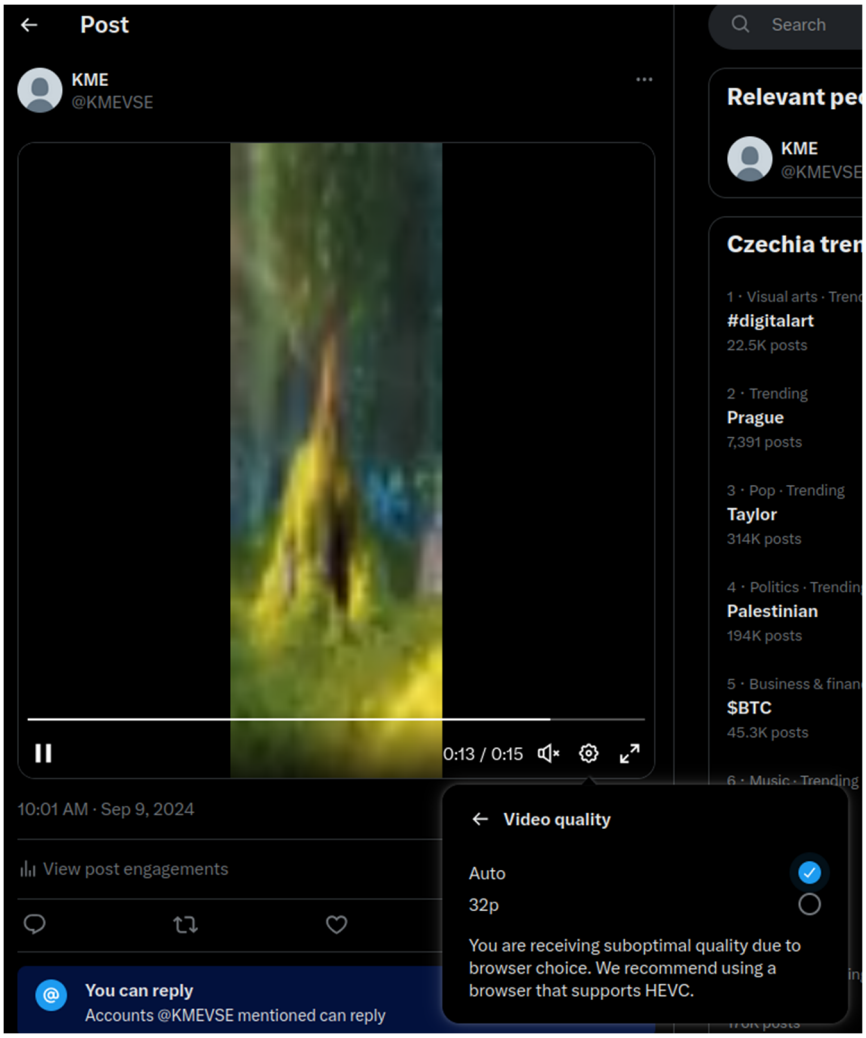

Figure 1: Erroneous video encode of social platform X

The first sequence was taken from the open-source animated film *Big Buck Bunny*. A 15-second clip was extracted from the UHD 30 fps version. The second sequence, *SquareAndTimelapse* (Netflix), is a 10 seconds long live-action clip produced by Netflix, featuring dynamic scenes with a moving crowd and fast-paced motion typical of timelapse videos. To match the supported 16:9 aspect ratio of the selected platforms, the original horizontal resolution of 4096 pixels was cropped to 3840 pixels. The third, 30 seconds long sequence, featured jellyfish footage, showcasing slow-paced motion scenes with tiny particles moving in the water. These sequences were uploaded to each video platform using a web browser, except for Instagram, where the official Android application was required. Since the Instagram app operates as a "black box," there was no control or visibility over any potential local video processing before the upload. The "Upload at highest quality" option was enabled to ensure the best possible quality during the upload process. Once the platforms had encoded the sequences, yt-dlp was used to download the resulting videos. Most platforms provided H.264-encoded versions of the videos, with the exception of Instagram, which used the VP9 codec. A VP9-encoded 360p version was also downloaded from YouTube. Although these platforms support additional video encoding formats, they were not available for the videos uploaded for this study. All downloaded videos maintained the same resolution and frame rate as the original uploads (reference files), with one exception: Instagram's 360 p variant, where the horizontal resolution was altered from 640 pixels to 638 pixels. This discrepancy caused issues during the quality evaluation, which had to be addressed. Additionally, Instagram's videos were converted from the ITU-R BT.709 color space to ITU-R BT.601, a standard originally designed for standard-definition video in the early digital era. This unexpected

color space conversion further complicated the quality evaluation and required additional adjustments to ensure consistency across the tested videos.

3.2 Metrics Used for Objective VQA

Three objective quality metrics were selected for this study: VMAF, PSNR-HVS-M, and MS-SSIM. All of these metrics are full-reference, meaning they require both the distorted video and its corresponding reference video for comparison. The video pairs being compared must have the same resolution, frame rate, subsampling, and bit depth to ensure accurate measurement. However, VMAF provides methods to compare videos even when these parameters do not match, though this feature was not utilized in this paper. For consistency, only the luma channel (Y) of the YC_BC_R color space was used for the quality evaluation. The official VMAF implementation was used for calculating VMAF scores, which involved decoding both the downloaded and reference videos into the YUV4MPEG2 format to obtain uncompressed YCbCr video. For PSNR-HVS-M and MS-SSIM, Python libraries were employed to perform the calculations.

4. RESULTS AND DISCUSSION

4.1 Objective Video Quality

Figures 2, 3 and 4 show results of the selected metrics in all three sequences.

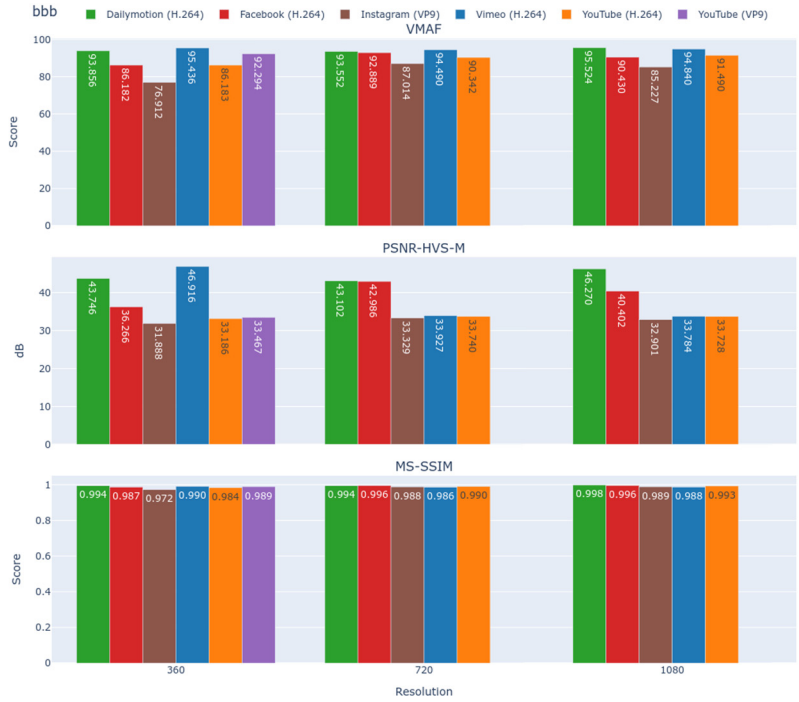

Figure 2: Big Buck Bunny sequence results

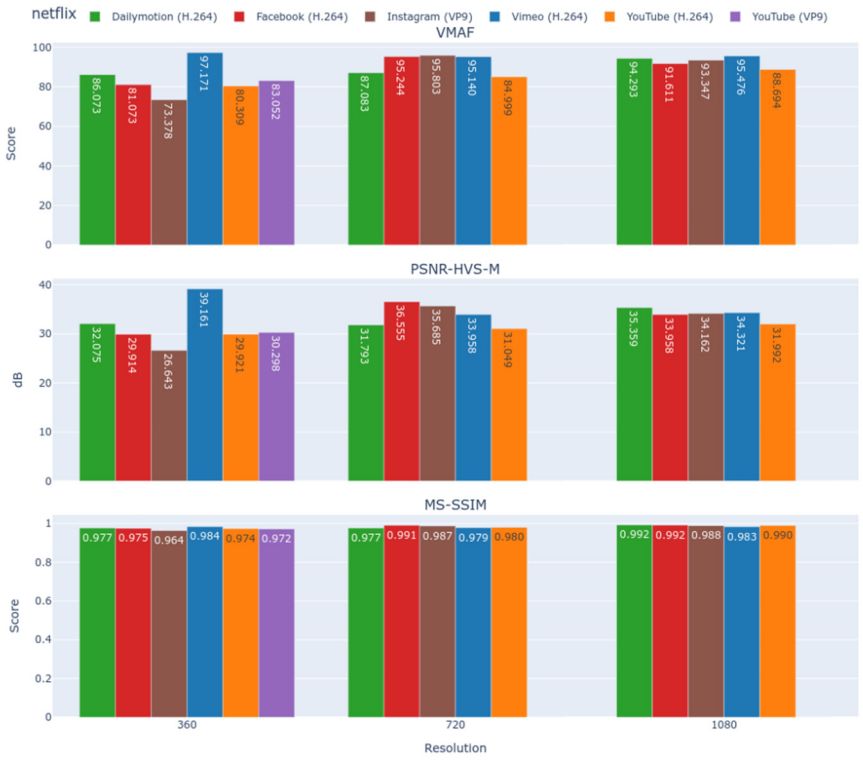

Figure 3: Netflix sequence results

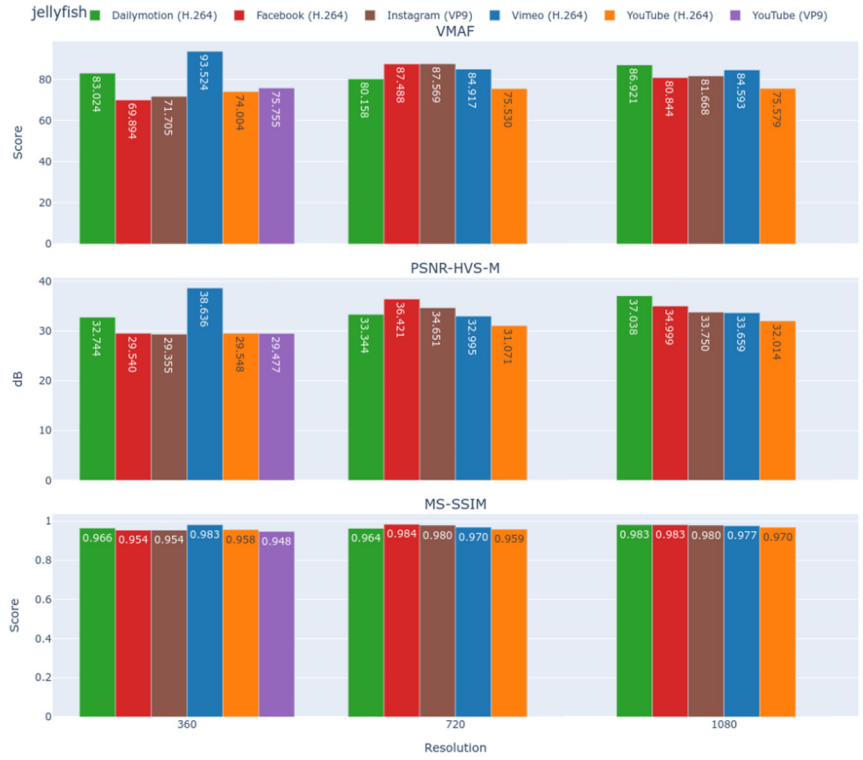

Figure 4: Jellyfish sequence results

The VMAF results indicate consistently high scores across all platforms and resolutions, with Vimeo and Dailymotion generally performing the best, except at 720 p, where Instagram surprisingly excels. In contrast, Instagram typically ranks last in other resolutions, and YouTube's H.264 performance is also

relatively low. PSNR-HVS-M shows greater variation between platforms. For *Big Buck Bunny*, Dailymotion achieves the highest scores, while Instagram performs the worst. In the *Jellyfish* and *Netflix* sequences, Vimeo leads at 360 p with scores of 38.636 dB and 39.161 dB, respectively. However, at higher resolutions, Vimeo's performance declines, with Dailymotion or Facebook often taking the lead. MS-SSIM results show minimal differences across platforms, with near-perfect scores for most resolutions. For *Big Buck Bunny*, the majority of platforms achieve scores above 0.980, although Instagram consistently scores the lowest. Overall, Vimeo and Dailymotion outperform other platforms in most metrics, while Instagram lags behind, particularly in VMAF and PSNR-HVS-M, and struggles at the 360p resolution. MS-SSIM results suggest that the perceptual visual differences between platforms are small. YouTube's VP9 codec generally performs as well or slightly better than the older H.264 variant.

4.2 Bitrate variance

Figures 5, 6, and 7 illustrate the bitrate variations for each resolution across all tested video platforms. For 360p resolution, video platforms generally use bitrates ranging from 0.5 to 1.5 Mbps, 2 to 3 Mbps for 720 p, and 4 to 6 Mbps for 1080 p. Vimeo exhibits the highest bitrate and largest variance at 360p, which likely explains its superior performance across all metrics. At higher resolutions, Instagram shows the most significant variance in bitrate. In contrast, Dailymotion maintains consistent bitrate usage across all resolutions and sequences. Notably, YouTube's VP9 codec achieves the same quality as its H.264 counterpart but at a lower bitrate. Also, both Facebook and Instagram show similar bitrate values and variance while using different video coding standards.

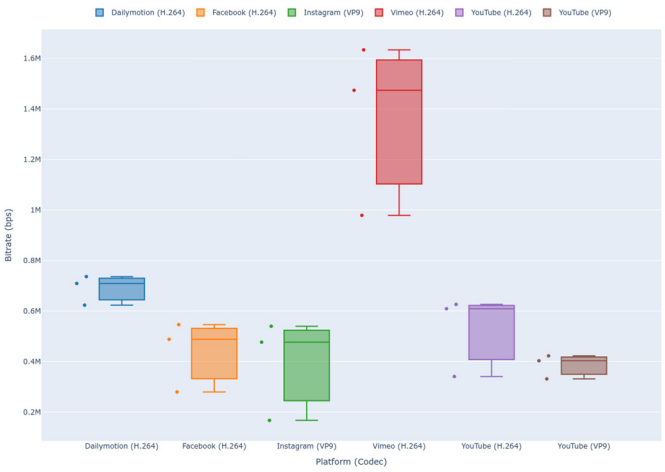

Figure 5: 360 p resolution bitrate variance

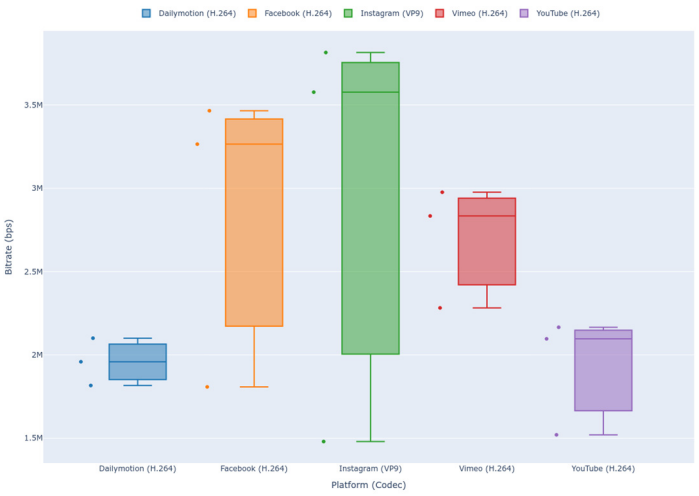

Figure 6: 720 p resolution bitrate variance

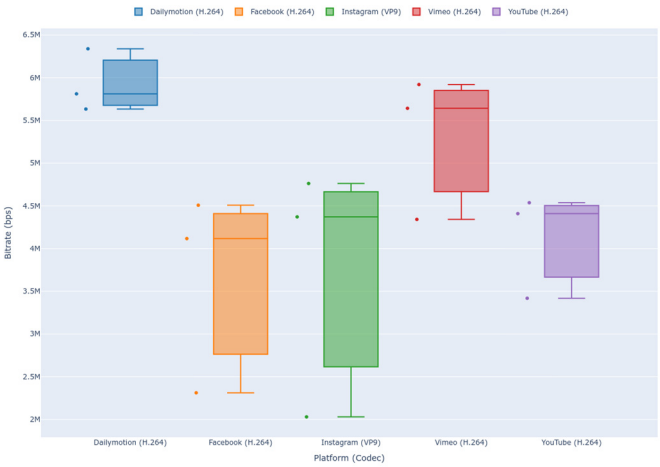

Figure 7: 1080 p resolution bitrate variance

3.3 Future Improvements in Objective VQA for User-Generated Content Platforms

The main goal of this paper is to establish a method for objectively evaluating the video quality of platforms that offer user-generated content and offer initial results. Based on the findings and empirical results, several areas for improvement in the evaluation process have been identified. First, both the variety and quantity of testing sequences can be significantly expanded. To facilitate this, automating the entire process would be highly beneficial. This would involve automating the preparation of reference sequences, video uploads and downloads, objective quality evaluation, and the presentation of results. Such automation would enhance efficiency and ensure consistency in the evaluation process. Further research is needed to explore the impact of clip length on quality evaluation. Currently, the evaluation is constrained by platform-specific limits, with the shortest maximum video length allowed across the selected platforms determining the clip duration. Understanding how the length of a clip affects both perceptual and objective video quality would provide deeper insights. Expanding the evaluation to include additional video platforms could also yield more comprehensive results. This could involve platforms that offer adult content, those requiring premium or affiliate accounts, and subscription-based services. Including such platforms would provide access to a broader range of resolutions and video standards, offering a more complete analysis of video quality across various formats. Partial comparisons could be conducted for resolutions not uniformly supported by all platforms. Additionally, expanding the evaluation to cover more advanced formats, such as HDR content, interlaced video, 360° videos, and stereoscopic (3D) content, would be possible within a subset of platforms. This would allow for a more detailed assessment of how platforms handle cutting-edge video formats, further refining the methodology for objective video quality evaluation.

5. CONCLUSIONS

It is not possible to definitively rank the video platforms from best to worst based on this dataset, but several general conclusions can be drawn. For users who prioritize high visual quality and advanced features, and are not concerned with higher bitrates, Vimeo stands out as a strong choice. YouTube, while not always delivering the highest visual quality, offers a wide range of features and is the most versatile platform. In contrast, Instagram and Facebook do not excel in either visual quality or advanced video features, as video-on-demand (VOD) services are only a small part of their broader social networking offerings. Dailymotion, though a smaller player in the market with limited features (such as the lack of HDR support), still delivers sufficient visual quality for standard use cases.

7. REFERENCES

Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F. & Carli, M. (2006) New full-reference quality metrics based on HVS. In: *Proceedings of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 22-24 January 2006, Scottsdale, Arizona, USA.*

Gonzalez, R.C. (2009) *Digital image processing*. Pearson education India.

Li, Z., Aaron A., Katsavounidis, I., Moorthy, A. & Manohara, M. (2016) *Toward A Practical Perceptual Video Quality Metric.* Available from: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652 [Accessed 14th September 2024].

Ponomarenko, N., Silvestri, F., Egiazarian, K., Carli, M., Astola, J.T. & Lukin, V.V. (2007) On between-coefficient contrast masking of DCT basis functions. In: *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 25-26 January 2007, Scottsdale, USA.*

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P. (2004) Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*. 13(4), pp. 600–612. Available from: doi: 10.1109/TIP.2003.819861

Wang, Z., Simoncelli, E.P. & Bovik A.C. (2003) Multiscale structural similarity for image quality assessment. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, *9-12 November 2003, Pacific Grove, California, USA.* pp. 1398–1402. Available from: doi: 10.1109/ACSSC.2003.1292216