QoE-based Performance Comparison of AVC, HEVC, and VP9 on Mobile Devices with Additional Influence Factors

Omer Nawaz *, Markus Fiedler, Siamak Khatibi

Posted Date: 30 December 2023

doi: 10.20944/preprints202312.2342.v1

Keywords: QoE metrics; Video quality assessments; HEVC and AVC comparison; mobile codecs efficiency; Multimedia Streaming; QoE IFs

Article

QoE-based Performance Comparison of AVC, HEVC, and VP9 on Mobile Devices with Additional Influence Factors

Omer Nawaz *, Markus Fiedler and Siamak Khatibi

Blekinge Tekniska Högskola, Karlshamn, Sweden; markus.fiedler@bth.se; siamak.khatibi@bth.se
* Correspondence: omer.nawaz@bth.se

Abstract: While current video quality assessment research predominantly revolves around resolutions of 4K and beyond, targeted at Ultra High-Definition (UHD) displays, effective video quality for mobile video streaming remains primarily within the range of 480p to 1080p. In this study, we conducted a comparative analysis of the Quality of Experience (QoE) for widely implemented video codecs on mobile devices, specifically Advanced Video Coding (AVC), its successor High-Efficiency Video Coding (HEVC), and Google's VP9. Our choice of 720p video sequences from a newly developed database, all with identical bitrates, aimed to maintain a manageable subjective assessment duration, capped at 35-40 minutes. To mimic real-time network conditions, we generated stimuli by streaming original video clips over a controlled emulated setup, subjecting them to eight different packet loss scenarios. We evaluated the quality and structural similarity of the distorted video clips using objective metrics, including Video Quality Metric (VQM), Peak Signal-to-Noise Ratio (PSNR), Video Multi-Method Assessment Fusion (VMAF), and Multi-Scale Structural Similarity Index (MS-SSIM). Subsequently, we collected subjective ratings through a custom mobile application developed for Android devices. Our findings revealed that VMAF accurately represents the degradation in video quality compared to other metrics. Moreover, in most cases, HEVC exhibits an advantage over both AVC and VP9 under low packet loss scenarios. However, it is noteworthy that in our test cases, AVC outperformed HEVC and VP9 in scenarios with high packet loss, based on both subjective and objective assessments. Our observations further indicate that user preferences for the presented content contribute to video quality ratings, emphasizing the importance of additional factors that influence the perceived video quality of end-users.

Keywords: QoE metrics; Video quality assessments; HEVC and AVC comparison; mobile codecs efficiency; Multimedia Streaming; QoE IFs

1. INTRODUCTION

The affordability of handheld mobile devices with internet availability has resulted in the tremendous growth of multimedia traffic. More brands are turning towards content creators for the promotion of their products and the quality of content with good end-user experience is the key to success due to severe competition among monetized offerings. The anticipated global monthly data usage is 19-GB in 2023 as per the latest mobile data traffic forecasts [1,2]. There is a shift towards the consumption of multimedia especially video streaming and gaming on mobile devices. The global mobile gaming market has already reached around 185 billion US dollars as of 2022 and is expected to grow threefold by the end of 2027 [3–5]. The ascent of multimedia and gaming on mobile phones is emblematic of a transformative shift in how individuals engage with digital content and entertainment. The convergence of powerful hardware, intuitive user interfaces, and a thriving app ecosystem has turned mobile devices into multifunctional entertainment hubs. This paradigm shift can be attributed to several key factors. Firstly, the increasingly sophisticated mobile hardware, equipped with high-resolution displays, robust processors, and advanced graphics capabilities, has created an ideal platform for delivering visually immersive multimedia and gaming experiences [6,7]. Secondly,

 doi:10.20944/preprints202312.2342.v1

the availability of high-speed internet connectivity, particularly the widespread deployment of 4G and the emergence of 5G networks, has ensured seamless streaming and online multiplayer gaming experiences [8,9]. Moreover, the app marketplaces, such as Apple's App Store and Google Play, have fostered an environment where developers can innovate and publish an extensive array of multimedia and gaming applications, meeting diverse user preferences. As a result, mobile phones have become a primary medium for entertainment consumption, reshaping the dynamics of the media and gaming industries [10,11]. Thus, the success of any social media application with video at its core relies on choosing the correct platform, codecs, and optimizations to adapt to end-to-end network quality parameters.

Taking into consideration the aforementioned factors, we conducted an analysis of the performance of the most widely employed video codecs across diverse hardware and software platforms, particularly in the context of video streaming and mobile gaming. Video codecs are fundamental in the compression and transmission of digital video content, and the comparison between the AVC (Advanced Video Coding, H.264), HEVC (High-Efficiency Video Coding, H.265), and VP9 codecs has attracted significant attention. While HEVC and VP9 may indeed offer superior coding efficiency with respect to compression, AVC maintains a dominant market share of nearly 80% in the mobile device sector [12]. This prevalence can be attributed to its lower processing overhead, resulting in reduced computational demands, a vital consideration for mobile devices. This paper presents a comprehensive codec comparison, evaluating both objective and subjective Quality of Experience (QoE) metrics to gauge user perception. Additionally, we analyzed the impact of human-influencing factors, such as user contentment with the displayed material, on video quality assessments. Beyond the conventional Mean Opinion Score (MOS), we incorporate the Good or Better (GoB) and Poor or Worse (PoW) metrics, which are highly recommended for delivering a more transparent and well-rounded assessment of user ratings [13,14]. Our findings affirm that, regardless of its age, AVC demonstrates superior performance atleast on mobile devices, owing to its optimization and lower computational demands. Moreover, our results underscore the significance of human-influencing factors and their potential to influence video quality ratings.

The paper is structured as follows: In Section 2, we offer an overview of the background and a concise exploration of relevant technologies. Section 3 provides comprehensive details of the experimental setup, including all parameters. Section 4 presents the assessment outcomes, accompanied by essential explanations. Lastly, Section 5 outlines the conclusions drawn from our study.

2. BACKGROUND

In this section, we have focused on the video codec standards, the QoE perspective for quality comparison, and the role of additional impact factors. A summary of previous work relevant to this paper is also provided.

Significant research has been conducted to evaluate the performance disparities between HEVC and AVC across various scenarios encompassing bitrate, quality optimization, and computational overhead. These investigations spanned from QoE assessments on high-definition displays to crowdsourced studies. However, a noticeable research gap exists regarding the impact of these codecs' performance specifically on mobile devices, utilizing a database obtained within an emulated network for future benchmarking. Song et al. [15] modeled the performance of AVC, HEVC, and VP9, observing that VP9 and HEVC exhibit a superior bitrate performance of 33% to 44% respectively across multiple resolutions compared to AVC. However, the database used in their study comprised solely of encoded videos generated locally, lacking consideration for computational overhead. Casas et al. [16] measured QoE provision on smartphones, employing lab-based assessments and user ratings garnered through crowdsourcing. Their evaluation encompassed platforms like YouTube, WhatsApp, and Facebook accessed via Chrome and Google Maps. Notably, this study did not delve into the analysis of underlying codecs or specialized applications for quality assessment. Some other research

 doi:10.20944/preprints202312.2342.v1

studies have focused on codec performance within vehicular networks and live video broadcasting on mobile devices, exploring platforms such as Periscope and Facebook Live. These studies emphasized the performance evaluation of communication protocols like Real-time Messaging Protocol (RTMP) and HTTP Live Streaming (HLS) and are not focused on underlying codec performance [17,18].

2.1. Video Codecs

Video coding standards are primarily evaluated based on compression performance alongside the ability to maintain video quality. The implementation of coding standards within an application is extremely important by maintaining the defined syntax of bit-stream and the decoding process, while encoders generate standard-compliant bitstream and thus determine compression performance. This is primarily the reason that old codec with a relatively higher degree of optimizations within an application over the years may outperform the new codec with better theoretical compression efficiency.

2.1.1. AVC and HEVC

H.264/MPEG-4-AVC [19] was launched back in 2004 and is still the widely used video coding standard over diverse platforms. According to Bitmovin's Video Developer Report, H.264 (AVC) is the world's most popular codec with an estimated 90% of video using it. Most modern devices with heterogeneous platform support AVC, and due to the low computational overhead as compared to its rivals, makes it a first choice for mobile-based applications [12]. AVC leverages both spatial and temporal redundancy reduction techniques to compress video data effectively. Spatial compression is facilitated through intra-frame coding, which encodes each frame independently. Temporal compression, on the other hand, is achieved through inter-frame prediction, where subsequent frames are predicted based on preceding frames, and only the differences (residuals) are encoded. Additionally, AVC employs predictive coding, which involves predicting pixel values based on neighboring pixels, thereby reducing the amount of information that needs to be transmitted. The codec also features advanced entropy coding techniques, such as context-adaptive binary arithmetic coding (CABAC), which adaptively encodes symbols based on the context, resulting in more efficient coding. AVC's versatility is evident in its support for a wide range of video resolutions and bit rates, rendering it suitable for diverse applications, from low-resolution video conferencing to high-definition video streaming [20,21].

HEVC, also known as H.265, represents a significant advancement in video compression standards. Developed through a collaborative effort between the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG), HEVC was introduced to address the growing demand for more efficient video compression while maintaining high-quality video [22]. HEVC introduces coding tree units (CTUs), enabling more flexible partitioning of coding units and improving compression efficiency. It employs quad-tree block partitioning, variable block sizes, and a wider range of prediction directions, enhancing its ability to capture intricate motion in video [23]. HEVC's advanced compression capabilities make it well-suited for ultra-high-definition (UHD) video content, as it can deliver the same quality video at approximately half the bit rate of its predecessor, H.264.

2.1.2. VP9

Google's VP9 is an open-source video codec that has emerged as a significant player in the domain of video compression, particularly in web-based video streaming and mobile gaming. Designed as a royalty-free alternative to established standards like H.264 and H.265, VP9 is part of Google's WebM project, which aims to provide efficient and high-quality video codecs for web applications. VP9 employs a variety of advanced techniques to achieve its compression goals. These include both intra-frame and inter-frame coding, enabling efficient compression by encoding individual frames independently and using inter-frame prediction to reduce redundancy. The codec supports a range of block sizes, allowing it to adapt to various types of video content. A notable advantage

 doi:10.20944/preprints202312.2342.v1

of VP9 is its capability to handle higher resolutions and bit depths, making it particularly suited for ultra-high-definition (UHD) content [24,25]. This feature, combined with its open-source nature, has made VP9 a popular choice for streaming and web video applications like YouTube, etc. However, the codec may face challenges related to hardware decoder support on different devices as Apple devices do not support VP9. Despite this, VP9 remains a significant contender in the landscape of video codecs, offering efficient compression while avoiding licensing costs, particularly advantageous for web-based video streaming.

2.2. QoE Perspective

The term QoE is defined by ITU-T as *'The degree of delight or annoyance of the user of an application or service.'* [26], with reference to the full definition that continues with *'It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state'* [27]. QoE offers a spectrum of objective metrics aimed at gauging user perception regarding video quality through various tools. However, the most dependable measure resides in direct user interaction, achieved by conducting subjective assessments.

2.3. Objective Metrics

Evaluations of video quality can be categorized as full-reference, reduced-reference, or no-reference, contingent upon the accessibility of original videos for reference. The complexities inherent in aggregating video databases and the limited availability of full-reference necessitate the development of diverse methodologies for gauging no-reference bitstream data such as progressive downloads, image quality assessments, and adaptive video streaming [28–30]. Our research focuses on mobile devices, prioritizing lifelike resolutions for video streaming. We've constructed a video stimuli database within an emulated network environment, guaranteeing the availability of extensive full-reference data for our evaluations. The objective metrics employed in evaluating QoE span various families, such as structural similarity metrics encompassing methodologies like the Multi-Scale Structural Similarity Index (MS-SSIM), among others. Additionally, within the realm of principle quality metrics, notable members include the Peak Signal-to-Noise Ratio (PSNR) and Video Multi-method Assessment Fusion (VMAF), etc [31]. These full-reference metrics primarily function to measure the disparity between the original frames and those received, constituting the video stream. Notably, different video codecs tend to yield distinct distortions, posing a challenge for these metrics. This study integrates MS-SSIM, VMAF, PSNR, and Video Quality Metric (VQM) methodologies to ascertain video quality, considering the nuanced implications of codec-induced distortions.

2.3.1. PSNR and MS-SSIM

SSIM and PSNR stand as the prevailing objective metrics for quantifying image and video quality owing to their computational simplicity and extensive historical benchmarking, among other factors. A multitude of scientific publications have evaluated the merits and demerits of these metrics. A comprehensive synthesis of these deliberations is available via the MSU Graphics and Media Lab Video Group link, drawing insights from an analysis encompassing 378 articles [32]. MS-SSIM is an enhancement to the traditional structural similarity index by using sub-sampling on multiple stages and involves structural distortion measurement instead of the error [33]. The MS-SSIM is defined as:

$$\text{MS-SSIM} = \frac{1}{N}\sum_{i=1}^{N}\left[\text{SSIM}_i(L)\cdot(\text{MS-SSIM}_i(C))^{\beta}\right] \tag{1}$$

 doi:10.20944/preprints202312.2342.v1

Where:

$$\begin{aligned} \text{MS-SSIM} &= \text{Multi-Scale Structural Similarity Index} \\ N &= \text{Number of scales} \\ i &= \text{Scale index, ranging from 1 to } N \\ \text{SSIM}_i(L) &= \text{SSIM at scale } i \text{ for the luminance (luma) component} \\ \text{MS-SSIM}_i(C) &= \text{Contrast component of MS-SSIM at scale } i \\ \beta &= \text{Weighting parameter} \end{aligned}$$

2.3.2. VQM and VMAF

VQM assesses the perceptual impact of video distortions, encompassing aspects such as blurring, irregular motion, global noise, block distortion, and color aberrations, combining these factors into a unified metric. Empirical testing outcomes demonstrate a strong alignment between VQM scores and subjective evaluations of video quality, leading to its adoption by ANSI as an benchmark for objective video quality assessment [31,34].

VMAF is a video quality metric developed by Netflix in collaboration with multiple research groups notably the Laboratory for Image and Video Engineering (LIVE) at The University of Texas. The metric measures information fidelity loss, loss of details, impairments, and temporal difference based on luminance [35]. It tends to outperform other metrics in both Netflix tests and other video quality tools benchmarks [36,37].

2.4. Subjective Metrics

The Mean Opinion Score (MOS) serves as the prevailing subjective metric employed to quantize user perception regarding a stimulus and has found extensive adoption within the industry for assessing speech and video quality [26]. Several studies delve into the efficacy of MOS, debating whether the disparity between ratings like 'Good' and 'Excellent' equates to the difference observed between 'Poor' and 'Bad.' Conversely, other research explores the relevance of MOS in accurately predicting user perception towards stimuli, highlighting its potential in assessing acceptability [38–40]. On the other hand, many factors can influence the user ratings like user background, the test environment, etc. The term Influence Factor (IF) is defined as 'Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user' [27].

In this paper, we have analyzed the impact of user liking (delight) towards the shown content on his video quality ratings which resulted in multiple sub-groups. So, apart from standard statistical computations, we have calculated one-way ANOVA to analyze the difference and statistical relevance of our results where there are more than three sub-categories. The one-way ANOVA can be easily calculated from mean-squared error and is commonly derived as shown below:

$$MSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{N - k}$$

where N be the total number of observations. - k be the number of groups or treatments. - n_i be the number of observations in the i-th group. We can compute the overall sample mean $\bar{X}$ and the between-group variability, also known as the mean squared treatment (MST), is given by:

$$MST = \frac{\sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

The test statistic for one-way ANOVA is the ratio of between-group variability to within-group variability, known as the F-statistic:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

$$F = \frac{MST}{MSE}$$

Under the null hypothesis, the F-statistic follows an F-distribution with $(k-1, N-k)$ degrees of freedom. We can compare the calculated F-statistic to the critical value from the F-distribution table to make a decision. If $F > F_{\alpha}$, we reject the null hypothesis in favor of the alternative.

3. EXPERIMENTAL SETUP

This section describes the reasons for the choice of video stimuli for this paper, the details about the emulated network, induced transmission impairments, and the methodology used for conducting the subjective assessment.

3.1. Video Selection

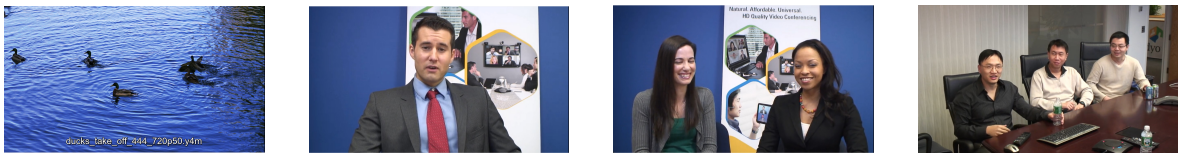
Drawing from our prior experiences and ongoing discussions pertaining to prolonged subjective assessments involving numerous video stimuli lasting a few seconds each [41,42], it was observed that the users tend to get bored and start losing focus which may affect their ratings. Consequently, we created a database by selecting four original video clips from the xiph.org test suite, resulting in a database of 112 video stimuli after distortions to maintain a reasonable duration for the subjective assessment process. The technical specifications of these videos are available in the Table 1.

Table 1. Reference Videos Specifications

Name	Length seconds	fps	Resolution
Ducks	10	50	1280×720
Johnny	10	60	1280×720
KristenAndSara	10	60	1280×720
Vidyo1	10	60	1280×720

The source video sequences are in color-sensitive raw format i.e., YUV4Mpeg (.y4m). All four videos belong to YUV 4:2:0 color space with three different temporal and spatial characteristics as per ITU-T P.910 recommendations [43]. All four videos are different from each other in terms of spatial characteristics. The spatial and temporal details of these videos can be found in these articles [44–46]. AVC and HEVC videos were encoded using libx264 and libx265 video coding libraries respectively with Matroska [47] as the container. Whereas the VP9 videos are encoded using the libvpx-vp9 library and WEBM [24] as the container. Although the most common resolution for video streaming over mobile devices is still hovering around 480p, we have chosen the 720p resolution which is the most common resolution for video streaming on mobile devices for applications like YouTube over high-speed 4G/5G networks or Wifi [48,49].

The sample frame from the original source sequences are shown in the Figure 1.



(a) Ducks **(b)** Johnny **(c)** KristenAndSara **(d)** Vidyo1

Figure 1. Frames from reference videos

As the focus of the research was on additional influence factors and to remain focused on the benchmarking of objective and subjective metrics so apart from the Ducks video, the remaining three videos fall in the low temporal index. The Ducks video has medium spatial and temporal effects while Johnny, KristenAndSara, and Vidyo1 belong to different categories in terms of spatial index.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

3.2. Emulated Testbed and Network Impairments

The experimental setup used for collecting the video stimuli can be categorized into four groups [40]:

1. to compress the test media using different codecs, resolutions, bps, and frame rates into the local machine. The benchmarking of the resultant video and compression efficiency is used to determine the quality of the video codec
2. to use the real-time network and streaming videos using different parameters mentioned above and collecting the stimuli on the end-devices
3. to use simulation software to stream stimuli with different codec settings on a depicted network with varying transmission impairments
4. to use an emulated testbed for real network experience with an opportunity for repeatable results

We have adopted the emulated network approach to replicate the real network conditions and manage issues like priorities associated with video packets in a repeatable environment. The setup used for establishing the emulated network is shown in the Figure 2.

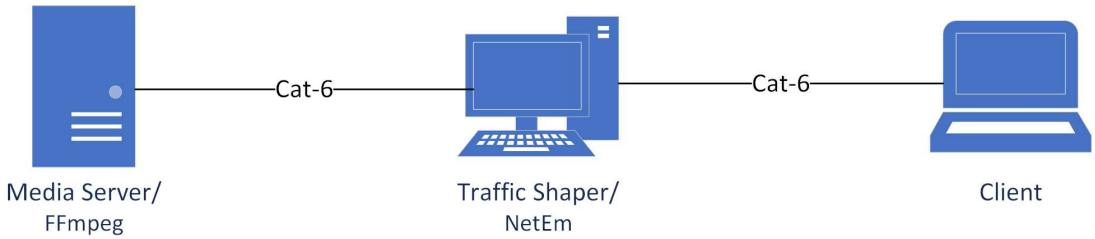


Figure 2. Video Streaming Setup: Emulated Approach

The emulated testbed consists of a streaming media server, a network emulator, and a client. The test network is designed to identify the effect of packet loss on the quality of videos encoded in H.264, H.265, and VP9. The apparatus is designed to stream HD videos from the FFmpeg [50] streaming server to a client encoded with different video codecs using RTSP protocol through a network emulator (NetEm) [51], where different network impairments have been imposed on the passing traffic and the distorted video stream is captured at the client. FFmpeg supports all the latest codecs including H.264, H.265, and VP9. It also supports all types of UDP and TCP streaming protocols. NetEm is responsible for routing between server and client using multiple network interfaces and hence, it acts as an interface between the media server and streaming client. The details of the used hardware/software can be found in Table 2.

Table 2. Hardware Specifications for the emulated testbed

Specifications	Streaming Server	Traffic Shaper	Client
Platform	HP ProLiant DL120	Dell OptiPlex 9020	Dell OptiPlex 3050 (AIO Series)
Processor	Intel Xeon E5-1620	Intel Core i7 3.6 GHz-Quad Core	Intel Core i7 3.4 GHz-Quad Core
RAM	16GB DDR4	8GB DDR3	8GB DDR4
OS	Windows 10Pro	Fedora 31(Server Edition)	Windows 10Pro
Storage	512GB SSD	500 GB HDD 7200 RPMs	256GB SSD
Software	FFMPEG 4.3.1	NetEm Kernel Version 5.5.8	FFMPEG 4.3.1

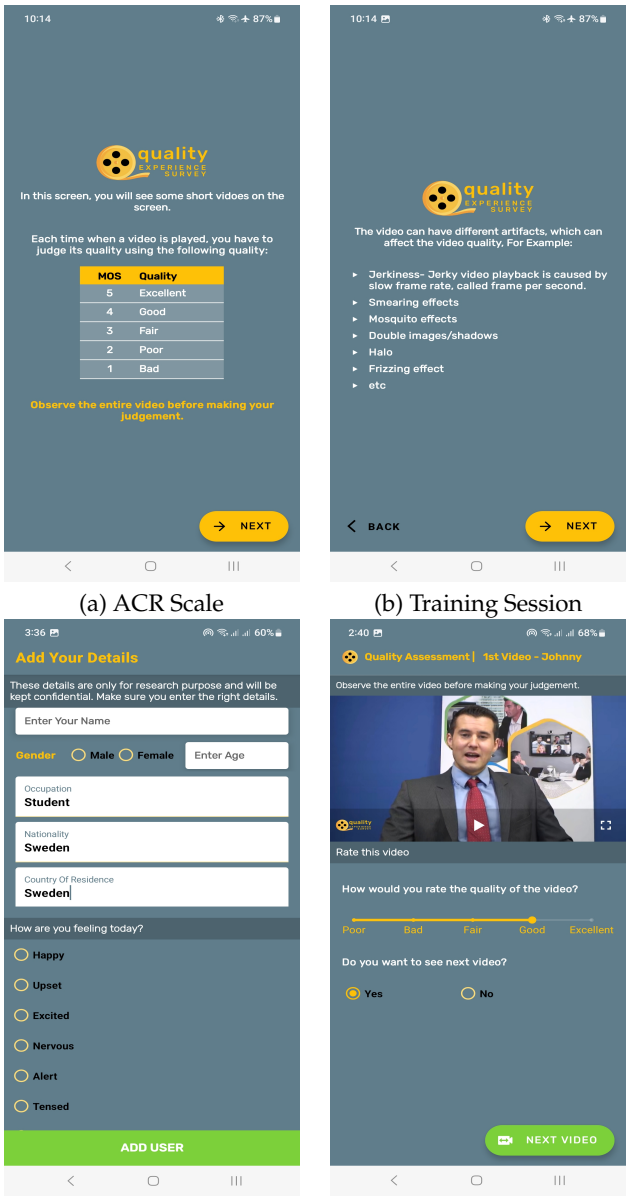
We have chosen nine packet loss scenarios ranging from 0% to 10% for streaming videos on the emulated setup. There was no delay or jitter associated with the video stimuli used in this experiment. The main reason for this choice was the length of the subjective assessment due to the benchmarking of three codecs. In order to obtain the objective metrics mentioned in Section 2.3. The objective metrics are obtained using the Video Quality Measurement Tool (VQMT) developed by Moscow State University.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

We have used the 13.1 free version which has some limitations as compared to the Pro version but provides correct results, unlike the Demo version [52].

3.3. Subjective Assessment

As we were analyzing the codecs with both objective and subjective metrics along with the additional influence factor of delight on a mobile device, we developed an Android application for conducting the subjective assessment and Google Firebase [53] is used for the collection of the results. The assessments were conducted on Samsung Note 10 Lite (SM-N770F/DS) mobile with 8GB of RAM. Due to the nature of the experiment, it was not possible to conduct the assessment in the perceptual lab but the method of conduction of individual assessment and rating scales were used as per the ITU-T P.910 [43] and ITU-R BT.500-14 [54] recommendations. A training session was conducted before every assessment, and users were provided with both verbal and written instructions as shown in the Figure 3.



(a) ACR Scale (b) Training Session

(c) User Info (d) MOS

Figure 3. Mobile App for Subjective Assessment

 doi:10.20944/preprints202312.2342.v1

In the next step, test media without distortions was shown to the user, and they were asked whether they liked it or not based on a binary scale of 'Yes' and 'No'. The users were also requested to provide their ranking of the content based on their delight towards the shown content on the 1–9 point scale. These ratings were scaled down to five-point scale with a step size of 0.5 using the formula $5 - (9 - R)/2$ where R stands for delight rating on the nine-point scale. This conversion helped us in comparing the effect of content delight with corresponding MOS. This categorization also helped us to increase number of users in each sub-category. The basic user data along with influence factors is taken from the users as shown in the Figure 3. The stimuli encoded with different codecs and packet loss ratios but belonging to the same group i.e., Johnny, etc. were shown randomly to the subjects. The users were provided the choice to take a break after watching the first two sets of videos and the majority of users took a 3–5 minute break. The user ratings for video quality were obtained using the Single-Stimulus method on a 5-point Absolute Category Rating (ACR) scale.

The selection of test media for this paper meant a total of 112 videos, including the originals resulting in a total assessment time of around 35–40 minutes. A total of 51 participants rated the test media resulting in the compilation of 5712 assessments. During the post-evaluation, three outliers were detected and subsequently excluded from the dataset. Out of the remaining 48 subjects, 20 males and 28 females participated with a mean age of 22.52 and mode 21. The majority of the participants were Bachelor students of information technology and were aware of the issues related to the quality of multimedia streaming.

4. RESULTS AND DISCUSSION

The foremost thing to mention is the encoding delay experienced by the streaming server. As the processing of videos is done in batches, we have noticed that AVC has the least computing overhead as compared to HEVC and VP9. In our experience, the encoding delay of HEVC was 3-4 times higher than the AVC. The time taken was VP9 was slightly better than the HEVC but still far greater than the AVC. This may be one of the foremost reasons for the resilience of AVC for low computing devices nowadays. The dataset of the streamed videos was analyzed before the subjective assessment. It was noted that due to the high temporal and spatial effects of the Ducks video, the results were not consistent with the other three videos having somewhat similar characteristics. It is important to mention that some videos like Duck_AVC at 0.7%, Johnny_VP9 at 0.5%, KristenAndSara_AVC at 0.7%, and Vidyo1_HEVC at 0.1% have low quality as compared to the proceeding high loss video. This shows the effect of high-priority packet loss associated with different types of encoding frames.

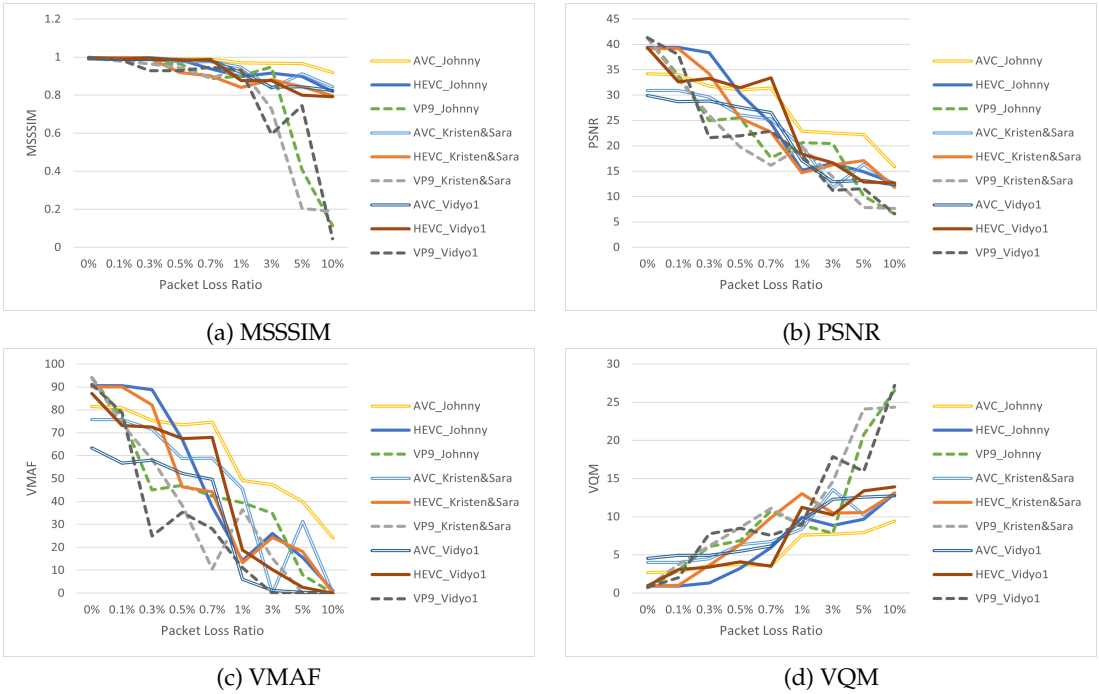
4.1. Objective Assessment

The results of objective metrics are shown in the Figure 4. AVC has outperformed VP9 and is slightly better or on par with HEVC for most of the cases. It is important to mention that the VMAF metric has the most reliable estimation of video quality as compared to other metrics. Based on our results, the MSSSIM is the most unreliable metric as it failed to quantize the video degradation for different packet loss scenarios and the results are mostly flat. In the case of KristenAndSara_AVC at 0.7% and Vidyo1_HEVC at 0.1%, we can observe that VMAF can correctly depict the change as compared to both PSNR and VQM. We can easily say that within the context of the VQMT implementation used in this paper, VMAF emerges as the optimal metric for quantifying video quality in multimedia streaming. These findings are in harmony with related research [36,55].

4.2. Subjective Ratings

The comparison of the three codecs for packet loss concerning MOS is shown in the Figure 5. There are three major observations from the results:

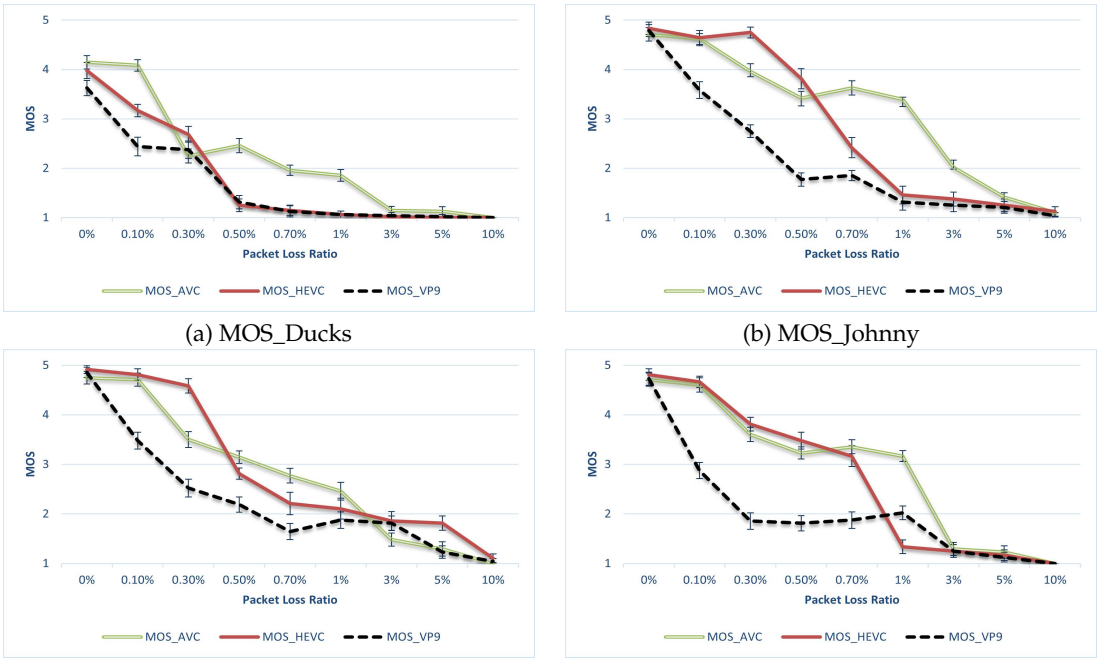
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1



(a) MSSSIM (b) PSNR

(c) VMAF (d) VQM

Figure 4. Ratings of objective metrics



(a) MOS_Ducks (b) MOS_Johnny

(c) MOS_KristenAndSara (d) MOS_Vidyo1

Figure 5. MOS of subjective assessment

1. The AVC overall outperforms both HEVC and VP9 in terms of human ratings.
2. The HEVC outperforms AVC or is on par in low packet loss scenarios, but its performance deteriorates when the packet loss reaches around 0.7%.
3. The performance VP9 tend to perform better in higher packet loss environment but still can't match the ITU-T codecs.

One key observation is that the MOS ratings reflect the actual degradation of the video quality for low packet loss ratios as compared to most objective metrics in the previous section. If the amount

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

of overhead associated with HEVC and VP9 is considered as described above in Section 4, we can understand that the performance gains for HEVC are not significant and that is the reason for the popularity and resilience of AVC as of 2023. Moreover, it also shows that for mobile devices with the requirements of computational and power efficiency, the AVC will remain the most popular codec in the near future. The values of GoB and PoW metrics are shown in the Table 3.

Table 3. GoB and PoW metrics

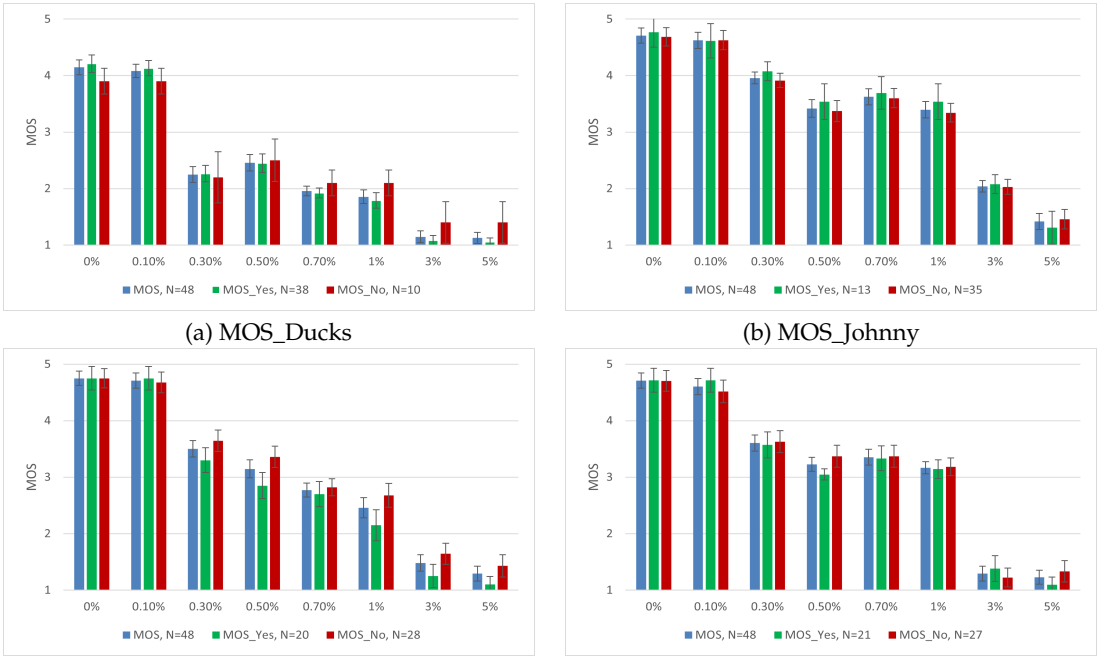
		AVC		HEVC		VP9	
Stimuli	PL%	GoB%	PoW%	GoB%	PoW%	GoB%	PoW%
Johnny	0%	100%	0%	100%	0%	100%	0%
	0.1%	100%	0%	100%	0%	62.5%	4.2%
	0.3%	91.7%	0%	100%	0%	0%	25%
	0.5%	43.8%	2.1%	64.6%	0%	0%	97.9%
	0.7%	62.5%	0%	2.1%	50%	0%	100%
	1%	39.6%	0%	0%	93.7%	0%	95.8%
	3%	0%	91.7%	0%	100%	0%	100%
	5%	0%	100%	0%	100%	0%	100%
	10%	0%	100%	0%	100%	0%	100%
KristenAndSara	0%	100%	0%	100%	0%	100%	0%
	0.1%	100%	0%	100%	0%	52.1%	4.2%
	0.3%	50%	0%	100%	0%	2.1%	52.1%
	0.5%	22.9%	8.3%	0%	18.8%	0%	75%
	0.7%	0%	22.9%	0%	66.7%	0%	95.8%
	1%	0%	47.9%	0%	68.7%	0%	89.6%
	3%	0%	100%	0%	77.1%	0%	95.8%
	5%	0%	100%	0%	95.8%	0%	100%
	10%	0%	100%	0%	100%	0%	100%
Vidyo1	0%	100%	0%	100%	0%	100%	0%
	0.1%	100%	0%	100%	0%	10.4%	22.9%
	0.3%	60.4%	0%	81.3%	0%	0%	89.6%
	0.5%	22.9%	0%	52.1%	4.2%	0%	93.8%
	0.7%	35.4%	0%	35.4%	18.8%	0%	87.5%
	1%	16.7%	0%	0%	100%	0%	89.6%
	3%	0%	100%	0%	100%	0%	100%
	5%	0%	100%	0%	100%	0%	100%
	10%	0%	100%	0%	100%	0%	100%

It is evident that the GoB percentage remains satisfactory up to a 0.5% packet loss. Additionally, the figures reveal a significant degradation in video quality when using the VP9 implementation in our testbed, even at low packet loss ratios. These metrics offer a more meaningful and easily comprehensible insight compared to MOS values in our evaluation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

4.3. Impact of delight of shown video content

Apart from the video quality, numerous human and system influence factors may affect user ratings [56,57]. We have taken the input from the users towards their delight for the shown content on a binary and ordinal scale as described in Section 2. The results of the MOS values based on user delight with error bars are shown in the Figure 6.



(a) MOS_Ducks (b) MOS_Johnny

(c) MOS_KristenAndSara (d) MOS_Vidyo1

Figure 6. MOS of subjective assessment based on Delight for AVC

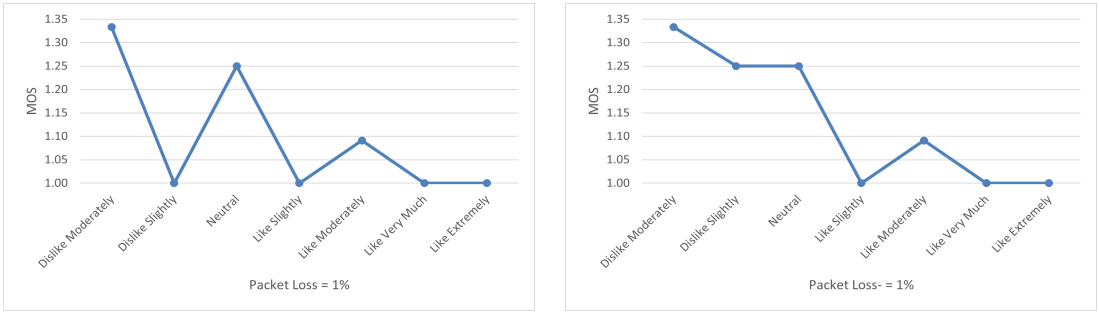
The effect of Delight on the shown content and its impact on MOS is evident. Although there are only two groups of results for the binary scale, we have calculated one-way ANOVA to observe the statistical relevance of the results. In the case of the *Ducks* video for AVC, apart from packet loss ratios of 0.3 and 0.5%, the remaining ratings are either statistically relevant or very close to being relevant, i.e., α=0.05. It is very important to mention that for *Johnny* video where the majority of subjects didn't like the content, the impact is negligible for most of the scenarios. In the case of *KristenAndSara* and *Vidyo1* with almost the same number of delights, the impact is statistically significant for some medium and high packet loss ratios. These results are in line with our previous work where we have observed that people showing delight towards a particular content are more critical if the video quality is degraded to a level where their viewing experience is disturbed [58]. To investigate this behavior, we have also used the user ratings towards the shown content on a 1—9 ordinal scale. The results for the Ducks video for HEVC and VP9 are shown in the Table4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 December 2023 doi:10.20944/preprints202312.2342.v1

Table 4. One way ANOVA for *Ducks* Video

Codec	Packet Loss Ratio	Delight Ratings_Sig	Delight_Sig
HEVC	0.5%	0.077	0.041
	1%	0.227	0.044
	3%	0.011	0.050
VP9	0.3%	0.20	0.056
	0.5%	0.360	0.028
	0.7%	0.018	0.011
	1%	0.273	0.005
	3%	0.430	0.004
	5%	0.011	0.050

The issue here is a smaller number of subjects in each group due to nine subgroups i.e., *Dislike Extremely, Dislike Very Much, Dislike Moderately (n=3), Dislike Slightly (n=4), Neutral (n=4), Like Slightly (n=5), Like Moderately (n=11), Like Very Much (n=10), Like Extremely (n=11)* where *n* shows users within the each group. Due to such a small sample size, we can observe the higher values of alpha for most of the scenarios which are significant on the binary scale. So, we have plotted the mean values for user ratings of one scenario where we have a higher significance level as compared to the binary scale, i.e., packet loss at 1%. The results are shown in the Figure 7.



(a) Ducks_HEVC (b) Ducks_VP9

Figure 7. Mean plot of delight for Ducks video

It can be observed that the users who have shown extreme delight towards the content are more critical when the quality degrades below a certain level. We have observed this trend in most of the remaining results. This shows that delight towards the content can influence user ratings and people are more sensitive to the video quality of the content which they like extremely.

5. CONCLUSION

In this study, we conducted performance benchmarking of the AVC, HEVC, and VP9 codecs on a mobile device. This evaluation encompassed both objective metrics and subjective assessments, considering various packet loss ratios within an emulated network environment. Our analysis focused on the prevailing video streaming resolution for high-speed networks on handheld devices. Our findings revealed that videos streamed using the AVC codec exhibited superior quality and greater resilience to transmission impairments. While HEVC occasionally matched or surpassed AVC performance under low packet loss conditions, its substantial computational overhead and subsequent power consumption offset this advantage. Conversely, VP9 consistently underperformed in comparison to other codecs across the spectrum of test scenarios. Additionally, we observed limitations in the accuracy of objective metrics, with VMAF emerging as the most reliable metric in our study.

In line with existing research, we have found that user delight plays a significant role in the user ratings of the shown video. Statistical significance was evident in most cases, and users exhibited

heightened scrutiny of video quality degradation once they had a strong affinity for the content. This underscores the significance of human-related factors and the necessity for collecting user input beyond conventional MOS ratings. Ultimately, our work highlights the importance of incorporating these user-related influences into future models to enhance the accuracy of user perception predictions.

Acknowledgments: This paper is part of the work carried out at the Video Quality Assessment Group at PUCIT headed by the 1^{st} author of this paper. The authors would especially like to thank Usman Aamer and Aqsa Aslam for their contributions during the experimental setup and mobile application development. The authors would also like to acknowledge the contribution of Muhammad Armaghan Rashid (Aitchison College) during the subjective assessments.

Abbreviations

Abbreviations

The following abbreviations are used in this manuscript:

QoE	Quality of Experience
AVC	Advanced Video Coding
HEVC	High-Efficiency Video Coding
VQM	Video Quality Metric
PSNR	Peak Signal-to-Noise Ratio
VMAF	Video Multi-Method Assessment Fusion
MS-SSIM	Multi-Scale Structural Similarity Index
MOS	Mean Opinion Score
GoB	Good or Better
PoW	Poor or Worse
VQMT	Video Quality Measurement Tool
ACR	Absolute Category Rating

References

1. Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, accessed on 2023-02-01.
2. Ericsson Mobility Report November 2022. https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2022, accessed on 2023-01-21.
3. Precedence Research: Mobile Gaming Market Size. https://www.precedenceresearch.com/mobile-gaming-market, accessed on 2023-09-16.
4. Most used devices for digital videos in the U.S. 2023. https://www.statista.com/forecasts/997109/most-used-devices-for-digital-videos-in-the-us, accessed on 2023-09-13.
5. World Telecommunication/ICT Indicators Database. https://www.itu.int:443/en/ITU-D/Statistics/Pages/publications/wtid.aspx, accessed on 2023-09-13.
6. Sattarov, A.; Khaitova, N. MOBILE LEARNING AS NEW FORMS AND METHODS OF INCREASING THE EFFECTIVENESS OF EDUCATION. *European Journal of Research and Reflection in Educational Sciences* **2019**, *7*.
7. Zerman, E.; Kulkarni, R.; Smolic, A. User Behaviour Analysis of Volumetric Video in Augmented Reality. 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), 2021, pp. 129–132. ISSN: 2472-7814, doi:10.1109/QoMEX51781.2021.9465456.
8. Shangguan, Z. The impacts of 5G technology and cloud computing on the gaming accessories industry. International Conference on Electronic Information Engineering and Computer Technology (EIECT 2021). SPIE, 2021, Vol. 12087, pp. 272–275. doi:10.1117/12.2624838.
9. Huang, H.S.; Su, Y.S. A Practical Study of QoE on Cloud Gaming in 5G Networks. 2023 International Wireless Communications and Mobile Computing (IWCMC), 2023, pp. 638–643. ISSN: 2376-6506, doi:10.1109/IWCMC58020.2023.10182439.

10. Twenge, J.M.; Martin, G.N.; Spitzberg, B.H. Trends in U.S. Adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print. *Psychology of Popular Media Culture* **2019**, *8*, 329–345. Place: US Publisher: Educational Publishing Foundation, doi:10.1037/ppm0000203.
11. Newzoo Global Games Market Report 2022 | Free Version. https://newzoo.com/resources/trend-reports/newzoo-global-games-market-report-2022-free-version, accessed on 2023-09-12.
12. Inc, B. Bitmovin's 4th Annual Video Developer Report 2020. https://go.bitmovin.com/video-developer-report-2020, accessed on 2023-01-29.
13. Hoßfeld, T.; Heegaard, P.E.; Varela, M.; Möller, S. QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Qual User Exp* **2016**, *1*, 1–23. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer International Publishing, doi:10.1007/s41233-016-0002-1.
14. Hoßfeld, T.; Heegaard, P.E.; Skorin-Kapov, L.; Varela, M. Deriving QoE in systems: from fundamental relationships to a QoE-based Service-level Quality Index. *Qual User Exp* **2020**, *5*, 1–17. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer International Publishing, doi:10.1007/s41233-020-00035-0.
15. Song, W.; Xiao, Y.; Tjondronegoro, D.; Liotta, A. QoE Modelling for VP9 and H.265 Videos on Mobile Devices. Proceedings of the 23rd ACM international conference on Multimedia; Association for Computing Machinery: New York, NY, USA, 2015; MM '15, pp. 501–510. doi:10.1145/2733373.2806256.
16. Casas, P.; Seufert, M.; Wamser, F.; Gardlo, B.; Sackl, A.; Schatz, R. Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices. *IEEE Transactions on Network and Service Management* **2016**, *13*, 181–196. Conference Name: IEEE Transactions on Network and Service Management, doi:10.1109/TNSM.2016.2537645.
17. Siekkinen, M.; kämäräinen, T.; Favario, L.; Masala, E. Can You See What I See? Quality-of-Experience Measurements of Mobile Live Video Broadcasting. *ACM Trans. Multimedia Comput. Commun. Appl.* **2018**, *14*, 34:1–34:23. doi:10.1145/3165279.
18. Iza Paredes, C.; Mezher, A.M.; Aguilar Igartua, M. Performance Comparison of H.265/HEVC, H.264/AVC and VP9 Encoders in Video Dissemination over VANETs. Smart Objects and Technologies for Social Good; Gaggi, O.; Manzoni, P.; Palazzi, C.; Bujari, A.; Marquez-Barja, J.M., Eds.; Springer International Publishing: Cham, 2017; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 51–60. doi:10.1007/978-3-319-61949-1_6.
19. ITU-T Rec: Advanced video coding for generic audiovisual services. Recommendation H.264 (08/2021), ITU-T.
20. Wiegand, T.; Sullivan, G.J.; Bjontegaard, G.; Luthra, A. Overview of the H.264/AVC video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on* **2003**, *13*, 560–576.
21. Iain E, R. *The H.264 Advanced Video Compression Standard*, 2nd ed.; Wiley, 2010.
22. ITU-T Rec: H.265 : High efficiency video coding. Recommendation H.265 (08/2021), ITU-T.
23. Sullivan, G.J.; Ohm, J.R.; Han, W.J.; Wiegand, T. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology* **2012**, *22*, 1649–1668. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology, doi:10.1109/TCSVT.2012.2221191.
24. The WebM Project | VP9 Video Codec Summary. https://www.webmproject.org/vp9/, accessed on 2023-01-28.
25. Ozer, J. VP9 Codec: Google's Open-Source Technology Explained. https://www.wowza.com/blog/vp9-codec-googles-open-source-technology-explained, accessed on 2023-10-18.
26. ITU-T Rec. P.10/G.100. Vocabulary for performance, quality of service and quality of experience. Recommendation (11/2017), ITU-T, 2017.
27. Qualinet. Qualinet White Paper on Definitions of Quality of Experience. Technical report, 2013. Library Catalog: www.qualinet.eu.
28. Heikkilä, G.; Gustafsson, J. Video QoE: leveraging standards to meet rising expectations. *ERICSSON TECHNOLOGY REVIEW ARTICLES* **2017**.
29. ITU-T P.1203 : Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport. Technical report.

30. Min, X.; Gu, K.; Zhai, G.; Liu, J.; Yang, X.; Chen, C.W. Blind Quality Assessment Based on Pseudo-Reference Image. *IEEE Transactions on Multimedia* **2018**, *20*, 2049–2062. Conference Name: IEEE Transactions on Multimedia, doi:10.1109/TMM.2017.2788206.
31. Wang, Y. Survey of Objective Video Quality Measurements. Technical Report MA 01748, EMC Corporation Hopkinton, USA.
32. PSNR and SSIM: application areas and criticism. https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html, accessed on 2023-09-14.
33. Wang, Z.; Simoncelli, E.; Bovik, A. Multiscale structural similarity for image quality assessment. The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, 2003, Vol. 2, pp. 1398–1402 Vol.2. doi:10.1109/ACSSC.2003.1292216.
34. Pinson, M.; Wolf, S. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting* **2004**, *50*, 312–322. Conference Name: IEEE Transactions on Broadcasting, doi:10.1109/TBC.2004.834028.
35. Netflix/VMAF · Features. https://github.com/Netflix/vmaf/blob/master/resource/doc/features.md, accessed on 2023-10-12.
36. Blog, N.T. Toward A Practical Perceptual Video Quality Metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, accessed on 2023-01-29.
37. Video Quality Metrics Benchmark Methodology. https://videoprocessing.ai/benchmarks/video-quality-metrics_frm.html, accessed on 2023-10-12.
38. Pinson, M.H.; Janowski, L.; Pepion, R.; Huynh-Thu, Q.; Schmidmer, C.; Corriveau, P.; Younkin, A.; Callet, P.L.; Barkowsky, M.; Ingram, W. The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing* **2012**, *6*, 640–651. doi:10.1109/JSTSP.2012.2215306.
39. Streijl, R.C.; Winkler, S.; Hands, D.S. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Syst.* **2016**, *22*, 213–227. doi:10.1007/s00530-014-0446-1.
40. Nawaz, O. A Holistic View of QoE for Multimedia Streaming. Licentiate Dissertation, 2023. Publisher: Blekinge Tekniska Högskola.
41. Fröhlich, P.; Egger, S.; Schatz, R.; Mühlegger, M.; Masuch, K.; Gardlo, B. QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment? 2012 Fourth International Workshop on Quality of Multimedia Experience, 2012, pp. 242–247. doi:10.1109/QoMEX.2012.6263851.
42. Nawaz, O.; Minhas, T.N.; Fiedler, M. QoE based comparison of H.264/AVC and WebM/VP8 in an error-prone wireless network. Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium, Lisbon, 2017, pp. 1005–1010.
43. ITU-T Rec. P.910. Subjective video quality assessment methods for multimedia applications, 2021.
44. Bender, I.; Borges, A.; Agostini, L.; Zatt, B.; Correa, G.; Porto, M. Complexity and compression efficiency analysis of libaom AV1 video codec. *Journal of Real-Time Image Processing* **2023**, *20*, 1–14. Company: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer Berlin Heidelberg, doi:10.1007/s11554-023-01308-5.
45. Mercat, A.; Arrestier, F.; Pelcat, M.; Hamidouche, W.; Menard, D. Probabilistic Approach Versus Machine Learning for One-Shot Quad-Tree Prediction in an Intra HEVC Encoder. *Journal of Signal Processing Systems* **2019**, *91*, 1021–1037. Company: Springer Institution: Springer Label: Springer Number: 9 Publisher: Springer US, doi:10.1007/s11265-018-1426-z.
46. Bender, I.; Palomino, D.; Agostini, L.; Correa, G.; Porto, M. Compression Efficiency and Computational Cost Comparison between AV1 and HEVC Encoders. 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5. ISSN: 2076-1465, doi:10.23919/EUSIPCO.2019.8903006.
47. Matroska Media Container Homepage. https://www.matroska.org/index.html, accessed on 2023-02-01.
48. Hamer, A. The ultimate guide to YouTube video sizes | Descript. https://www.descript.com/blog/article/the-ultimate-guide-to-youtube-video-sizes, accessed on 2023-06-21.
49. Ankita. The Best Resolution For YouTube: A Complete Guide | OFFEO. https://offeo.com/learn/best-resolution-for-youtube, accessed on 2023-06-21.
50. FFmpeg (cross-platform solution to record, convert and stream audio and video). https://ffmpeg.org/, accessed on 2023-07-02.
51. NetEm - Network Emulator | The Linux Foundation.

52. MSU Video Quality Measurement Tool (VMAF, PSNR, VQM, SSIM, NIQE, etc). http://compression.ru/video/quality_measure/video_measurement_tool.html, accessed on 2023-07-03.
53. Google Firebase. https://firebase.google.com/, accessed on 2023-07-02.
54. ITU-R Rec. BT.500-14. Methodology for the subjective assessment of the quality of television pictures, 2019.
55. Bampis, C.G.; Bovik, A.C. Learning to Predict Streaming Video QoE: Distortions, Rebuffering and Memory **2017**. Publisher: arXiv Version Number: 1, doi:10.48550/ARXIV.1703.00633.
56. Nawaz, O.; Fiedler, M.; De Moor, K.; Khatibi, S. Influence of Gender and Viewing Frequency on Quality of Experience. 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), 2020, pp. 1–4. ISSN: 2472-7814, doi:10.1109/QoMEX48832.2020.9123106.
57. Reiter, U.; Brunnström, K.; De Moor, K.; Larabi, M.C.; Pereira, M.; Pinheiro, A.; You, J.; Zgank, A. Factors Influencing Quality of Experience. In *Quality of Experience: Advanced Concepts, Applications and Methods*; Möller, S.; Raake, A., Eds.; T-Labs Series in Telecommunication Services, Springer International Publishing: Cham, 2014; pp. 55–72. doi:10.1007/978-3-319-02681-7_4.
58. Nawaz, O.; Fiedler, M.; Khatibi, S. Impact of Human and Content Factors on Quality of Experience of Online Video Streaming:. Proceedings of the 17th International Joint Conference on e-Business and Telecommunications; SCITEPRESS - Science and Technology Publications: Lieusaint - Paris, France, 2020; pp. 59–66. doi:10.5220/0009831400590066.