Poster: User-Oriented QoE Model for Video Streaming on Mobile Devices

Wangyu Choi and Jongwon Yoon[*]
{wangyu92,jongwon}@hanyang.ac.kr
Hanyang University, ERICA, Republic of Korea

ABSTRACT

The shift from traditional PCs and TVs to mobile devices such as smartphones and tablets has significantly transformed the video streaming landscape. Traditional Quality of Experience (QoE) models, predominantly designed for larger screens, fall short in addressing the nuances of mobile consumption, often misguiding bandwidth usage and quality delivery. This paper introduces a novel user-oriented QoE model tailored for the mobile environment, which accounts for heterogeneous viewing environments. Unlike conventional models that estimate QoE based solely on bitrate and resolution, our approach encompasses the entire video streaming pipeline, from server transmission to the user's perception. In addition, we design lightweight but effective QoE models for mobile devices. This work bridges the gap between user experience and QoE modeling, offering a path toward more adaptive and efficient video streaming services for the increasingly mobile-centric world.

CCS CONCEPTS

• **Information systems** → **Multimedia streaming**; **Mobile information processing systems**.

KEYWORDS

QoE model, Perceived visual quality, Machine learning

ACM Reference Format:
Wangyu Choi and Jongwon Yoon. 2024. Poster: User-Oriented QoE Model for Video Streaming on Mobile Devices. In *ACM International Conference on Mobile Systems, Applications, and Services (Mobisys '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3643832.3661410

1 INTRODUCTION

We have seen the proliferation of smartphones and the development of wireless networks shift the viewing experience of video streaming from PCs and TVs to mobile devices such as tablets and smartphones. Globally, many users are consuming video content on their smartphones and tablets, either on the go or in their spare time. This trend highlights the importance of optimizing the experience on mobile devices for video streaming providers.

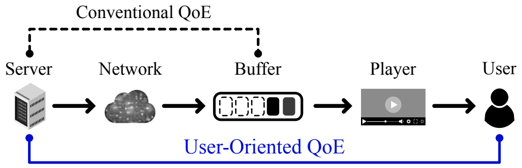


Figure 1: The difference between conventional QoE models and our approach, user-oriented QoE model.

However, traditional QoE (Quality of Experience) models for video streaming have largely focused on large enough screen sizes, such as PCs and TVs, and have failed to account for the mobile device experience. These differences have a direct impact on the quality users experience when consuming video content. Even at the same bitrate and resolution, the experience on mobile devices can be different, and sometimes a lower resolution can provide a satisfactory viewing experience.

Moreover, unnecessarily high bitrates don't just waste bandwidth, they also cause more rebuffering events by a cascade effect. This degrades the user's experience, especially when mobile networks are highly variable. Therefore, we argue for a new QoE model that can guide video streaming services for mobile devices to ensure efficient bandwidth usage while maximizing user experience quality.

In this work, we focus on bridging the gap between user experience and the QoE model in a mobile device environment with heterogeneous screen sizes/resolutions. Figure 1 shows the difference between the user's conventional QoE model and our core idea, the user-oriented QoE model. To measure QoE, conventional QoE models use the bitrate and resolution of a video segment in the playback buffer, while our approach considers the process of a video segment from being rendered to the viewport to entering the human eye (i.e., viewing environments). Our approach considers the entire video streaming pipeline from the server to the user's eyes. Moreover, our proposed QoE model has a lightweight yet effective design that takes into account the computational power of mobile devices. It is better suited to mobile device environments with heterogeneous screen sizes and resolutions.

2 DESIGN

The goal of this work is to accurately estimate the user's perceived quality by considering the user's viewing environment. We calibrate the resolution and frame rate of the video by considering the viewport (i.e., the screen) on which it is rendered and the distance of the

[*] This work was supported by Basic Science Research Program through the National Research Foundation of South Korea (NRF) funded by the Ministry of Education NRF-2022R1A2C1008743 (50%) and Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2024-2020-0-01741, 50%).

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06
https://doi.org/10.1145/3643832.3661410

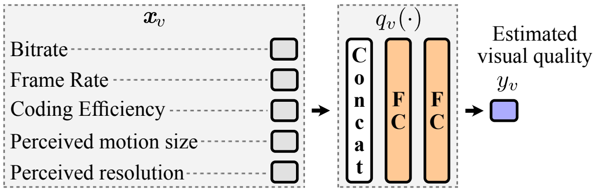


Figure 2: Neural network architecture for the proposed QoE model.

screen from the user's eyes (i.e., the viewing environment), rather than the video frames that are fed into the playback buffer. We then define the inputs to the model and introduce a deep learning-based regression model to estimate the user's QoE.

Perceived resolution and motion size. We define the perceived resolution of the video rendered in the player's viewport as follows: $(w_p, h_p) = (\min(w, w_{\text{eye}}), \min(h, h_{\text{eye}}))$, where w and h are width and height of the video's resolution, respectively. w_{eye} and h_{eye} are the maximum resolutions of the human eye. The perceived resolution cannot surpass the maximum possible resolution human can perceive. The maximum resolution of the human eye can be calculated $(w_{\text{eye}}, h_{\text{eye}}) = \frac{\theta_{\text{viewport}}}{\theta_{\text{vision}}}$, where θ_{viewport} and θ_{vision} are the angles of the viewport size and human vision acuity, respectively. We can calculate the angle size of the viewport using the size of the viewport and the distance between the viewport and the eye:

$$\theta_{\text{viewport}}(w_v, h_v, d) = (2\arctan(\frac{w_v}{2d})\frac{180}{\pi}, 2\arctan(\frac{h_v}{2d})\frac{180}{\pi}) \tag{1}$$

where, w_v, h_v, and d represent the width and height of the viewport and the distance between the viewport and the eye, respectively.

Similarly, we define the perceived motion size m_p using the following formula:

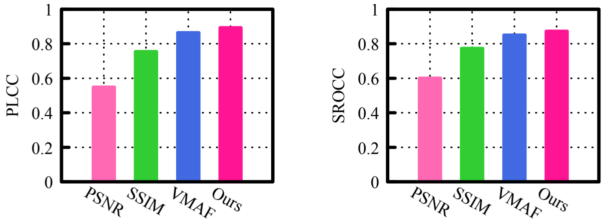
$$m_p = \frac{\theta_{\text{motion}}}{\theta_{\text{vision}}} = \frac{(2\arctan(\frac{m_v}{2d})\frac{180}{\pi})}{\theta_{\text{vision}}} \tag{2}$$

$$m_v = (w_v, h_v) \times (p_{\text{motion}}) = \sqrt{{w_v}^2 + {h_v}^2} \times \frac{m}{\sqrt{w^2 + h^2}} \tag{3}$$

where, m and m_v represent the motion size at the pixel level and the size of motion within the viewport, respectively.

Inputs of the model. Based on the perceived resolution and motion size described earlier, we list the inputs to the model as follows. *(i) Bitrate*: A widely used metric for evaluating visual quality, higher bitrates generally provide higher quality, although other factors must be taken into account. *(ii) Frame rate*: Determines the video's smoothness, typically ranging from 24 fps to 120 fps. *(iii) Coding efficiency*: Refers to the video compression efficiency determined by the codec, pixel pattern of video frames, etc. Video coding methods compress videos based on similarity between frames, therefore the greater the difference between frames, the less efficient it is. *(iv) Perceived motion size*: Refers to the size of object movements in videos considering the user's viewing environment. Users perceive varying degrees of motion based on their viewing environment. *(v) Perceived resolution of viewport*: Refers to the resolution perceived by the user's eyes considering the environment.

Network architecture. Shown in Figure 2, a neural network for visual quality estimation is composed of an input layer, two hidden layers with dropout to mitigate overfitting, a fully connected layer, and an output layer that uses ReLU as the activation function. Note that it is designed to be lightweight to account for the low computational power of mobile devices. We adopt standard supervised learning to train a visual quality estimation model while minimizing the mean squared error (MSE) between the output visual quality and the subjective scores provided by actual users using stochastic gradient descent. We utilize publicly available datasets [1–3] on subjective video streaming quality to determine the perceived visual quality of input $\mathbf{x}_v$.



(a) PLCC **(b) SORCC**

Figure 3: A comparison of existing QoE models and ours.

Training. We train our model on a dataset comprising 40,000 video segments, sourced from combined studies [1–3]. This dataset size is determined to be sufficient to achieve a converged model through extensive preliminary testing. We adopt Adam, which is widely adopted as an optimizer. To optimize the training performance and prevent overfitting, we employ an early stopping technique. Specifically, the training process is halted if the validation loss does not improve for 30 consecutive epochs. We adopt a linear decay strategy for the learning rate, starting at 1e-3 and decreasing to 1e-4 over the training period. This gradual reduction in the learning rate helps in fine-tuning the model's weights more precisely as it approaches the optimal values.

3 PRELIMINARY RESULT

To evaluate the proposed model, we compare ours with existing QoE models that are widely adopted in video streaming solutions: PSNR, SSIM [5], and VMAF [4]. Figure 3 shows the accuracy of PLCC and SROCC of the proposed model compared to the existing QoE models. The results show that the proposed model outperforms state-of-the-art VMAF in terms of accuracy. It achieves a PLCC of 0.89 and SROCC of 0.87 (VMAF is 0.86 and 0.85), indicating a higher level of accuracy compared to others. This validates the proposed model estimates visual quality more closely aligned with the user's perceived quality by utilizing the user-oriented indicator.

REFERENCES

[1] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik. 2018. Towards Perceptually Optimized End-to-end Adaptive Video Streaming. (Aug. 2018). arXiv:1808.03898 [eess.IV]

[2] Nabajeet Barman, Yuriy Reznik, and Maria G Martini. 2023. A Subjective Dataset for Multi-Screen Video Streaming Applications. (May 2023). arXiv:2305.03138

[3] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik. 2021. Subjective and Objective Quality Assessment of High Frame Rate Videos. *IEEE Access* 9 (2021), 108069–108082.

[4] Netflix. 2016. VMAF - Video Multi-Method Assessment Fusion.

[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik. 2003. Multi-scale Structural Similarity for Image Quality Assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, Vol. 2. 1398–1402.