IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 August 2022, accepted 7 September 2022, date of publication 12 September 2022, date of current version 20 September 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3206034*

RESEARCH ARTICLE

# Modeling Perceived Quality on 8K VVC Video Under Various Screen Sizes and Viewing Distances

**YASUKO SUGITO, (Member, IEEE), YUICHI KONDO, DAICHI ARAI, AND YUICHI KUSAKABE**
Science and Technology Research Laboratories, NHK (Japan Broadcasting Corporation), Setagaya-ku, Tokyo 157-8510, Japan

Corresponding author: Yasuko Sugito (sugitou.y-gy@nhk.or.jp)

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of NHK (Japan Broadcasting Corporation) under Approval No. 2021-27, and performed in line with the Declaration of Helsinki.

**ABSTRACT** Perceptual video quality considerably affects the quality of experience (QoE) of watching television (TV) broadcasts. Viewing conditions, such as the screen size and viewing distance, impact the perceived quality. We performed subjective evaluation experiments on 8K (7,680 × 4,320) ultra-high definition (UHD) compressed videos under seven viewing conditions (combinations of 31.5-, 55-, and 85-inch displays and 0.75, 1.5, and 3.0 H (times of screen height) of viewing distance). Distorted videos compressed by the versatile video coding (VVC)/H.266 were used in four types of encoding resolution, from 2K (1,920 × 1,080) to 8K, at a wide bitrate settings range of 3–80 Mbps. We derived a simple regression equation predicting the mean opinion score (MOS) using the hierarchical linear model (HLM), investigating the factors influencing subjective video quality. In this equation, MOS is expressed as a linear combination of terms including intercept and bitrate associated with sequence and encoding resolution, screen size, and viewing distance; it indicates that the smaller the screen, or the further the viewing distance, the fewer artifacts are perceived, as following empirical rules. Furthermore, we confirmed that the derived model is accurate as the Pearson linear and Spearman rank order correlation coefficients between predicted and actual MOS values were more than 0.97.

**INDEX TERMS** 8K ultra-high definition television (UHDTV), hierarchical linear model (HLM), subjective evaluations, versatile video coding (VVC)/H.266, video quality assessments, viewing conditions.

## I. INTRODUCTION

Ultra-high definition television (UHDTV) systems [1] are gradually becoming popular, as indicated by 4K (3,840×2,160) and 8K (7,680×4,320) satellite broadcasting in Japan [2] from 2018. In this first 8K broadcasting service, 8K 59.94-Hz (60-Hz) videos have been compressed in 85 Mbps using the high efficiency video coding (HEVC)/H.265 [3]. Meanwhile, the versatile video coding (VVC)/H.266 [4] was standardized in 2020 as the subsequent video coding scheme of HEVC. Bonnineau *et al.* [5] conducted subjective assessments on both 8K HEVC and VVC encoded videos and reported that VVC exhibits an average of approximately 41% of bitrate reduction over HEVC for the same visual quality. Thus, VVC can become a dominant technique for delivering high-quality UHD videos at considerably lower bandwidth, such as terrestrial transmission.

When watching television (TV) broadcasts, the perceived video quality significantly impacts the quality of experience (QoE) [6]. For the video quality assessments on practical broadcasting, it is necessary to consider a video degradation level caused by compression because video coding is inevitably applied, and the target bitrates differ depending on the transmission paths (e.g., satellite, terrestrial, the Internet). Several models have been proposed to predict the perceptual quality of compressed videos, designed to be well correlated to the subjective evaluation results. ITU-T Rec. P.1204 [7], the following standard of P.1203.1 [8], prescribes

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Bellan.

three models (P.1204.3, 4, and 5) that support up to 4K HEVC videos. These models assume two types of viewing conditions: one is for personal computers or TVs with 24–100 inches screen size at 1.5–3 H (times of screen height) viewing distance, and another is for mobiles or tablets with screens less than 13 inches at 4–6 H. Fremerey *et al.* [9] proposed a model for 360° compressed videos up to 8K watching on a head-mounted display (HMD). Notably, VVC, the latest video coding standard, was not yet considered in these aforementioned models.

Viewers watch TV under diverse viewing circumstances, affecting the perceived video quality. The subjective quality of compressed videos under various viewing distances and screen sizes has been studied [10], [11], [12]. Moreover, we empirically learned that fewer artifacts are observed with a smaller screen or further viewing distance. However, such studies are yet to be performed on 8K videos, whose optimal viewing distance is 0.75 H [13] for an immersive experience. For example, Bonnineau *et al.* [5] conducted subjective evaluations on 8K VVC encoded videos with only one condition, using an 85-inch 8K TV and viewing at the optimal viewing distance.

We performed subjective evaluation experiments on 8K VVC encoded videos under seven viewing conditions (combinations of three types of screen sizes and three types of viewing distances). We analyzed the experimental results using a statistical model to clarify factors that affect subjective visual quality.

The rest of the paper is organized as follows. Section II explains the subjective evaluation experiments on 8K videos. We detail the experimental results in Section III and discuss the results in Section IV. We derive a statistical model based on the subjective results in Section V and discuss the model in Section VI.

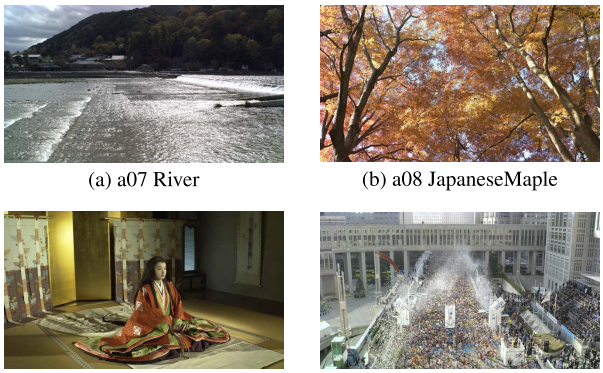
## II. 8K SUBJECTIVE EVALUATION EXPERIMENTS

### A. TEST VIDEOS

Four 8K 60 fps progressive (60p) BT.2020 [1] video sequences were selected from UHD/wide-color-gamut (WCG) standard test sequences - Series A[1] (the River, JapaneseMaple, and LayeredKimono sequences) and B[2] (the Marathon(start) sequence). The duration of the sequences was originally 15 s (900 frames), and we used 6 s (360 frames) for the experiments. Fig. 1 illustrates each sequence's thumbnail image (the first frame of the 6 s).

Their spatio-temporal characteristics determined the selection to ensure widely spread features. Fig. 2 details the mean spatial and temporal perceptual information (SI and TI) [13] that approximate each sequence's spatial and temporal complexity. We converted the 8K sequences from RGB 4:4:4 12 bits to the encoder input format $YC_bC_r$ 4:2:0 10 bits and calculated the SI and TI values from the 10-bit Y component.

[1]https://www.ite.or.jp/content/test-materials/uhdtv_a/
[2]https://www.ite.or.jp/content/test-materials/uhdtv_b/



(a) a07 River (b) a08 JapaneseMaple
(c) a11 LayeredKimono (d) b07 Marathon(start)

**FIGURE 1.** Thumbnail images of the four test sequences.

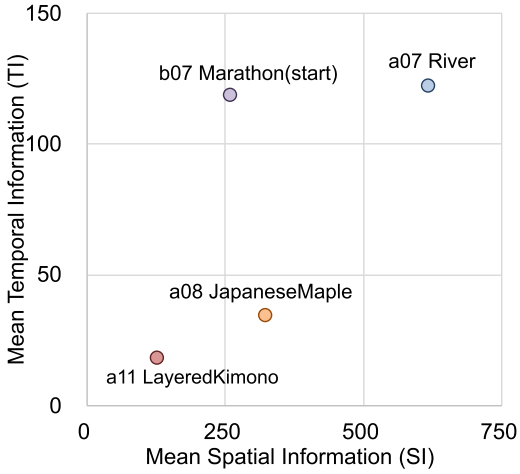



**FIGURE 2.** Spatial-temporal perceptual information of the four test sequences.

**TABLE 1.** Encoding conditions.

| | |
|---|---|
| Encoding resolution | 2K (1,920×1,080)<br>4K (3,840×2,160)<br>6K (5,760×3,240)<br>8K (7,680×4,320) |
| Frame frequency | 60p |
| Up- and down-scaling | Lanczos-3 filter [14] |
| Encoder | VVenC [16] ver. 1.1.0 |
| Configurations | random access, slow, 4:2:0 10 bits |
| Intra period | 32 |
| GOP size | 16 |
| Target bitrates (fixed QP setting) | 3, 5, 7, and 10 Mbps for 2K<br>10, 15, 20, and 30 Mbps for 4K<br>20, 30, 40, and 60 Mbps for 6K<br>25, 40, 60, and 80 Mbps for 8K |

We compressed the 8K sequences using VVC encoder software in a broadcasting set. We down-converted the 8K videos to 2K (1,920×1,080), 4K, and 6K (5,760×3,240) spatial resolutions with the same 60 Hz temporal framerate to generate distorted videos. For all the down- and up-conversion

**TABLE 2.** Actual bitrates and QP settings for test videos.

| Sequence | 2K (**Mbps**/QP for 4 bitrates) | | | | 4K (**Mbps**/QP for 4 bitrates) | | | | 6K (**Mbps**/QP for 4 bitrates) | | | | 8K (**Mbps**/QP for 4 bitrates) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a07 | **3.3**/42 | **5.0**/40 | **7.5**/38 | **9.1**/37 | **9.8**/45 | **15.2**/43 | **19.0**/42 | **28.9**/40 | **20.0**/46 | **31.1**/44 | **38.6**/43 | **59.3**/41 | **24.7**/48 | **38.6**/46 | **60.2**/44 | **74.8**/43 |
| a08 | **3.1**/41 | **5.0**/38 | **6.8**/36 | **9.4**/34 | **10.3**/41 | **14.2**/39 | **19.2**/37 | **30.9**/34 | **19.0**/41 | **29.6**/38 | **39.9**/36 | **63.3**/33 | **24.2**/42 | **37.9**/39 | **58.3**/36 | **79.5**/34 |
| a11 | **3.1**/26 | **4.8**/22 | **6.7**/19 | **10.0**/16 | **10.4**/23 | **15.2**/21 | **18.7**/20 | **31.6**/18 | **19.5**/24 | **25.4**/23 | **35.7**/22 | **67.1**/20 | **21.4**/27 | **42.0**/25 | **60.7**/24 | **87.2**/23 |
| b07 | **3.0**/45 | **5.4**/42 | **6.5**/41 | **9.3**/39 | **9.8**/45 | **13.8**/43 | **19.3**/41 | **30.6**/38 | **21.2**/44 | **28.9**/42 | **38.9**/40 | **59.3**/37 | **23.2**/46 | **42.1**/42 | **63.9**/39 | **82.9**/37 |

processes, the Lanczos-3 filter [14] using FFmpeg[3] was applied according to previous studies [5], [15]. Next, we load the down-converted and 8K original videos to the encoder. The encoding conditions are presented in Table 1. With the random access configuration, the group of picture (GOP) size was set at 16, and intra pictures were inserted every 32 frames (approximately 0.5 s).

The target bitrates were determined based on the lower threshold of the broadcasting service's required bitrate using HEVC: 10, 30, and 80 Mbps for 2K, 4K, and 8K videos, respectively [17]. Next, we estimated the bitrate for 6K videos as 60 Mbps, slightly higher than the mean of 4K and 8K, and 30, 50, 70, and 100% of the lower required bitrate for each encoding resolution were considered as the target bitrates in the experiment. We adjusted a quantization parameter (QP) value to be the closest bitrate to each target with a fixed QP setting. The actual bitrates and QP values are shown in Table 2. A previous study on 8K subjective evaluations revealed that VVC exhibits an average of approximately 41% of bitrate reduction over HEVC for the same visual quality [5]. Therefore, we considered that target bitrates widely cover video quality from low to high in this range.

We generated 16 encoded videos per sequence and up-converted 2K, 4K, and 6K compressed videos to 8K. In addition to the compressed videos, we prepared four uncompressed videos per sequence, one being the original 8K video. The other three were 8K videos that were up-converted from the 2K, 4K, and 6K down-converted original videos, referred to as 2K, 4K, and 6K original videos, respectively.

### *B. SUBJECTIVE EVALUATION EXPERIMENTS*

#### 1) VIEWING CONDITIONS

We equipped an 8K uncompressed recorder for the experiments that stored the test videos and three distinct (31.5, 55, and 85-inch) 8K liquid crystal display (LCD) monitors. Before the experiments, the luminance, white point of D65, and contrast of the displays were adjusted using a color luminance meter while presenting the PLUGE signal [18]. The peak luminance was set to 100 cd/m$^2$, which is a professional setting for standard dynamic range (SDR) videos [19].

We set seven viewing conditions to investigate differences in the perception of video quality with the monitor size and viewing distance, as presented in Table 3. Regarding the viewing distance, 0.75, 1.5, and 3.0 H represent the optimal viewing distance for 8K, 4K, and 2K videos, respectively [13]. A viewing point was set for a subject sitting on a chair directly in front of each screen.

[3]https://ffmpeg.org/

**TABLE 3.** Seven experimental viewing conditions.

| Screen size (width×height) | Viewing distance in H (in m) |
|---|---|
| 31.5 inch (0.70 m×0.39 m) | 0.75, and 1.5 H (0.3, and 0.6 m) |
| 55 inch (1.22 m×0.68 m) | 0.75, and 1.5 H (0.5, and 1.0 m) |
| 85 inch (1.88 m×1.06 m) | 0.75, 1.5, and 3.0 H (0.8, 1.6, and 3.2 m) |

#### 2) EVALUATION METHOD

Subjective evaluation experiments were performed using the single-stimulus (SS) method prescribed in ITU-R Rec. BT.500 [13]. An equivalent method called absolute category rating (ACR) is defined in ITU-T Rec. P.913 [20]. We selected the SS method because it is practically appropriate. When watching TV programs, a compressed video is solely displayed, but the uncompressed reference video is not presented. In the experiments, mid-gray with a video number (1 s), a test video (6 s), and mid-gray with "VOTE" (3 s) were presented, and subjects graded the video quality at a five-Likert scale (5, Excellent; 4, Good; 3, Fair; 2, Poor; 1, Bad) by the end of the display of "VOTE."

First, each subject signed a consent form after receiving the experimental overview information. A verbal instruction based on a sample instruction for ACR described in Appendix II of P.913 was provided. Subjects were encouraged (1) to evaluate a part in front of them, (2) carefully observe the entire clip before judging, (3) rate the general quality of the video rather than the content, and (4) frankly answer a query on video quality when they saw this clip on a TV screen.

Subsequently, a training session was conducted, including the highest and lowest quality 8K compressed videos. Subjects evaluated five test items generated from three sequences that differed from those introduced in Section II-A, namely, the SteelPlant, Festival, and Water polo(scrolling text) sequences from the UHD/WCG test sequences A and B.

Eighteen video experts familiar with 8K videos for research purposes participated in the evaluations. Each of them assessed a 13-min session consisting of 80 test videos ((16 compressed + 4 uncompressed videos)×4 sequences) under seven viewing conditions. Two sessions were conducted simultaneously to relieve subjects' fatigue and increase the convenience of the executions, and observers

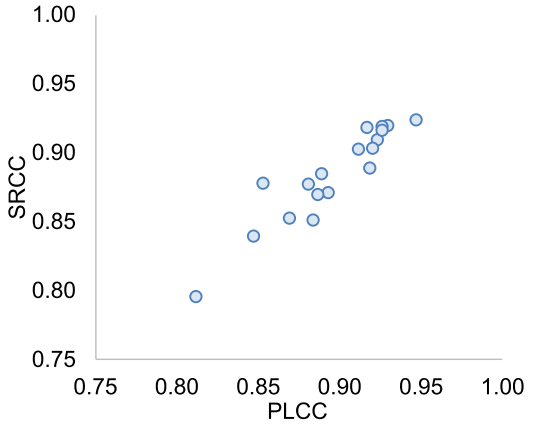



**FIGURE 3.** Correlation coefficients for 18 subjects.

took at least a 30-min rest before the following sessions. Considering the order effect, we prepared four types of playlists with distinct order of the test videos and randomly played one of the four for each session. Furthermore, each subject observed the test videos in a different order of the viewing conditions.

## III. EXPERIMENTAL RESULTS

### A. SCREENING OF SUBJECTS

A screening method described in BT.500-14 [13] section A7-5.3 was applied. In this method, the rejection threshold was determined based on the Pearson linear correlation coefficient (PLCC) and the Spearman rank order correlation coefficient (SRCC) between individual scores and the mean scores of all 18 subjects for all 560 items (80 test videos × seven viewing conditions). Fig. 3 is the scatter plot of PLCCs and SRCCs of the 18 evaluators, distributed between 0.80 and 0.95. Because all the CCs are more significant than 0.70, the maximum correlation threshold of the SS method, no outlier was detected. Thus, all experimental results were used in the following sections.

### B. SUBJECTIVE EVALUATION RESULTS

The mean opinion score (MOS), a subjective quality, was calculated from the results of 18 viewers. The four graphs in Fig. 4 denote the MOS values for each test sequence. The horizontal axis revealed the encoding resolution and bitrate in Mbps or the original video described as "ori." The marker shapes indicate the screen sizes: triangle, diamond, and circle markers correspond to 31.5, 55, and 85 inches, respectively. The line shapes express the viewing distances: 0.75, 1.5, and 3.0 H are represented by the dotted, dashed, and solid lines, respectively. The error bars denote a 95% confidence interval (CI) using Student's t-distribution. The graph legends are sorted in the descending order of the viewing distance and the screen size ascending order, roughly the descending order of MOS values.

Tables 4, 5, and 6 reveal the PLCC, SRCC, and the root-mean-square error (RMSE) between each combination of the seven types of viewing conditions, respectively.

**TABLE 4.** PLCC between each combination of viewing conditions.

| PLCC | **31-1.5H** | **55-1.5H** | **85-1.5H** | **31-0.75H** | **55-0.75H** | **85-0.75H** |
|---|---|---|---|---|---|---|
| **85-3.0H** | 0.934 | 0.932 | 0.920 | 0.901 | 0.865 | 0.850 |
| | **31-1.5H** | 0.988 | 0.984 | 0.981 | 0.958 | 0.945 |
| | | **55-1.5H** | 0.992 | 0.988 | 0.975 | 0.964 |
| | | | **85-1.5H** | 0.988 | 0.975 | 0.964 |
| | | | | **31-0.75H** | 0.987 | 0.979 |
| | | | | | **55-0.75H** | 0.993 |

**TABLE 5.** SRCC between each combination of viewing conditions.

| SRCC | **31-1.5H** | **55-1.5H** | **85-1.5H** | **31-0.75H** | **55-0.75H** | **85-0.75H** |
|---|---|---|---|---|---|---|
| **85-3.0H** | 0.939 | 0.940 | 0.948 | 0.938 | 0.938 | 0.939 |
| | **31-1.5H** | 0.983 | 0.978 | 0.985 | 0.984 | 0.976 |
| | | **55-1.5H** | 0.980 | 0.982 | 0.984 | 0.984 |
| | | | **85-1.5H** | 0.985 | 0.978 | 0.975 |
| | | | | **31-0.75H** | 0.986 | 0.984 |
| | | | | | **55-0.75H** | 0.988 |

**TABLE 6.** RMSE between each combination of viewing conditions.

| RMSE | **31-1.5H** | **55-1.5H** | **85-1.5H** | **31-0.75H** | **55-0.75H** | **85-0.75H** |
|---|---|---|---|---|---|---|
| **85-3.0H** | 0.529 | 0.639 | 0.707 | 0.807 | 1.025 | 1.128 |
| | **31-1.5H** | 0.234 | 0.314 | 0.375 | 0.613 | 0.738 |
| | | **55-1.5H** | 0.181 | 0.237 | 0.449 | 0.575 |
| | | | **85-1.5H** | 0.209 | 0.394 | 0.508 |
| | | | | **31-0.75H** | 0.293 | 0.421 |
| | | | | | **55-0.75H** | 0.198 |

### C. OBJECTIVE QUALITY METRICS

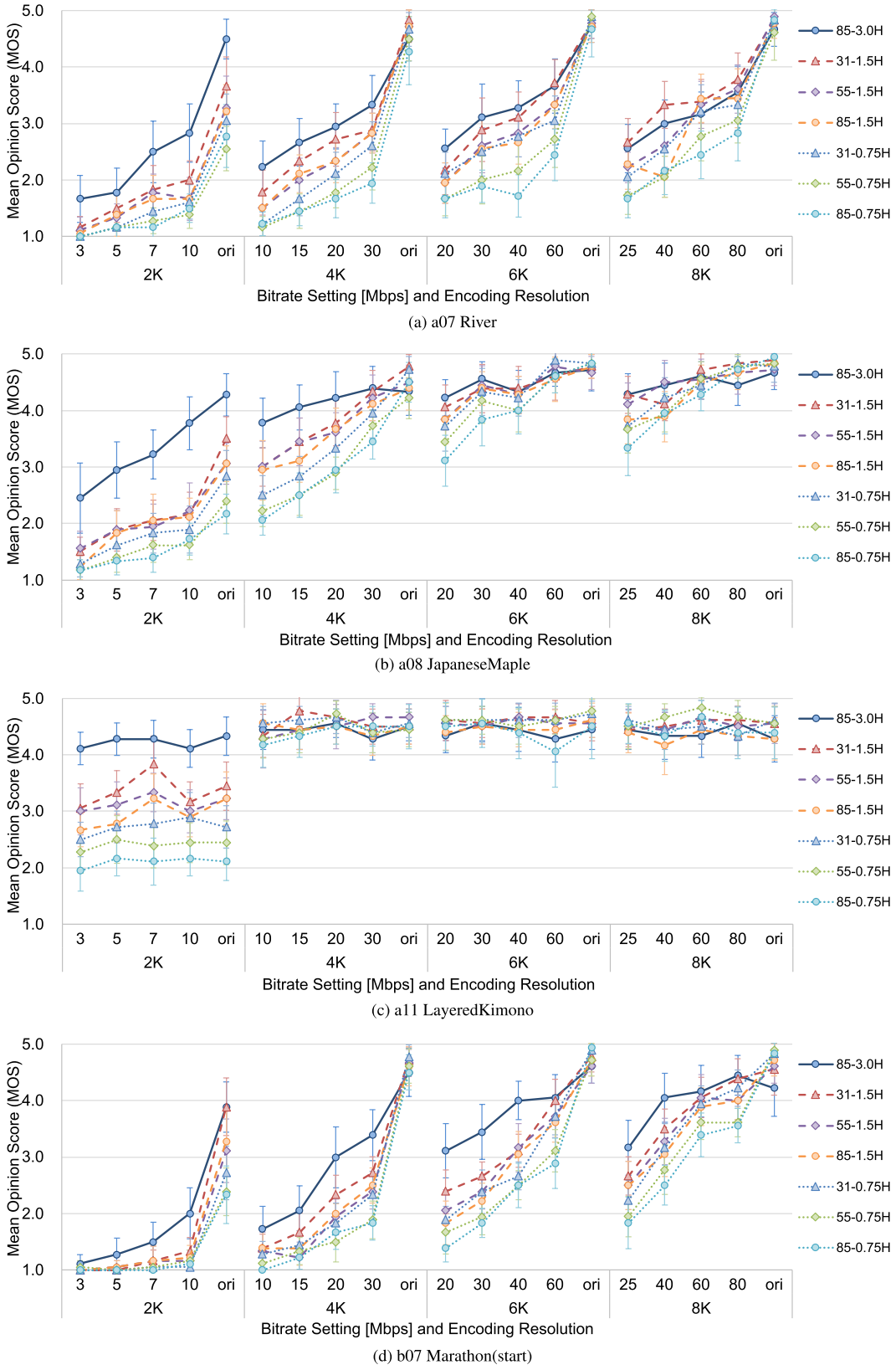
Four standard objective quality metrics, namely, peak signal-to-noise ratio (PSNR), structure similarity index (SSIM) [21], multi-scale SSIM (MS-SSIM) [22], and video multimethod assessment fusion (VMAF) [23], were considered for compressed videos (64 videos per viewing condition) and compared to the subjective evaluation results. For the computations, we used VMAF v2.3.0[4] (October 2021) and the v.0.6.1 model, intended for 2K videos. Since all the four metrics are full reference, loading both reference and distorted images is necessary for the calculations. We used the 8K original videos as a reference input of the metrics.

The performance of the objective metrics was evaluated similarly to previous related studies [5], [24]. The consistency between the metric values and the subjective evaluation results was investigated by the logistic curve fitting based on the least square method as follows:

$$\hat{y} = a + \frac{b}{1 + \exp(-c(x - d))}, \tag{1}$$

where $x$ and $\hat{y}$ denote the objective metric value and the predicted MOS, respectively. The true MOS $y$ corresponding to $x$ was obtained from the subjective evaluation. The variables

[4]https://github.com/Netflix/vmaf




(a) a07 River

(b) a08 JapaneseMaple

(c) a11 LayeredKimono

(d) b07 Marathon(start)

**FIGURE 4.** **Subjective evaluation results for each sequence.**

$a$, $b$, $c$, and $d$ are selected to minimize $\sum_i (y_i - \hat{y}_i)^2$ for all items $i$.

We assessed the performance in terms of PLCC, SRCC, and RMSE concerning the corresponding relationship

**TABLE 7.** Correlations and RMSEs to objective quality metrics.

| viewing | PSNR | | | | SSIM | | | | MS-SSIM | | | | VMAF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| condition | PLCC | SRCC | RMSE | RMSE* | PLCC | SRCC | RMSE | RMSE* | PLCC | SRCC | RMSE | RMSE* | PLCC | SRCC | RMSE | RMSE* |
| 85-3.0H | **0.862** | **0.857** | **0.492** | **0.282** | **0.945** | **0.902** | **0.319** | **0.101** | **0.946** | 0.911 | **0.314** | **0.111** | **0.928** | 0.912 | **0.361** | **0.154** |
| 31-1.5H | 0.836 | 0.814 | 0.644 | 0.442 | 0.884 | 0.878 | 0.550 | 0.304 | 0.909 | 0.904 | 0.489 | 0.261 | 0.914 | 0.912 | 0.477 | 0.250 |
| 55-1.5H | 0.852 | 0.818 | 0.631 | 0.411 | 0.890 | 0.881 | 0.551 | 0.307 | 0.911 | 0.904 | 0.497 | 0.278 | 0.916 | 0.909 | 0.484 | 0.262 |
| 85-1.5H | *0.790* | *0.800* | 0.716 | 0.497 | 0.879 | *0.864* | 0.557 | 0.327 | 0.901 | *0.888* | 0.506 | 0.297 | 0.907 | *0.895* | 0.492 | 0.278 |
| 31-0.75H | 0.802 | 0.814 | *0.749* | *0.543* | 0.879 | 0.876 | 0.597 | 0.379 | 0.903 | 0.901 | 0.539 | 0.342 | 0.913 | 0.912 | 0.510 | 0.311 |
| 55-0.75H | 0.827 | 0.832 | 0.726 | 0.520 | 0.876 | 0.889 | 0.624 | 0.413 | 0.901 | **0.915** | 0.561 | 0.370 | 0.916 | **0.924** | 0.519 | 0.321 |
| 85-0.75H | 0.818 | 0.823 | 0.723 | 0.537 | *0.864* | 0.882 | *0.634* | *0.438* | *0.889* | 0.904 | *0.576* | *0.402* | *0.906* | 0.916 | *0.533* | *0.356* |

The table displays the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SRCC), the root-mean-square error (RMSE), and the epsilon-insensitive RMSE (RMSE*) between subjective evaluation results represented as MOS and an objective quality metric for each viewing condition. RMSE* considers a 95% confidence interval (CI) of MOS, shown as the error bar in Fig. 4, when calculating errors.

between $y_i$ and $\hat{y}_i$. Furthermore, we calculated the epsilon-insensitive RMSE (RMSE*) [25], which considers a 95% CI of MOS shown as the error bar in Fig. 4. The error between $\hat{y}$ and $y$ will become zero if $\hat{y}$ is within the 95% CI of MOS $y$. PLCC, SRCC, and RMSEs measure linearity, monotonicity, and accuracy, respectively. The CCs should be 1, whereas the RMSEs should be 0.

We calculated the performance results of the metrics as presented in Table 7. As subjective evaluation was conducted under the seven viewing conditions, seven MOS values were obtained for each test video. The viewing conditions are sorted in the same order as that of Fig. 4. The figures in **bold** and *italic* indicated the best and worst results in the seven conditions, respectively.

## IV. DISCUSSION ON EXPERIMENTAL RESULTS

### A. SUBJECTIVE EVALUATION RESULTS

The experimental results in Fig. 4 and Table 6 exhibited a trend that complies with empirical rules. The MOS is increasing with the viewing distance for the same screen size and is decreasing with the screen size for the same viewing distance. Furthermore, the high SRCCs (0.938 or greater) in Table 5 imply the magnitude relationship of MOS is nearly consistent for each viewing condition.

To investigate comprehensively, we plotted a bitrate ladder in Fig. 5 for each sequence at the lowest MOS case with 0.75 H of 85 inches. In the graphs, the blue, red, green, and purple circle points indicate the MOS values for the 2K, 4K, 6K, and 8K encoding resolutions at actual bitrates, respectively. The error bars denote a 95% CI using the Student's t-distribution. The dashed lines in the same colors correspond to the MOS values of the original videos for each spatial resolution.

The graphs revealed encoding resolution changes affect the perceived video quality and depend on sequences. For instance, for all four sequences, the MOS on the 2K original video in the blue dashed line is less than 3; thus, 2 (Poor) or 1 (Bad) was graded by some evaluators. Among them, the LayeredKimono sequence in Fig. 5 (c) is peculiar. Overall, the encoding complexity of this sequence is low as MOS values on 4K, 6K, and 8K encoding resolutions are more than 4 (the MOS for the 4K original video equals 6K). However, the MOS values drastically decreased on 2K, even in the originally uncompressed video. Observers perceived the aliasing on diagonal edges annoying, which is associated with

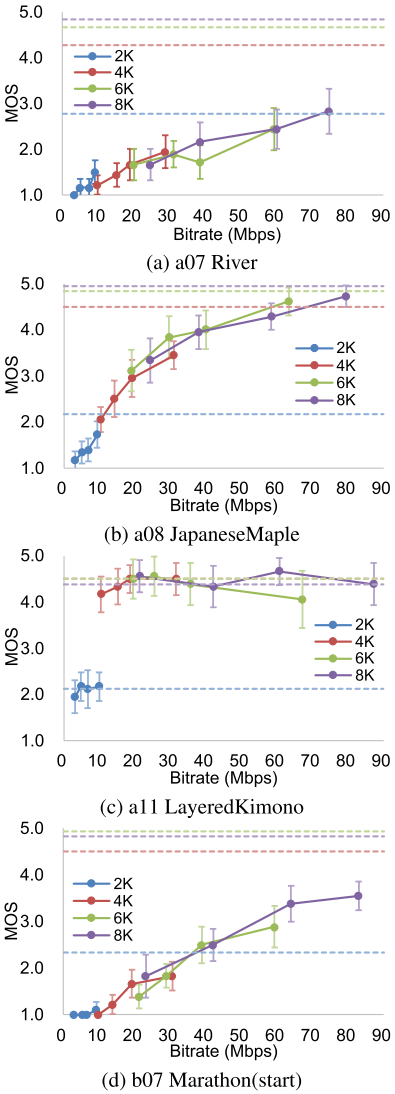



**FIGURE 5.** Bitrate ladder for each sequence at 85-0.75 H.

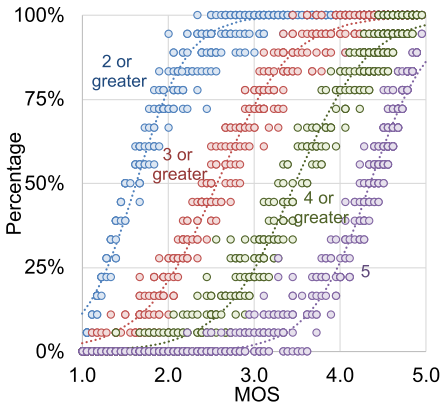



**FIGURE 6.** Score distribution per MOS.

the down-conversion process from 8K to 2K. Furthermore, the down-conversion filter could be improved.

In VVC, adaptive resolution change (ARC), which includes reference picture resampling (RPR), allowing the spatial resolution of intra- and inter-coded pictures to change, has been newly adopted. This technique can effectively improve coding efficiency [26]. For example, at bitrates between 4K, 6K, and 8K encoding resolutions (20–60 Mbps), no significant difference between the experimental results was observed for all sequences (see Fig. 5). In contrast, as mentioned previously, the LayeredKimono sequence at 2K exhibited poor results. Such findings should help determine an appropriate spatial resolution for ARC at a target bitrate.

### B. SCORE DISTRIBUTION

Subjective assessments on 8K compressed videos have typically been conducted using a double-stimulus method [5], [17], [27]. For example, in the double-stimulus impairment scale (DSIS) method [13], a test image is presented after the corresponding reference image, and subjects evaluate the fidelity of the test image relative to the reference by using a five-Likert scale. In contrast, we adopted the SS method described in Section II-B. Thus, we anticipated expert viewers familiar with the DSIS method to grade a score with a distinct tendency when using the SS method. We hypothesized that they hardly choose "3," which denotes "Fair," because of the difficulty in grading at an absolute scale.

We investigated the distribution of the scores as in our previous studies [27], [28] to verify the assumption. Fig. 6 presents the relationship between the MOS values (horizontal axis) and the percentages of scores (vertical axis).

From left to right, the circles in blue, red, green, and purple correspond to the score ranges of 2 or greater (2–5), 3 or greater (3–5), 4 or greater (4–5), and 5, respectively. The dotted line indicates the fitted curve of a logistic function for each score range using the least-squares method:

$$\hat{y}_X = \frac{1}{1 + \exp(-a_X(x - b_X))}, \tag{2}$$

**TABLE 8.** Variables of the logistic functions (2).

| | $a_3$ | $b_3$ | $a_4$ | $b_4$ | $a_5$ | $b_5$ |
|---|---|---|---|---|---|---|
| SS | 2.34 | 2.57 | 2.35 | 3.48 | 2.86 | 4.35 |
| DSIS [27] | 2.31 | 2.55 | 2.27 | 3.45 | 3.03 | 4.38 |

where $x$ and $\hat{y}_X$ denote a MOS value and a predicted proportion of scores X or greater, respectively. The actual proportion $y_X$ corresponding to $x$ is plotted as a circle in the graph. The variables $a_X$ and $b_X$ are selected to minimize $\sum_{\text{all conditions } i}(y_{Xi} - \hat{y}_{Xi})^2$: $a_X$ determines the distribution width of the scores, whereas $b_X$ indicates the MOS value that results in $\hat{y}_X = 0.5$. Table 8 shows the specific values of the variables $a_X$ and $b_X$, $X = 3 - 5$. For comparison, we also arranged those of the DSIS case from our previous study [27] in the table. We did not show the values for $X = 2$ because of the lack of MOS values less than 2 in the DSIS case, and more than half of the evaluators overlapped in the two experiments. The values in Table 8 revealed that the distributions of scores 3–5 are like one another, which is contrary to our prediction.

Recently, Pinson [29] proposed $\Delta S_{CI}$, a novel method for measuring the precision of subjective tests. In this method, for each pair of stimuli A and B, the absolute difference of MOS $\Delta S$ (i.e., $\Delta S = |\text{MOS}(A) - \text{MOS}(B)|$) is measured, and a Student's paired t-test is conducted between individual scores of A and B at a 95% confidence level. Then, bin $\Delta S$ by 0.1 MOS intervals (0 to 0.05, 0.1 $\pm$ 0.05, 0.2 $\pm$ 0.05, ...) and compute $\pi$, the percentage of pairs A and B that show the statistical difference. $\Delta S_{CI}$ is defined as the $\Delta S$ that comes closest to producing $\pi = 95\%$. Through the investigations over various datasets mostly evaluated by non-experts, $\Delta S_{CI} = 0.5$ for 24 subjects and $\Delta S_{CI} = 0.7$ for 15 subjects when the 5-level ACR scale was used. As our previous studies indicated that expert results differ from those of non-experts [27], [28], we calculated $\pi$ for each 0.1 of $\Delta S$ using our results obtained from 18 video experts: $\pi = 88\%$ for $\Delta S = 0.6$ and $\pi = 98\%$ for $\Delta S = 0.7$. For comparison, we randomly selected 15 subjects from the 18 subjects and calculated the mean $\pi$ of 100 trials: $\pi = 92\%$ for $\Delta S = 0.7$; and $\pi = 98\%$ for $\Delta S = 0.8$. We confirmed that our experimental results follow the existing $\Delta S_{CI}$ rule.

### C. OBJECTIVE QUALITY METRICS

As presented in Table 7, the MOS values of 85-3.0H exhibited the best correlations with the four objective quality metrics among the seven viewing conditions, and the results in the viewing conditions with 0.75 H were inferior to others. This phenomenon could be attributed to the following reasons. (1) We applied the VMAF 2K model trained by subjective evaluation results observed from the viewing distance of 3 H [23], and (2) the viewing distance of a dataset used to determine the parameters of MS-SSIM was 32 pixels per degree of visual angle [22], which should be more than 3 H considering the test patches were 64 $\times$ 64 pixels.

**TABLE 9.** Explanatory variables considered for the MOS modeling.

| Variables | Scale | Descriptions |
|---|---|---|
| br (bitrate) | ratio | actual bitrate (3.023 to 87.248 Mbps) |
| seq (sequence) | nominal | a07, a08, a11, and b07 |
| res (encoding resolution) | ratio | 2K, 4K, 6K, and 8K |
| inch (screen size) | ratio | 31.5, 55, and 85 inches |
| dist (viewing distance) | ratio | 0.75, 1.5, and 3.0 H |

These findings proved the necessity of objective quality metrics suited to 8K observed from a viewing distance of 0.75 H. For example, the VMAF 4K model, which predicts the subjective quality of video displayed on a 4K TV and viewed from 1.5 H, was developed [30]. Such an approach can be a solution.

## V. STATISTICAL MODELING OF MOS

### A. HIERARCHICAL LINEAR MODEL

We conducted a regression analysis using the hierarchical linear model (HLM) [31] to investigate factors that affect the MOS values on compressed videos. This is also called the multi-level and mixed model, which can treat nested structure data (e.g., students within classrooms within schools). Although such a model is a standard method in psychology or sociology, it is typically not applied to analyze subjective assessments, as pointed out in a previous related study [32].

Conventionally, the general linear model, including t-test, F-test, and analysis of variance (ANOVA), has been used for statistical analyses on subjective evaluation results. For example, in a simple linear regression (3), the error term $\epsilon_i$ is assumed independent, and each sample is uncorrelated to others.

$$Y = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{3}$$

However, nested data can stray from this independent assumption because samples in a group (for example, students in the same school) may have a similar tendency. In such a case, HLM is appropriate, but the general linear model is not suitable. HLM is applicable to analyze MOS values on compressed videos, which should be varied depending on sequences, encoding conditions, and subjects [32], [33].

### B. MODEL DERIVATION

We considered five parameters for the modeling, as presented in Table 9 as explanatory variables of MOS. The above three are video coding conditions, and the following two are viewing conditions.

In a regression model using HLM, regression coefficients can be expressed as a summation of fixed and random effects. The fixed effect denotes the expected value calculated from all data. The random effect differs from a fixed effect and varies with groups (e.g., sequences) or individuals.

Firstly, we focused on the relationship between MOS and bitrate. MOS enlarges with the increasing bitrate. Secondly, to investigate the necessity of HLM, we calculated the intra-class correlation coefficient (ICC) that measures the similarity within a group for the rest of the considered variables in Table 9. Equation (4) is the definition of ICC, where $\sigma_b^2$ and $\sigma_w^2$ is the between-group and within-group variances, respectively.

**TABLE 10.** ICC results.

| Variables | ICC |
|---|---|
| seq | 0.417 |
| res | 0.346 |
| inch | 0.000 |
| dist | 0.050 |

$$\mathrm{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}. \tag{4}$$

As denoted in Table 10, the ICCs of the encoding conditions (sequence and encoding resolution) were more significant than those of the viewing conditions (screen size and viewing distance). Thus, we applied HLM to the encoding but not the viewing conditions.

We studied several candidate models using R[5] ver.4.1.3 (March 2022) and selected a simple yet sufficient performance model. The performance was measured by a goodness of fit in terms of Akaike's information criterion (AIC), Bayesian information criterion (BIC), and log-likelihood (logLik). Appendix A outlines details of the candidate models and their performance.

Equation (5) detailed the derived model in Wilkinson notation. In the formula, the MOS value on the left-hand side is defined as a linear combination of intercept (denoted as "1"), bitrate (br), the screen size (inch), and viewing distance (dist). Furthermore, HLM is applied to the left-hand side of the parenthesized terms, intercept and bitrate, and the variances separated by "/" after "|" indicate levels: a level of resolution (res) under that of sequence (seq) exists.

$$\mathrm{MOS} \sim 1 + \mathrm{br} + \mathrm{inch} + \mathrm{dist} + (1 + \mathrm{br}|\mathrm{seq/res}). \tag{5}$$

Equation (6) describes the specific regression formula of our model for estimating the MOS value for bitrate $i$, screen size $j$, and viewing distance $k$.

$$\mathrm{MOS}_{ijk} = \beta_0 + \beta_1 i - 0.005j + 0.332d_{k,D2} + 0.899d_{k,D3}. \tag{6}$$

Here, we explain the terms on the right-hand side of (6) from left to right. Table 11 details the specific values of $\beta_0$ and $\beta_1$, which are the intercept and the slope of bitrate $i$, respectively. In this model, $\beta_0$ and $\beta_1$ resulted in distinct values depending on the sequence (seq) and encoding resolution (res) because these were separately derived as the fixed effect and the two types of the random effect that vary with seq and res in seq (denoted as res:seq). We provided specific values for the fixed and random effects in Appendix B, and the

[5]https://www.R-project.org/

**TABLE 11.** Specific values of $\beta_0$ and $\beta_1$ in (6).

| seq | res | $\beta_0$ | $\beta_1$ |
|---|---|---|---|
| a07 | 2K | 2.368 | 0.047 |
| | 4K | 2.450 | 0.052 |
| | 6K | 2.157 | 0.029 |
| | 8K | 2.120 | 0.024 |
| a08 | 2K | 3.623 | 0.082 |
| | 4K | 3.744 | 0.055 |
| | 6K | 3.839 | 0.019 |
| | 8K | 3.777 | 0.015 |
| a11 | 2K | 3.997 | 0.053 |
| | 4K | 4.251 | 0.002 |
| | 6K | 4.277 | -0.001 |
| | 8K | 4.231 | 0.001 |
| b07 | 2K | 2.240 | 0.056 |
| | 4K | 2.234 | 0.056 |
| | 6K | 2.246 | 0.040 |
| | 8K | 2.464 | 0.029 |

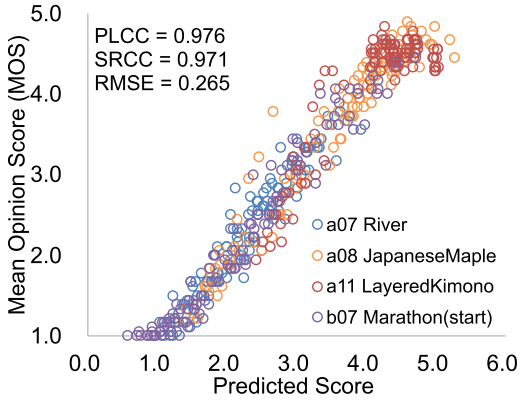



**FIGURE 7.** Relationship between predicted score and MOS.

figures in Table 11 were their summation. The bitrate $i$ was centralized at 30 Mbps, e.g., $i = -5$, 0, and 10 correspond to 25, 30, and 40 Mbps, respectively.

Regarding the terms related to the viewing conditions, the screen size $j$ was centralized at 55 inches. The viewing distance $k$ was considered as a categorical value: D1, D2, and D3 are compatible with 0.75, 1.5, and 3.0 H, respectively. Equation (7) represents the dummy variable $d$, and the last two terms of (6) express the change amount of the MOS when the viewing distance has been altered from D1 to D2 and D3.

$$d_{x,y} = \begin{cases} 1 & (x = y) \\ 0 & (x \neq y) \end{cases} \tag{7}$$

### C. MODEL EVALUATION

We evaluated the derived model's performance. The scatter plot in Fig. 7 illustrates the relationship between predicted scores of the model (horizontal axis) and the true scores, MOS values (vertical axis). As described in the graph, PLCC, SRCC, and RMSE between the predicted and true scores were 0.976, 0.971, and 0.265, respectively. Note that these similarities were calculated without the curve fitting of (1). The evaluation results revealed that the proposed model can predict MOS with sufficient accuracy.

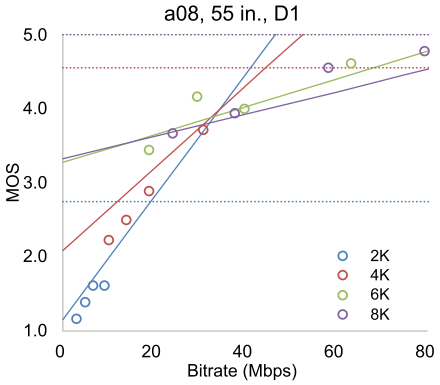



**FIGURE 8.** Example of the regression lines and MOS for a08, 55-0.75H.

## VI. DISCUSSION ON DERIVED MODEL

We confirmed that the regression formula in (6) reflects the empirical rules on viewing conditions. The regression coefficient of the display size $j$ denotes MOS decreasing by 0.005 if the display size increased one inch from 55 inches. The last two terms in (6) related to the viewing distance $k$ indicate the MOS increases by 0.332 and 0.899 if the viewing distance was expanded to 1.5 and 3.0 H from 0.75 H, respectively. Also, $\beta_0$ signifies MOS in the bitrate of 30 Mbps ($i = 0$) viewing at 0.75 H ($d_{k,D2} = d_{k,D3} = 0$ for $k = D1$) with a 55-inch display ($j = 0$), and MOS increases by $\beta_1$ if the bitrate was increased 1 Mbps from 30 Mbps.

Although the MOS values for the original videos were not used to derive (6), those values in each spatial resolution should become the upper limit of MOS in the corresponding encoding resolution. For example, Fig. 8 presents an example of the regression lines (in solid lines), MOS+95%CI of the original videos, the upper limit is 5.0, (in dotted lines), and MOS values at actual bitrates (in circles). The lines and circles in blue, red, green, and purple correspond to the results of spatial resolutions at 2K, 4K, 6K, and 8K, respectively, and both dotted lines of 6K and 8K are in 5.0. In the graph, the slope of the 2K regression line in blue is considerably steeper than others (also see $\beta_1$ of a08 in Table 11). However, the MOS values in the 2K encoding resolution could be saturated at the bitrate crossed to the blue dotted line, approximately 20 Mbps. A knee point will be observed at approximately 20 Mbps if we conduct subjective assessments at higher bitrates on the 2K encoding resolution. In the proposed model, a simple linear combination works well if the bitrate range is limited.

## VII. CONCLUSION

We encoded four 8K sequences in 2K, 4K, 6K, and 8K encoding resolutions at four bitrates for each resolution using VVC and conducted subjective evaluation experiments under seven viewing conditions with distinct screen sizes and viewing

**TABLE 12.** Goodness of fit for each candidate model.

| Model | AIC | BIC | logLik |
|---|---|---|---|
| M0 | 1457.4 | 1465.6 | -726.7 |
| M1 | 1235.2 | 1247.5 | -614.6 |
| M2 | 785.6 | 802.0 | -388.8 |
| M3 | 546.9 | 583.9 | -264.5 |
| M4 | 242.1 | 291.4 | -109.1 |
| M5 | 212.9 | 262.1 | -94.4 |
| M6 | 210.0 | 259.2 | -93.0 |

distances. The subjective results proved the empirical rules of the relationship between the perceived video quality, screen size, and viewing distance. The smaller the screen, or the further the viewing distance, the fewer artifacts are observed. Moreover, the results on encoding resolutions that vary in the sequences can be applied for ARC, a newly adopted technique of VVC.

From the experimental results, we derived a simple regression equation that predicts MOS using HLM, and MOS could be formulated by using a simple linear combination of the terms (intercept and bitrate associated with sequence and encoding resolution, screen size, and viewing distance). We evaluated the derived model's performance regarding the similarities between the predicted and actual MOS values and confirmed the high accuracy as both PLCC and SRCC are more than 0.97 and RMSE is less than 0.30.

From this study, we reconfirmed that subjects feel limited deterioration in the 31.5-inch 8K display than for the larger 8K displays. However, with the smaller screen, observers may feel less "sense of being there," which is a feature of 8K [34]. For our future study, we plan to extend our evaluation of the video quality to an evaluation of QoE on 8K compressed videos.

## APPENDIX A
## CANDIDATE MODELS

In this Appendix, the goodness of fit for the seven candidate models M0–M6 was detailed. First, the models were denoted in Wilkinson notation as follows:

- M0: MOS $\sim 1$
- M1: MOS $\sim 1 + (1|\text{seq})$
- M2: MOS $\sim 1 + (1|\text{seq/res})$
- M3: MOS $\sim 1 + \text{br} + (1 + \text{br}|\text{seq/res})$
- M4: MOS $\sim 1 + \text{br} + \text{inch} + \text{dist} + (1 + \text{br}|\text{seq/res})$
- M5: MOS $\sim 1 + \log(\text{br}) + \text{inch} + \text{dist} + (1 + \log(\text{br})|\text{seq/res})$
- M6: MOS $\sim 1 + \log(\text{br}) + \log(\text{inch}) + \text{dist} + (1 + \log(\text{br})|\text{seq/res})$

Table 12 describes the goodness of fit for each model in terms of AIC, BIC, and logLik. The smaller AIC or BIC is, or the larger logLik is, the higher the goodness of fit.

Among them, we selected M4 in Section V, though the goodness of fit for M5 and M6 were superior to that of M4. The reason for this was that M4 is simple, with effortlessly comprehended regression coefficients, and adequate accuracy, as displayed in Fig. 7.

**TABLE 13.** Fixed and random effects of $\beta_0$ in (6).

| (Fixed effect) | (Random effect) | | | |
|---|---|---|---|---|
| | seq | | res:seq | |
| 3.126 | a07 | -0.853 | 2K:a07 | 0.096 |
| | | | 4K:a07 | 0.177 |
| | | | 6K:a07 | -0.116 |
| | | | 8K:a07 | -0.153 |
| | a08 | 0.636 | 2K:a08 | -0.139 |
| | | | 4K:a08 | -0.018 |
| | | | 6K:a08 | 0.077 |
| | | | 8K:a08 | 0.016 |
| | a11 | 1.038 | 2K:a11 | -0.167 |
| | | | 4K:a11 | 0.087 |
| | | | 6K:a11 | 0.113 |
| | | | 8K:a11 | 0.067 |
| | b07 | -0.820 | 2K:b07 | -0.066 |
| | | | 4K:b07 | -0.072 |
| | | | 6K:b07 | -0.060 |
| | | | 8K:b07 | 0.158 |

**TABLE 14.** Fixed and random effects of $\beta_1$ in (6).

| (Fixed effect) | (Random effect) | | | |
|---|---|---|---|---|
| | seq | | res:seq | |
| 0.035 | a07 | 0.008 | 2K:a07 | 0.004 |
| | | | 4K:a07 | 0.009 |
| | | | 6K:a07 | -0.014 |
| | | | 8K:a07 | -0.019 |
| | a08 | -0.006 | 2K:a08 | 0.053 |
| | | | 4K:a08 | 0.027 |
| | | | 6K:a08 | -0.010 |
| | | | 8K:a08 | -0.014 |
| | a11 | -0.010 | 2K:a11 | 0.028 |
| | | | 4K:a11 | -0.023 |
| | | | 6K:a11 | -0.026 |
| | | | 8K:a11 | -0.024 |
| | b07 | 0.008 | 2K:b07 | 0.013 |
| | | | 4K:b07 | 0.013 |
| | | | 6K:b07 | -0.002 |
| | | | 8K:b07 | -0.014 |

## APPENDIX B
## FIXED AND RANDOM EFFECTS

In this Appendix, we present the fixed and random effects of $\beta_0$ and $\beta_1$ in (6) in Tables 13 and 14, respectively.

## REFERENCES

[1] *Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange*, document ITU-R BT.2020-2, Recommendation, Oct. 2015.
[2] Y. Narita, S. Hara, and A. Hanada, "4K/8K UHDTV satellite broadcasting: Advanced technologies and services in Japan," in *Proc. NAB Broadcast Eng. Inf. Technol. Conf.*, Apr. 2019, pp. 57–64.
[3] *High Efficiency Video Coding*, document ITU-T H.265 (V7), Recommendation, Nov. 2019.
[4] *Versatile Video Coding*, document ITU-T H.266, Recommendation, Aug. 2020.
[5] C. Bonnineau, W. Hamidouche, J. Fournier, N. Sidaty, J.-F. Travers, and O. Deforges, "Perceptual quality assessment of HEVC and VVC standards for 8K video," *IEEE Trans. Broadcast.*, vol. 68, no. 1, pp. 246–253, Mar. 2022.

[6] K. Brunnström *et al.*, "Qualinet white paper on definitions of quality of experience," Eur. Netw. Qual. Exper. Multimedia Syst. Services (QUALINET), Lausanne, Switzerland, Tech. Rep., Version 1.2, Mar. 2013.
[7] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K*, document ITU-T P.1204, Recommendation, Jan. 2020.
[8] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Video Quality Estimation Module*, document ITU-T P.1203.1, Recommendation, Jan. 2019.
[9] S. Fremerey, S. Göring, R. R. R. Rao, R. Huang, and A. Raake, "Subjective test dataset and meta-data-based models for 360° streaming video quality," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.
[10] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," *Proc. SPIE*, vol. 9394, pp. 27–37, Feb. 2015.
[11] H. Amirpour, R. Schatz, C. Timmerer, and M. Ghanbari, "On the impact of viewing distance on perceived video quality," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.
[12] J. Lin, N. Birkbeck, and B. Adsumilli, "Translation of perceived video quality across displays," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.
[13] *Methodologies for the Subjective Assessment of the Quality of Television Images*, document ITU-R BT.500-14, Recommendation, Oct. 2019.
[14] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.
[15] A. V. Katsenou, F. Zhang, K. Swanson, M. Afonso, J. Sole, and D. R. Bull, "VMAF-based bitrate ladder estimation for adaptive streaming," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.
[16] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "VVenC: An open and optimized VVC encoder implementation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2021, pp. 1–2.
[17] A. Ichigaya and Y. Nishida, "Required bit rates analysis for a new broadcasting service using HEVC/H.265," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 417–425, Jun. 2016.
[18] *Specifications of PLUGE Test Signals and Alignment Procedures for Setting of Brightness and Contrast of Displays*, document ITU-R BT.814-4, Recommendation, Jul. 2018.
[19] *Reference Electro-Optical Transfer Function for Flat Panel Displays Used in HDTV Studio Production*, document ITU-R BT.1886-0, Recommendation, Mar. 2011.
[20] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*, document ITU-T P.913, Recommendation, Jun. 2021.
[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst., Comput.*, Mar. 2003, pp. 1398–1402.
[23] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (Jun. 2016). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[24] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, pp. 1–18, Dec. 2015.
[25] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T P.1401, Recommendation, Jan. 2020.
[26] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.
[27] Y. Sugito and Y. Kusakabe, "A comparison of non-experts and experts using DSIS method," in *Proc. Electron. Imag. Symp. (EI)*, Jan. 2022, pp. 1–6.
[28] Y. Sugito and M. Bertalmio, "Non-experts or experts? Statistical analyses of MOS using DSIS method," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2732–2736.
[29] M. H. Pinson, "Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality," NTIA, Washington, DC, USA, NTIA Rep., 21-550, Oct. 2020.
[30] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. De Cock. (Oct. 2018). *VMAF: The Journey Continues*. [Online]. Available: https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12
[31] S. W. Raudenbush and A. S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, vol. 1. Thousand Oaks, CA, USA: SAGE, 2002.
[32] A. van Kasteren, K. Brunnström, J. Hedlund, and C. Snijders, "Quality of experience of 360 video—Subjective and eye-tracking assessment of encoding and freezing distortions," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 9771–9802, Mar. 2022.
[33] R. Schatz, A. Zabrovskiy, and C. Timmerer, "Tile-based streaming of 8K omnidirectional video: Subjective and objective QoE evaluation," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.
[34] K. Masaoka, M. Emoto, M. Sugawara, and Y. Nojiri, "Contrast effect in evaluating the sense of presence for wide displays," *J. Soc. Inf. Display*, vol. 14, no. 9, pp. 785–791, 2006.

**YASUKO SUGITO** (Member, IEEE) received the M.E. degree in computer engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2004. She is currently a Principal Research Engineer at Japan Broadcasting Corporation (NHK) Science and Technology Research Laboratories (STRL), Tokyo, researching video compression algorithms and image processing on 8K. Since 2010, she has been working at NHK STRL. She was a Visiting Researcher with Universitat Pompeu Fabra (UPF), Barcelona, Spain, from 2016 to 2017. Her current research interests include image quality assessment, both subjectively and objectively, for 8K videos with high-frame-rate (HFR) 120-Hz, high dynamic range (HDR), and wide color gamut (WCG).

**YUICHI KONDO** received the M.S. and Ph.D. degrees from The University of Tokyo, Tokyo, Japan, in 2016 and 2019, respectively. Since 2019, he has been working at Japan Broadcasting Corporation (NHK), Tokyo, in 2019, and is engaged in the research of video coding for ultrahigh definition television and immersive media.

**DAICHI ARAI** received the B.E. and M.E. degrees from the Tokyo Institute of Technology, Tokyo, Japan, in 2015 and 2017, respectively. He joined Japan Broadcasting Corporation (NHK), in 2017. Since 2020, he has been working at the NHK Science and Technology Research Laboratories (STRL). He is currently engaged in research on artificial intelligence applied to video compression.

**YUICHI KUSAKABE** received the M.S. degree in applied physics from The University of Tokyo, Tokyo, Japan, in 1999. Since 1999, he has been working for Japan Broadcasting Corporation (NHK) and is engaged in the research on video systems, displays, and coding systems for ultrahigh definition television and immersive media. He has been working on the standardization activity of video systems, such as HDR in international telecommunication union-radiocommunication (ITU-R) WP6C and the association of radio industries and businesses (ARIB). He is currently a Research Producer at NHK Science and Technology Research Laboratories.