https://doi.org/10.2352/EI.2022.34.11.HVEI-131

A Comparison of Non-Experts and Experts using DSIS Method

Yasuko Sugito and Yuichi Kusakabe; NHK; Tokyo, Japan

Abstract

Subjective evaluations are necessary to learn how expected viewers perceive the quality of a system. Traditionally, non-expert subjective tests are preferred rather than expert tests. In this study, we conducted subjective evaluation experiments for non-experts and experts on compressed 8K videos using the double stimulus impairment scale (DSIS) method and analyzed the experimental results expressed in terms of the mean opinion score (MOS), which is the average of individual scores. Furthermore, we investigated the differences between non-experts and experts by considering a new method in P.913 that estimates an improved MOS and a new experimental method using experts, called expert viewing protocol (EVP). Our contribution shows advantages of conducting expert subjective tests, such as EVP: expert tests allow to perform experiments with fewer subjects, to distinguish between original and distorted images, to determine a lower threshold for the image quality, to distribute scores in an appropriate range, and to constantly gain MOS values equal to improved MOS values.

Introduction

Subjective evaluations are essential in verifying how expected observers perceive the quality of a system under consideration. There are some established evaluation methods prescribed in international standards. For compressed images, the double stimulus impairment scale (DSIS) method described in Recommendation (Rec.) ITU-R BT.500 [1] is frequently used to assess a distorted test image relative to the uncompressed original image. The DSIS method is also called the degradation category rating (DCR) method as described in Rec. ITU-T P.913 [2]. This method presents a test image following its corresponding reference image, and each subject evaluates the deterioration level of the test image relative to the reference image using a five-Likert scale: 5, imperceptible; 4, perceptible, but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying. Then, a subjective result of the DSIS method is expressed as the mean opinion score (MOS), the average of individual scores for each test image indicated from 1–5 on a continuous scale, after a screening of subjects. From BT.500, the number of subjects is at least fifteen.

For selecting observers, the latest version of the recommendation, BT.500-14, allows choosing non-expert or expert viewers depending on the objectives of the assessment. However, the traditional subjective evaluation experiments with non-experts are preferable as opposed to experts (e.g., BT.500-12 published in September 2009). Previously [3], we statistically analyzed MOS values in three subjective evaluation results using the DSIS method (two non-expert and one expert datasets), investigated the differences in MOS values between the non-expert and expert subjects, and showed the benefits of conducting tests by experts. However, no direct comparison of the differences in MOS values for the same stimuli was conducted because all datasets were assessed by either non-experts or experts.

The last couple of years, some descriptions have been added to standards for subjective evaluations. BT.500-14 [1] Annex 8 to Part 2 defines the expert viewing protocol (EVP) for conducting assessments faster than the DSIS method that leverages at least nine experts. Nevertheless, the effectiveness of conducting experiments by experts has not yet been sufficiently discussed. P.913 [2] section 12.6 explains a calculation method to estimate an improved MOS using the maximum likelihood estimation method [4]. Although proposed as an alternative to existing screening methods including BT.500 and P.913, the relationship to a specific screening procedure described in BT.500-14 section A7-5.3 has not been considered.

Here, we conducted subjective assessments on 8K videos for the expert and non-expert observers, analyzed the differences in the MOS values as in our previous study [3], and considered the usefulness of the newly added methods.

8K Subjective Evaluation Experiments

We conducted the subjective evaluation experiments for the non-expert and expert viewers using 8K high dynamic range (HDR) sequences.

Test Sequences

Six 8K 119.88-Hz (120-Hz) HDR sequences were selected for this study. Experts first evaluated twelve sequences, then six that showed perceptually detectable deterioration even at a higher bit rate were used for non-experts. Three of the six sequences were sports contents (two for swimming and one for athletics), one contained contemporary dance, and the thumbnails of the other two sequences are shown in Fig. 1.

Figure 1. *Thumbnails of two of the 8K 120-Hz HDR sequences.*

The sequences were shot by 8K Hybrid Log-Gamma (HLG) [5] cameras and slightly compressed by a mezzanine codec for production. HLG is an HDR standard adopted for 8K broadcasting. Hereafter, HDR will refer to the HLG standard. Each sequence was originally 8 s long (960 frames), and 6 s (720 frames) were clipped after the encoding process described in the next subsection, and the statistics of the sequences indicate the results for that 6 s.

The left of Fig. 2 represents the mean spatial information (SI) and temporal information (TI) [1] for the six sequences, cor-

responding to the spatial and temporal complexities, respectively. The right of Fig. 2 shows the maximum colorfulness (CF) and dynamic range (DR). Each characteristic was calculated for every frame using 10-bit precision, and the mean or maximum value of whole frames was plotted on the graphs. The definition of DR is shown in eq. (1).

$$\mathrm{DR} = \log 10(\mathrm{L_{max}}/\mathrm{L_{min}}) \tag{1}$$

where $\mathrm{L_{max}}$ and $\mathrm{L_{min}}$ are the maximum and minimum luminance after excluding 1% of the brightest and darkest pixels, respectively. For calculating the absolute luminance, the peak luminance denoted by $\mathrm{L_W}$ in Table 5 of Rec. ITU-R BT.2100 [5] was set as 1,000 cd/m^2, which was equal to that of the monitor used for the subjective evaluations. Eq. (2) describes the formula of CF [6].

$$\mathrm{CF} = \sqrt{\sigma_{\mathrm{rg}}^2 + \sigma_{\mathrm{yb}}^2} + 0.3\sqrt{\mu_{\mathrm{rg}}^2 + \mu_{\mathrm{yb}}^2} \tag{2}$$

where $\mathrm{rg} = \mathrm{R} - \mathrm{G}$, $\mathrm{yb} = 1/2(\mathrm{R}+\mathrm{G}) - \mathrm{B}$, and σ^2 and μ are the variance and average of the pixel values, respectively. Although originally defined on the BT.709 RGB color space [7], we directly applied the formula to the 8K sequences that used the BT.2020 RGB color space [8], representing more vivid colors than that of BT.709. From Fig. 2, the characteristics of the selected six sequences were widely spread.

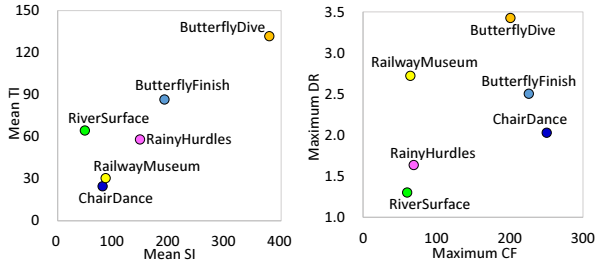


Figure 2. *Video characteristics for six test sequences.*

Video Encoding

An 8K 120-Hz real-time encoder [9] was used for the compression. The encoder complied with an 8K broadcasting standard [10]: the video coding scheme was high efficiency video coding (HEVC)/H.265 [11] Main 10 Profile (4:2:0/10 bits), intra pictures were inserted every 64 frames (approximately every 0.5 s), and 120-Hz video streams could be played as 59.94-Hz (60-Hz) videos by decoding every other frame of 120-Hz videos, (i.e., a temporal scalable coding).

Five wide-range bit rates of 64, 85, 110, 165, and 240 Mbps were set for the 8K 120-Hz videos. These bit rates are the minimum bit rate setting of the encoder, bit rate for 8K 60-Hz videos in an 8K satellite transmission experiment [12], maximum bit rate for 8K 60-Hz videos defined by the 8K broadcasting guideline (ARIB TR-B39 ver.2.5), upper bit rate for 8K 120-Hz videos in the 8K broadcasting standard [10], and bit rate for a high-quality transmission intended for public viewing [9], respectively. The 8K 120-Hz encoder outputs two streams, an 8K 60-Hz video stream and a differential stream between 8K 120- and 60-Hz videos, and automatically allocates bit amounts for them keeping the total amount to be equal to the bit rate setting. On average, a bit rate of a differential stream is less than 10% of that for the corresponding 8K 60-Hz video stream [13].

Experimental Setups

An 8K uncompressed recorder and liquid crystal display (LCD) monitor that supports the HLG format were equipped for the experiments. 8K 120- and 60-Hz HDR videos were stored in the recorder. Table 1 describes the specifications of the 8K monitor. The monitor was a prototype combined with an LCD panel equivalent to a consumer-grade 70-inch 8K TV and a high-precision backlight technology for HDR videos. Thus, it is suitable for simulating a home-use viewing condition.

Table 1. specifications of the 8K monitor

Size	70-inch diagonal (1.55 m wide and 0.87 m high)
Pixel count	7,680×4,320
Frame frequency	120; 60 Hz
Bit depth	10 bits
Peak luminance	1,000 cd/m^2
Color coverage	77% of BT.2020 [8]

Before the experiments, we measured the transition time of the monitor while displaying an 8K 120-Hz video with black and white frames. The time between black and white was less than 8.34 ms (= 1/119.88 Hz) for both black-white and white-black. Therefore, we confirmed that the monitor can accurately display 120-Hz videos. In addition, the luminance and contrast of the monitor were adjusted using the PLUGE signal [14] so that the peak luminance was set to 1,000 cd/m^2.

The experiments were conducted with reference to BT.500 [1] and were evaluated using the DSIS method, Variant I. An original video (6 s), mid-gray at approximately 50 cd/m^2 (3 s), the corresponding video to be evaluated (6 s), and mid-gray with "VOTE" (5 s) were presented. Subjects graded using a five-Likert score (5, imperceptible; 4, perceptible, but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying) from the beginning of the evaluation video to the end of the display of "VOTE." The determined duration of the videos was 6 s from a previous study on the optimal presentation duration for subjective evaluations [15].

The viewing conditions were compliant with Table 3 of BT.2100 [5], the viewing distance was set to 0.75 times the picture height (approximately 0.65 m), and the luminance of the surrounding was 5 cd/m^2. We prepared two viewing points, the left and right in front of the monitor, in the same manner as other 8K subjective evaluations [16] because its viewing field was designed as a small area of a screen to provide an immersive experience.

The experiments were conducted in two batches. The first batch from December 2020 to February 2021 was with fifteen video experts who are familiar with 8K videos for research purposes. The second batch in March 2021 was with twenty-four non-experts with a visual acuity of at least 20/20 and a normal color vision.

At first, an instruction for the assessment was given to each viewer, showing a video for a test session. We suggested that viewers should mainly evaluate a part in front of them. The test session including the highest and lowest quality of the 8K 120-Hz videos that are different from the six test sequences was held before formal sessions. Then, two sessions were conducted with

different frame rates (120 or 60 Hz). Between the sessions, evaluators took a break of at least 20 min. The evaluation was conducted individually. For experts, each session was divided into two parts, and they changed their viewing positions after the first part; however, non-experts assessed videos from a fixed position. Considering the order effect, the orders of the sessions, videos, and positions were altered depending on the evaluators. The same 66 videos ((5 bit rates $\times$ 6 sequences + 3 original sequences) $\times$ 2 frame-rates) were evaluated by non-experts and experts.

Experimental Results

First, we calculated MOS values for non-expert and expert observers after a screening process and similarly analyzed these values as our previous study [3] to demonstrate differences between non-experts and experts.

Screening of Subjects

We applied a screening method described in BT.500-14 [1] section A7-5.3. Since the screening procedure was originally proposed for the subjective assessment of multimedia video quality (SAMVIQ) method, it is called the SAMVIQ screening. Traditionally, BT.500 defines another screening procedure using Kurtosis coefficient, as described in BT.500-14 section A1-2.3, it was not applied in this study because of the note: *use of the procedure should be restricted to cases in which there are relatively few observers (e.g. fewer than 20), all of whom are non-experts.*

Here, we describe how to calculate the SAMVIQ screening. For each evaluation item j, we computed the mean score of all observers, $x_j = \sum_{i=1}^{N} y_{ij}/N$, where N is the number of observers and y_{ij} denotes an individual score for observer i on item j. Then, for each observer i, we calculated the Pearson linear correlation coefficient (PLCC) p_i and the Spearman rank order correlation coefficient (SRCC) s_i between x_j and y_{ij} for all items, respectively. Finally, we derived the rejection threshold (RT) using $r_i = \min(p_i, s_i)$. If the max correlation threshold (MCT) is less than $[\mathrm{mean}(r_i) - \mathrm{SD}(r_i)]$, where SD signifies the standard deviation, then RT = MCT and MCT = 0.7 for the DSIS method, else RT $= [\mathrm{mean}(r_i) - \mathrm{SD}(r_i)]$. If $r_i \leq$ RT, evaluation results of the subject i are discarded.

We conducted the SAMVIQ screening for 24 non-experts and 15 experts, respectively. Fig. 3 shows PLCC p_i and SRCC s_i on the horizontal and vertical axes, respectively, for all 39 evaluators. Non-experts and experts were plotted in orange and red, respectively. Note that p_i and s_i for non-expert and expert viewers were separately computed. The rejection threshold $\mathrm{RT} = [\mathrm{mean}(r_i) - \mathrm{SD}(r_i)]$ was $0.432 - 0.147 = 0.285$ for non-experts and $0.720 - 0.133 = 0.588$ for experts, and the three non-experts and two experts represented by triangles in Fig. 3 were discarded.

Assessment Results

We calculated MOS values for non-experts (represented as NE in equations) and experts (represented as EX in equations) from 21 and 13 viewers, respectively. The MOS values for the 8K 120-Hz and 60-Hz are denoted in Fig. 4. The horizontal axis holds sequence names and bit rate settings for 8K 120-Hz videos in Mbps or the original video described as "Org." Non-experts' and experts' MOS values (left-hand vertical axis) are respectively shown in orange squares and red diamonds, and pairs of black-bordered plots denote the statistical differences on Welch's t-test at a 5% significance level. The error bars denote 95% confidence

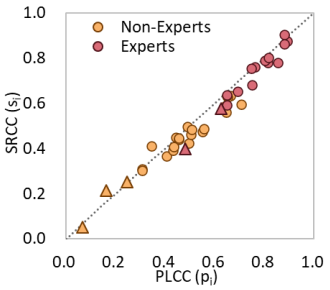
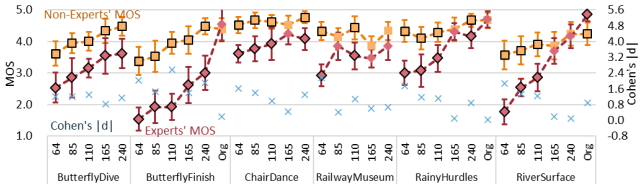


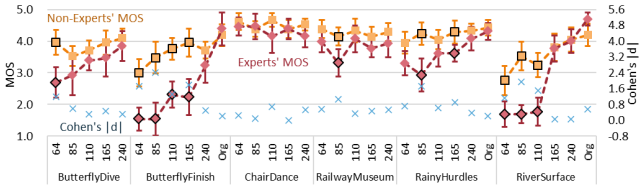
Figure 3. *Correlation coefficients of 39 subjects.*

interval (CI), $\mathrm{CI}(\mathrm{MOS}) = \mathrm{MOS} \pm 1.96 \times \sqrt{s^2/N}$, where s^2 is the unbiased variance of the individual scores and N denotes the number of subjects. The blue cross marks (right-hand vertical axis) show the absolute values of effect sizes (ESs), which report the magnitude of the difference between MOS values. Cohen's d is defined by eq. (3) [17].

$$d = \frac{\mathrm{MOS}_{NE} - \mathrm{MOS}_{EX}}{\sqrt{\frac{(N_{NE}-1)SD_{NE}^2 + (N_{EX}-1)SD_{EX}^2}{N_{NE}+N_{EX}-2}}} \tag{3}$$



(a) Results for 8K 120-Hz videos

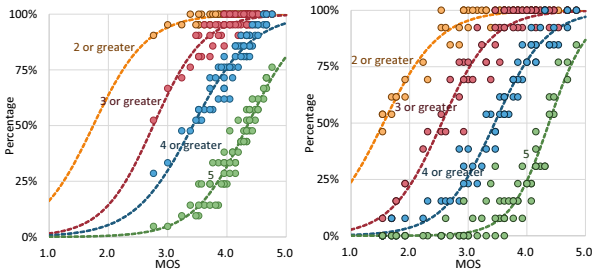


(b) Results for 8K 60-Hz videos

Figure 4. *MOS and ES for 8K 120- and 60-Hz videos.*

Score Distribution per MOS

Fig. 5 shows the relationship between the MOS values (horizontal axis) and the percentages of scores (vertical axis), plotted as circles.



(a) Non-experts (N=21) (b) Experts (N=13)

Figure 5. *Score distribution per MOS values.*

From the left to right, the circles in orange, red, blue, and green 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:

$$\hat{y}_X = \frac{1}{1+\exp(-a_X(x-b_X))} \tag{4}$$

where x and $\hat{y}_X$ denote a MOS value and a predicted proportion of scores X or greater, respectively. The true proportion y_X corresponding to x exists and is plotted as a circle in the graphs. The variables a_X and b_X are selected to minimize $\sum_{\text{all conditions i}}(y_{Xi}-\hat{y}_{Xi})^2$. When the variance of the scores for each MOS value is always at its minimum, a_X, which determines the distribution width of the scores, should be 4, and b_X, which indicates the MOS value that results in $\hat{y}_X = 0.5$, should be $X-0.5$ (see [3]). The specific values of the variables are shown in Table 2, and the figures in bold show that the value is closer to that of the lowest score variance case, i.e., $a_X = 4$ and $b_X = X-0.5$. Additionally, the coefficient of determination R^2, which measures the goodness of fit (the closer to 1, the better fitting), for each logistic function $\hat{y}_X$ can be seen in Table 3. Eq. (5) shows the definition of R^2 where $\bar{y}_X$ is the mean of y_{Xi}.

$$R^2 = \frac{\sum_i(\hat{y}_{Xi}-\bar{y}_X)^2}{\sum_i(y_{Xi}-\bar{y}_X)^2} \tag{5}$$

Table 2. Variables of the logistic functions

	a_2	a_3	a_4	a_5
Non-experts	**2.20**	**2.36**	2.01	2.22
Experts	2.10	2.31	**2.27**	**3.03**
	b_2	b_3	b_4	b_5
Non-experts	1.75	2.76	3.43	4.35
Experts	**1.55**	**2.55**	**3.45**	**4.38**

Table 3. Coefficient of determination for each logistic function

	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_5$
Non-experts	0.530	0.809	0.874	0.927
Experts	0.921	0.914	0.956	0.922

Score Variance per MOS

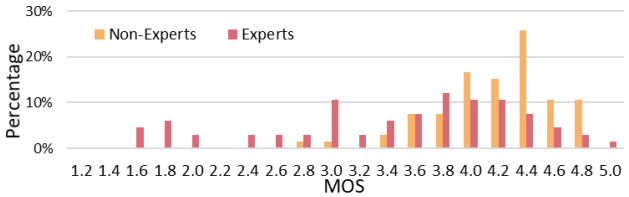
Fig. 6 shows the distribution of MOS values and the unbiased variance of individual scores for each 0.2 range of the MOS values, e.g., 1.6 on the horizontal axis is equivalent to MOS values between 1.4 and 1.6. In Fig. 6 (a), bars in orange and red denote the proportion of the MOS value range for non-expert and expert viewers, respectively. In Fig. 6 (b), plots in orange and red, respectively, represent the unbiased variance s^2 of non-experts and experts, error bars show 95% CI of the population variance σ^2 calculated using eq. (6) for sample size n, a pair of black-bordered plots denote the statistical difference in σ^2 at 5% significance level, and blue cross marks (right-hand vertical axis) show ESs $\eta_p^2 = r^2$ defined by eq. (7) [17].

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{6}$$

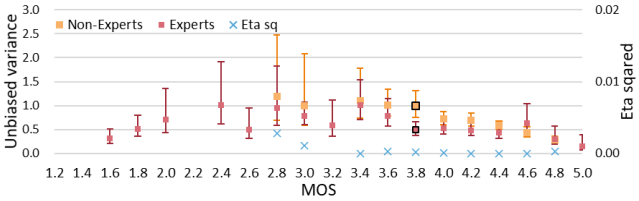
where χ^2_α is the upper $100\times\alpha$-th percentile of the chi-square distribution with the degree of freedom $n-1$.

$$r = \frac{d}{\sqrt{d^2+\frac{N^2-2N}{n_{NE}n_{EX}}}} \tag{7}$$

where d can be calculated from eq. (3) and $N = n_{NE}+n_{EX}$.



(a) Distribution of MOS values



(b) Unbiased variance and ES for each range of MOS values

Figure 6. *Analysis of score distribution for each 0.2 range of MOS values.*

Discussion

This section discusses the experimental results.

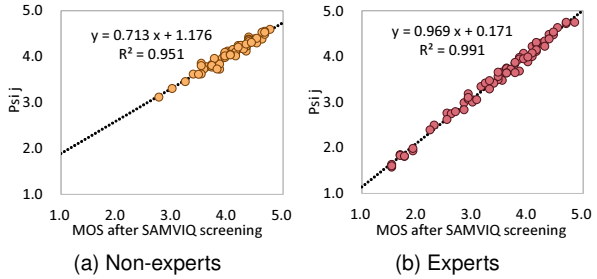
Relationship between the New P.913 Method and the SAMVIQ Screening

In June 2021, a new method to estimate an improved MOS under challenging test conditions using the maximum likelihood estimation method [4] was added to P.913 [2] section 12.6. In this method, U_{ij}, representing an individual score of subject i for the evaluation item j, is modeled as eq. (8).

$$U_{ij} = \psi_j + \Delta_i + v_i X \tag{8}$$

where ψ_j is a true quality score, called an improved MOS, of the item j, Δ_i is the bias of the subject i and $\sum_i \Delta_i = 0$, $v_i > 0$ is the inconsistency of the subject i, and $X \sim N(0,1)$ is i.i.d. Since this method is proposed as an alternative to traditional screening procedures using Kurtosis coefficient in BT.500 [1] and the bias of observers in P.913 [2], we calculated ψ_j for non-expert and expert viewers from the experimental results of all participants, i.e., 24 non-experts and 15 experts. For the calculations, we used a reference Python implementation introduced in Appendix III of P.913.

Next, we investigated the relationship between a MOS after the SAMVIQ screening MOS_j (horizontal axis) and an improved MOS ψ_j (vertical axis) for non-experts and experts, as shown in Fig. 7. The dashed line in each graph denotes a fitted linear function. We describe the specific function and the coefficient of determination R^2 calculated from eq. (5).



(a) Non-experts (b) Experts

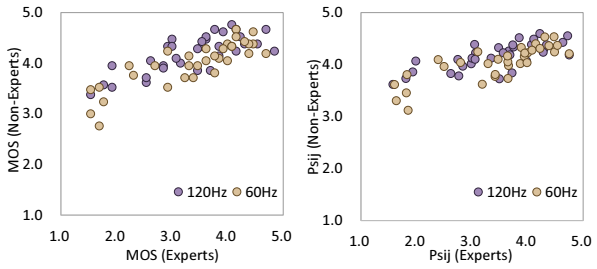
Figure 7. *Relationship between MOS and improved MOS.*

Results show that improved MOS values for the experts are almost equal to the corresponding MOS values after the SAMVIQ screening since the fitted function is similar to $y = x$, whereas that is untrue for non-experts.

Furthermore, we also conducted the same analyses on three datasets (two non-experts, namely, NE1 [18] and NE2 [19] with 24 and 14 participants, respectively, and one expert, namely, EE [20] with 16 participants) used previously [3]. For NE2, we calculated scores in the five-Likert scale, such that $\lfloor(\text{OriginalScore}-1)/20\rfloor+1$ where $1 \leq \text{OriginalScore} \leq 100$. One subject who did not assess all evaluation items was excluded. We applied the SAMVIQ screening procedure to each dataset and confirmed that the rejection threshold $\text{RT} = \text{MCT} = 0.7$ (i.e., $[\text{mean}(\text{r}_\text{i}) - \text{SD}(\text{r}_\text{i})] > \text{MCT}$) and no subjects were rejected in all datasets. This can be attributed to the difficulty level of the assessments. Since HDR still images were assessed in these subjective evaluations, the difficulty of evaluations was much lower than that of videos, and the variance of individual scores may have become small even in non-experts. Then, we calculated improved MOS values ψ_j and MOS_j and derived a fitted linear function for each dataset. All three fitted functions resulted in $y \sim x$.

From all analyses, if the $\text{mean}(\text{r}_\text{i})$ after the SAMVIQ screening is sufficiently large (empirically, 0.75 or greater), then $\psi_\text{j} \sim \text{MOS}_\text{j}$. Furthermore, experts constantly mark large $\text{mean}(\text{r}_\text{i})$, and do not require the new method of P.913.

Figures 8 (a) and (b) describe the relationship between the scores of non-expert (vertical axis) and expert (horizontal axis) subjects for MOS_j and ψ_j, respectively. Since the score differences between non-experts and experts in 120-Hz videos were more extensive than those in 60-Hz videos as shown in Fig. 4, we used the color-coded plots for each framerate for better visibility. Given $\psi_\text{j} \sim \text{MOS}_\text{j}$ for experts, the variance of ψ_j in non-experts is smaller than that of MOS_j, and that is an advantage of the new P.913 method.



(a) MOS after the SAMVIQ screening (b) improved MOS (Ψ_j)

Figure 8. *Relationship between scores of non-experts and experts.*

Differences between Non-Experts and Experts

Overall, the experts tended to grade lower scores for compressed videos and higher scores for original videos than the non-experts, as shown in MOS values in Fig. 4. Comparing the results of 120 and 60 Hz, 24 of 33 conditions of 120-Hz videos showed statistical differences, whereas 11 of 33 conditions of 60-Hz videos did. This can be attributed to two reasons. First, the non-experts were not familiar with high-frame-rate (HFR) videos, such as the 120-Hz videos, thereby having difficulty finding deterioration. Second, the experts who were familiar with HFR videos felt annoyed with HFR noise flickers in dark parts, especially for the ChairDance and RailwayMuseum sequences.

For the fitting curves in Fig. 5, expert variables were generally close to the lowest score variance case than the non-experts except for a_2 and a_3, as shown in Table 2. From Fig. 5 (a) and Table 3, the goodness of fit for $\hat{y}_2$ and $\hat{y}_3$ of the non-experts was too small, $R^2 << 0.9$, due to lack of low MOS values less than 3. Thus, the variance of experts was smaller than that of the non-experts as a general trend, as seen in Fig. 6 (b).

Previously [3], we confirmed the advantages of conducting subjective evaluation experiments with experts including:

1. Experts are useful to determine the lower threshold of image quality.
2. Experts better distinguish the differences between original and compressed images.
3. Experts help conduct experiments with fewer subjects and still see a general trend.

Items 1 and 2 were again confirmed by the experimental results in Fig. 4. Regarding Item 2, the variance in $\text{MOS} \sim 5$ of the non-experts was smaller than that of the experts as with our previous study, meaning that the non-expert subjects were unable to detect a subtle difference from the original image and they had a tendency to grade 5. Item 3 was rediscovered from Fig. 3: the correlations of the non-experts are relatively low and widely spread, while those of the experts are consistently high. Also, the smallness of the variance in experts' MOS in Figs. 5 and 6 (b) are useful for conducting tests with fewer observers. Moreover, from Fig. 6 (a), the MOS range for experts was much wider than that of non-experts especially in MOS values less than 3, showing the capability of experts to adapt the quality of test images to scores 1–5.

In October 2019, EVP, which allows conducting a quick subjective evaluation with fewer experts, was added to BT.500-14 [1] Annex 8 to Part2. In this protocol, two test images A and B are presented after the corresponding reference image. Then, each evaluator scores both A and B using an eleven-grade scale. That is, a skill to immediately evaluate images and grade scores with an appropriate scope are required. These findings strengthen the availability of such novel evaluating method.

So far, we have discussed the advantages of using experts for subjective assessments to dispel a traditional belief: conducting tests with non-experts is preferable rather than experts. We suggest that the type of viewers, non-experts or experts, should be selected depending on the objectives of the assessment as described in BT.500-14 [1]. For example, if we want to know the response of end-users, conducting an assessment by non-expert subjects would be desirable.

Conclusions

In this study, we conducted subjective evaluation experiments by non-experts and experts using 8K HDR videos. Also, we applied subject screening using the SAMVIQ screening, analyzed MOS values included in the meta analyses, and considered the differences between non-expert and expert viewers for the new methods adopted in BT.500 and P.913. Furthermore, we discussed the benefits of conducting experiments using experts as suggested in our previous study including determining lower threshold of the image quality, distinguishing between original and compressed images, and conducting experiments with fewer subjects and still finding a general trend. Additionally, we discovered the ability of experts to adapt the image quality into an appropriate range of scores as directly compared with MOS values for the same test videos. All these advantages justify the usefulness of EVP. Compared with MOS values after the SAMVIQ screening and improved MOS values estimated by the new P.913 method, applying the new method to subjective results of experts may be unnecessary due to the consistently high correlation between MOS values and individual scores.

For future work, we will continue to study the differences between non-expert and expert observers including subjective evaluation methods other than the DSIS method.

References

[1] Recommendation ITU-R BT.500-14, Methodologies for the subjective assessment of the quality of television images (2019).

[2] Recommendation ITU-T P.913, Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment (2021).

[3] Yasuko Sugito and Marcelo Bertalmío, Non-Experts or Experts? Statistical Analyses of MOS using DSIS Method, Proc. ICASSP 2020, pgs. 2732–2736. (2020).

[4] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis, A simple model for subject behavior in subjective experiments, J. Electron. Imaging, 11, 2020, 131–1–131–14 (2020).

[5] Recommendation ITU-R BT.2100-2, Image parameter values for high dynamic range television for use in production and international programme exchange (2018).

[6] David Hasler and Sabine E. Suesstrunk, Measuring colorfulness in natural images, Proc. HVEI VIII, pgs. 87 – 95. (2003).

[7] Recommendation ITU-R BT.709-6, Parameter values for the hdtv standards for production and international programme exchange (2015).

[8] Recommendation ITU-R BT.2020-2, Parameter values for ultra-high definition television systems for production and international programme exchange (2015).

[9] Yasuko Sugito, Shinya Iwasaki, Kazuhiro Chida, Kazuhisa Iguchi, Kikufumi Kanda, Xuying Lei, Hidenobu Miyoshi, and Yoshifumi Uehara, UHD-2/8K 120-Hz Realtime Video Codec, SMPTE Motion Imaging J., 7, 129, 41–49 (2020).

[10] ARIB STD-B32 Version 3.9-E1, Video coding, audio coding, and multiplexing specifications for digital broadcasting (2016).

[11] Recommendation ITU-T H.265 (V7), High efficiency video coding (2019).

[12] Yasuko Sugito, Kazuhisa Iguchi, Atsuro Ichigaya, Kazuhiro Chida, Shinichi Sakaida, Hiroharu Sakate, Yukinari Matsuda, Yukiyasu Kawahata, and Nobuaki Motoyama, HEVC/H.265 Codec System and Transmission Experiments aimed at 8K Broadcasting, The Best of IET and IBC 2015-2016, 7, 24–29 (2015).

[13] Shinya Iwasaki, Xuying Lei, Kazuhiro Chida, Yasuko Sugito, Kazuhisa Iguchi, Kikufumi Kanda, Hidenobu Miyoshi, and Yoshifumi Uehara, The required video bitrate for 8k120-hz real-time temporal scalable coding, Proc. ICCE 2020, pgs. 1–5. (2020).

[14] Recommendation ITU-R BT.814-4, Specifications of pluge test signals and alignment procedures for setting of brightness and contrast of displays (2018).

[15] Felix Mercer Moss, Ke Wang, Fan Zhang, Roland Baddeley, and David R. Bull, On the optimal presentation duration for subjective video quality assessment, IEEE Trans. Circuits Syst. Video Technol., 11, 26, 1977–1987 (2016).

[16] Yasuko Sugito, Shinya Iwasaki, Kazuhiro Chida, Kazuhisa Iguchi, Kikufumi Kanda, Xuying Lei, Hidenobu Miyoshi, and Kimihiko Kazui, Video bit-rate requirements for 8K 120-Hz HEVC/H.265 temporal scalable coding: experimental study based on 8K subjective evaluations, APSIPA Trans. Signal Inf. Process., 9, e5 (2020).

[17] Daniel Lakens, Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs, Front. Psychol., 4, 863 (2013).

[18] Pavel Korshunov, Philippe Hanhart, Thomas Richter, Alessandro Artusi, Rafał Mantiuk, and Touradj Ebrahimi, Subjective quality assessment database of HDR images compressed with JPEG XT, Proc. QoMEX 2015, pgs. 1–6. (2015).

[19] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux, An extensive performance evaluation of full-reference HDR image quality metrics, Quality and User Experience, 1, 2, 1–16 (2017).

[20] Yasuko Sugito and Marcelo Bertalmío, Practical use suggests a re-evaluation of HDR objective quality metrics, Proc. QoMEX 2019, pgs. 1–6. (2019).

Author Biography

Yasuko Sugito is a principal research engineer at the Japan Broadcasting Corporation (NHK) Science and Technology Research Laboratories (STRL), Tokyo, Japan, researching video compression algorithms and image processing particularly on 8K. Her current research interests focus on 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 Kusakabe is a senior research engineer at Japan Broadcasting Corporation (NHK) Science and Technology Research Laboratories (STRL), Tokyo, Japan. He is engaged in research of video system and displays for ultra-high definition system. He has been working on standardization of video systems such as HDR in ARIB and ITU-R SG6.