NON-EXPERTS OR EXPERTS? STATISTICAL ANALYSES OF MOS USING DSIS METHOD

Yasuko Sugito

NHK
Science and Technology Research Labs.
Tokyo, Japan
sugitou.y-gy@nhk.or.jp

Marcelo Bertalmío

Universitat Pompeu Fabra
Department of
Information and Communication Technologies
Barcelona, Spain

ABSTRACT

In image quality assessments, the results of subjective evaluation experiments that use the double-stimulus impairment scale (DSIS) method are often expressed in terms of the mean opinion score (MOS), which is the average score of all subjects for each test condition. Some MOS values are used to derive image quality criteria, and it has been assumed that it is preferable to perform tests with non-expert subjects rather than with experts. In this study, we analyze the results of several subjective evaluation experiments using the DSIS method. Our first contribution is to discuss the statistical meaning of the MOS values, which has not been previously addressed in the literature. Second, our results show that, contrary to the established belief, there are advantages when performing subjective tests with experts, in that they allow experiments to be performed with fewer subjects, and to better determine the lower threshold of image quality.

Index Terms— The double-stimulus impairment scale (DSIS) method, mean opinion score (MOS), non-expert, expert, statistical analysis

1. INTRODUCTION

Subjective evaluations are essential to verify how viewers will perceive the image quality of a system being tested. The double-stimulus impairment scale (DSIS) method described in Recommendation ITU-R BT.500 [1] is frequently applied to subjective quality assessments of compressed images. In this method, a test image is presented following the corresponding reference image, and each subject evaluates the deterioration level of the test image relative to the reference image using a five-grade scale (5, imperceptible; 4, perceptible, but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying). According to the recommendation, the number of subjects should be at least 15. Note that an equivalent subjective evaluation method, the degradation category rating (DCR) method, is described in Recommendation ITU-T P.910 [2].

Conventionally, the mean opinion score (MOS), which is the average score of all the subjects for each test condition, is treated as the subjective evaluation result. MOS values of the DSIS method are primarily used for two purposes: (1) measuring the subjective quality of the test images [3] and (2) measuring the performance of the objective image quality metrics [4]. For the first purpose, several criteria expressed as MOS values are traditionally used. For example, MOS=3.5 is referred to as the tolerance limit of deterioration [3]; the quality of an image is considered to be good if its MOS value is 3.5 or greater. However, such criteria are not mentioned in either BT.500 or P.910 and, to the best of our knowledge, the statistical meanings of such criteria have never been discussed. For the second purpose, the performance of the objective quality metrics can be evaluated using a correlation or an error in reference to the MOS values. Therefore, it is important to properly prepare the subjective experimental conditions. This is also true for the first purpose.

The selection of the subjects is one of the experimental conditions. It is traditionally believed that it is preferable to conduct subjective evaluation experiments with non-experts as opposed to experts, even though BT.500-13 [5], published in 2012, stated that "*Observers may be expert or non-expert depending on the objectives of the assessment*." This belief may be reflected in the general viewer condition in Rec. P.910 [2], which specifies a minimum of 15 non-experts: "*They should not be directly involved in picture quality evaluation as part of their work and should not be experienced assessors*." (BT.500-12 [6] said the same as well.) Meanwhile, P.910 also allows the use of a small number of experts (i.e., 4–8) for preliminary experiments. In other words, there has never been a sufficient discussion concerning the difference between non-experts and experts.

In this study, we discuss the statistical meanings of the MOS values and the difference between non-expert and expert subjects on the basis of the analysis results of subjective evaluation experiments using the DSIS method.

This work received partial funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 761544 (project HDR4EU) and under grant agreement number 780470 (project SAUCE) and from the Spanish government and FEDER Fund, grant ref. PGC2018-099651-B-I00 (MCIU/AEI/FEDER, UE).

2. STATISTICAL ANALYSIS OF THE MOS VALUES

In this section, we analyze the MOS values of three different experiments, in which the image quality of high dynamic range (HDR) compressed still pictures was assessed using the DSIS method.

2.1. Experimental conditions

In the non-expert experiment 1 (NE1) [7], 240 compressed HDR images (944×1,080) were assessed using a 2K (1,920× 1,080) SIM-2 monitor. The original (uncompressed) and compressed pictures were presented side by side for 10 s, and each subject scored the combination using the five-grade scale of the DSIS method. In total, 24 naïve subjects participated; however, only the results of 22 subjects were used after conducting the screening method described in BT.500.

Second, 100 HDR images (88 compressed, 6 tone-mapped and uncompressed, and 6 uncompressed 1,920× 1,080 images) were assessed using a 2K SIM-2 monitor in the non-expert experiment 2 (NE2) [8]. The original and test images were presented for 6 and 8 s, respectively, and each subject graded them using a 100-grade scale associated with the five-grade scale of the DSIS method. A total of 15 naïve subjects participated (40 out of 100 test images were assessed by 14 subjects), and the results of all 15 subjects were used, after conducting the screening method described in BT.500.

Finally, in the expert experiment (EE) [4], 260 HDR images (240 compressed and 20 hidden original 1,920×1,080 images) were evaluated by 16 experts familiar with HDR images. A 4K (4,096×2,160) HDR monitor was used for the assessment, and the presentation and grading methods were the same as in NE1. The screening was conducted in terms of the individual MOS values of the original images (4.55–5.00) and the Pearson linear correlation coefficient (PLCC) between the MOS values and individual scores for all 260 items (0.86–0.94); it was confirmed that there were no outliers.

2.2. Analysis of the score distribution per MOS value

Figure 1 shows the relationship between the MOS values (horizontal axis) and the percentages of scores (left-hand vertical axis), plotted as open circles, as well as the unbiased variance of the scores (right-hand vertical axis), plotted as x's, for each experiment. For NE2, we calculated scores in the five-grade scale, such that $\lfloor(\mathrm{OriginalScore}-1)/20\rfloor+1$ where $1 \leq \mathrm{OriginalScore} \leq 100$.

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. For each score range, the dotted line indicates the fitted curve of a logistic function of eq. (1) using the least squares method:

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

where x and $\hat{y}_X$ are a MOS value and a predicted proportion of scores X or greater, respectively. The true proportion y_X corresponding to x exists. The variables a_X and b_X are selected to minimize $\sum_{all\ conditions\ i}(y_{Xi}-\hat{y}_{Xi})^2$. The specific values of the variables are shown in Table 1.

Table 1. Variables of the logistic functions.

	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5
Non-expert 1	3.46	1.64	2.66	**2.53**	**2.76**	**3.50**	**3.63**	**4.39**
Non-expert 2	3.69	1.63	2.46	2.55	2.42	3.45	3.39	4.35
Expert	**3.98**	**1.57**	**2.94**	2.56	2.69	3.52	3.18	4.38

2.3. Analysis of the score variance per MOS value

Figure 2 illustrates the distribution of the MOS values and the unbiased variance of scores for each 0.2 range of the MOS values, e.g., 1.6 on the horizontal axis indicates the MOS values between 1.4 and 1.6. The analysis, including an F-test for the score variance, was conducted for the results of each of the 240 compressed images in NE1 and EE. In Fig. 2 (b), a pair of black-bordered bars marks the significant difference in the population variance at a 5% significance level.

2.4. Analysis of the correlation to MOS

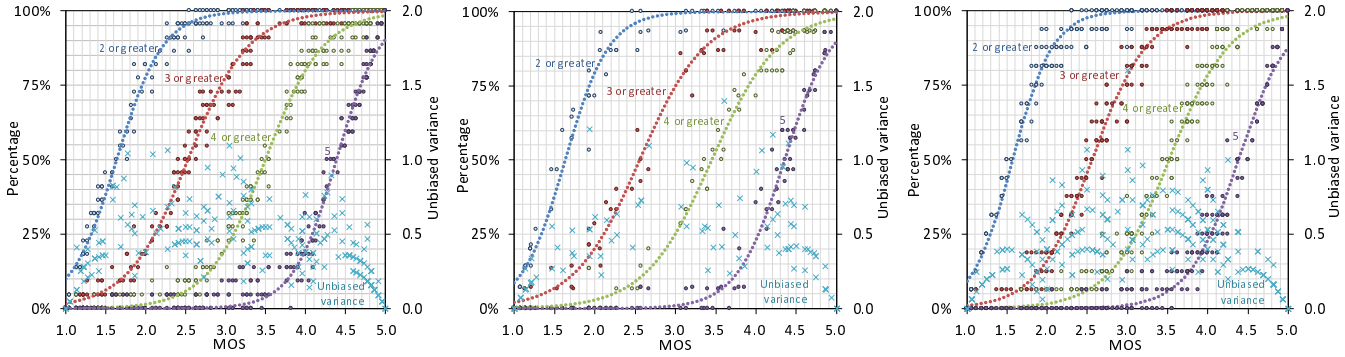
Figure 3 shows the correlation between the MOS values and the individual scores for NE1 and EE. In the same manner as in the previous section, a calculation was conducted for the results of each of the 240 compressed images. On the graph, the horizontal and vertical axes show the PLCC and the Spearman rank order correlation coefficient (SROCC), respectively. The dotted line indicates SROCC=PLCC. PLCC determines the linearity, whereas SROCC measures the monotonicity, or how much the order of the individual scores corresponds to that of the MOS values.

3. CONSIDERATIONS

In this section, we discuss the analysis results for the MOS values.

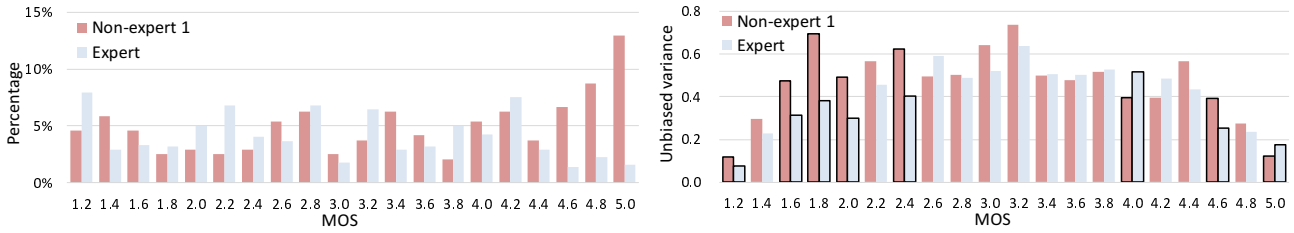
3.1. Statistical meanings of the fitted curves

In eq. (1), the variables a_X and b_X determine the distribution width of the scores (a larger a_X results in a narrower width) and the MOS value that results in $\hat{y}_X = 0.5$, respectively. If the variance of the scores for each MOS value is always at its minimum, the percentages of the scores and the unbiased variance per MOS should be as shown in Fig. 4. In such a case, the slope of the function X or greater shown by the dotted line is 1 for $X-1 \leq x \leq X$, and the proportion of scores, X or greater, becomes 0.5 when MOS is $X-0.5$. Regarding



(a) Non-expert 1 (N=22) (b) Non-expert 2 (N=14 or 15) (c) Expert (N=16)

Fig. 1. Score distribution per MOS value.



(a) Distribution of MOS values (b) Unbiased variance for each range of MOS values

Fig. 2. Analysis of the score distribution for each 0.2 range of the MOS values.

the logistic function of eq. (1), if $\hat{y}'_X = 1$ and $x = b_X$, a_X becomes 4.

In Table 1, the figures in bold indicate the variables that are the closest to the lowest score variance case, $a_X = 4$ and $b_X = X - 0.5$ for $X = 2, 3, 4$, and 5, for the three experiments. From this, it was found that the variance of EE was lower than that of NEs at lower MOS values; this can be seen for MOS values lower than 2.4 in Figs. 1 and 2 (b). Conversely, at higher MOS values, the variables of NE1 were the closest to the lowest variance case, as shown in Table 1. In addition, in Fig. 2 (b), the population variance of NE1 is significantly lower than that of EE for MOS values of 3.8–4.0 and 4.8–5.0.

3.2. Statistical meanings of the MOS values

From the considerations in the previous section and Fig. 1, the statistical meanings of the MOS values can be described as follows.

MOS=2.5, which is the intermediate value between the annoying and slightly annoying levels, is usually considered in practice as the lower level below which observers do not tolerate deterioration. At MOS=2.5, ∼95% of non-expert subjects gave scores ≥ 2 and nearly 50% of non-expert subjects gave scores ≥ 3, whereas nearly 100% of expert subjects gave scores ≥ 2 and ∼45% of expert subjects gave scores ≥ 3.

Reference [3] describes **MOS=3.0** as a lower limit for the broadcasting quality as assessed by experts. At MOS=3.0, nearly 100% of expert subjects gave scores ≥ 2, i.e., if the MOS value dips below 3.0, there is the possibility of a score of 1, which means an image at a very annoying level. Conversely, nearly 100% of non-expert subjects, but a slightly smaller percentage than that of experts, gave scores ≥ 2.

MOS=3.5, which is the intermediate value between the slightly annoying and non-annoying levels, is conventionally regarded as the tolerance limit of deterioration [3]. At MOS=3.5, more than 90% of non-expert subjects gave scores ≥ 3 and ∼50% of non-expert subjects gave scores ≥ 4, whereas ∼95% of expert subjects gave scores ≥ 3 and ∼50% of expert subjects gave scores ≥ 4. Therefore, if MOS exceeds 3.5, more than half of subjects will not consider the image as having an annoying level and nearly all of the remaining subjects will perceive the image as having a barely annoying level.

MOS=4.5, which is the intermediate value between the perceptible and imperceptible levels, is traditionally referred to as the detection limit of deterioration. At MOS=4.5, nearly 100% of both non-expert and expert subjects gave scores ≥ 3, more than 90% of subjects gave scores ≥ 4, and ∼60% of subjects gave 5. It can be said that a MOS value in which 50% of subjects gave 5 is ∼4.4 (see b_5 in Table 1).

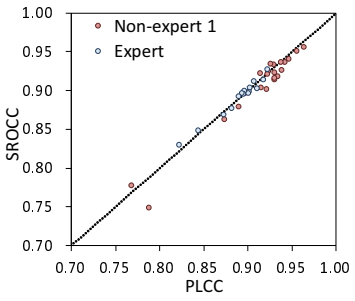


Fig. 3. Correlation to MOS values.

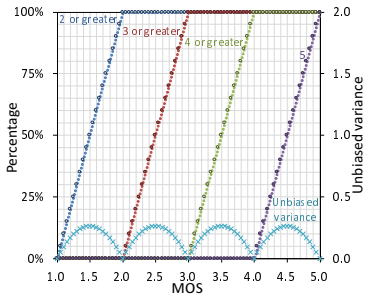


Fig. 4. Lowest score variance case (N=20).

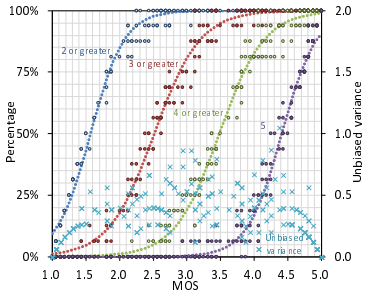


Fig. 5. Score distribution per MOS in NE1 for the 16 best PLCCs (N=16).

3.3. Difference between non-experts and experts

Here, we further consider the difference between non-expert and expert subjects on the basis of the analysis results. In NE1, 3 out of 240 compressed images were given MOS of 5.0. The reason for the small variance at MOS∼5.0 in Fig. 2 (b) might be that the subjects were not able to detect a subtle difference from the original image. In NE2, 2 out of 88 compressed images were given MOS of 5.0 (the original MOS values expressed from 1 to 100 were 95.93 and 96.71), and 2 out of 6 uncompressed images were given MOS of less than 4.5 (the original MOS values were 80.33 and 82.67). Meanwhile, in the EE, none of the compressed images were given MOS of 5.0 (maximum 4.81); however, 5 out of 20 original images were given MOS of 5.0 (minimum 4.63). Therefore, it is possible that experts can better distinguish the difference between original and compressed images.

Figure 3 shows that the correlations of the non-experts are widely spread: 0.77–0.96 in PLCC and 0.75–0.96 in SROCC, whereas those of the experts are consistently high: 0.82–0.92 in PLCC and 0.83–0.93 in SROCC. The very high correlations in NE1, such as 0.96, may be due to the high percentage of MOS values close to 5.0, as shown in Fig. 2 (a). In general, the possibility of the existence of scores of 5 sharply increases as MOS approaches 5.0, as can be seen in the purple circles close to MOS=5.0 in Fig. 1, and the correlation between the MOS values and the individual scores tends to be higher. Note that the PLCCs of the EE including the results of the original images were higher than those in the above figures, as described in Section 2.1.

In NE1, we extracted the results of 16 subjects whose PLCCs, 0.91–0.96, were better than those of the other six subjects and analyzed them in the same manner as in Section 2.2. Figure 5 and Table 2 show the percentages of the scores and the unbiased variance per MOS and the variables of the logistic functions, respectively. In the table, the figures in bold indicate values that are closer to the lowest score variance case than those in Table 1.

In Table 2, the variables a_2, b_2, and a_3 become closer to those of the EE; however, an opposite trend is observed for the other variables. Therefore, if we choose non-expert subjects whose correlation to MOS is very high, the expert trend might be predictable for lower MOS values, such as less than 2.4. However, that is not true for very high MOS values, such as approximately 4 or greater, in which the score variance of the experts can be higher than that of the non-experts.

Table 2. Variables of the logistic functions in NE1 for the 16 best PLCCs.

	a_2	b_2	a_3	b_3	a_4	b_4	a_5	b_5
NE1 best 16	**4.01**	**1.57**	**2.95**	**2.52**	**3.02**	3.48	**3.88**	**4.41**

Overall, our analysis shows that we can perform image quality tests with fewer observers if they are experts. In addition, experts are helpful to determine the lower limit of the image quality, due to their lower score variance at lower MOS values, and to detect a slight degradation from the original image.

4. CONCLUSIONS

In this study, we analyzed the results of subjective evaluation experiments using the DSIS method as assessed by non-experts and experts, showed the statistical meanings of the MOS values traditionally used as criteria of the image quality, and considered the difference between non-expert and expert subjects. The results show that, in terms of the statistical meanings of the criteria, there is a slight difference between non-experts and experts for the criteria of MOS=3.5 and lower. We also found that experts can be useful to determine the lower threshold of the image quality, to distinguish between original and compressed images, and to conduct experiments with a smaller number of subjects and still see a general trend.

In general, it is preferable that the type of subjects, non-experts or experts, should be chosen depending on the application of the system being tested as mentioned in BT.500. In addition, expert subjects can be useful for specific purposes as suggested in P.910. We will continue to analyze other experimental results using the DSIS method.

5. REFERENCES

[1] Recommendation ITU-R BT.500-14, "Methodology for the subjective assessment of the quality of television pictures," 2019.

[2] Recommendation ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[3] A. Ichigaya and Y. Nishida, "Required bit rates analysis for a new broadcasting service using hevc/h.265," *IEEE Transactions on Broadcasting*, vol. 65, no. 2, pp. 417–425, 2016.

[4] Y. Sugito and M. Bertalmío, "Practical use suggests a re-evaluation of hdr objective quality metrics," in *2019 11th International Conference on Quality of Multimedia Experience (QoMEX)*, Berlin, Germany, 2019, pp. 1–6.

[5] Recommendation ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[6] Recommendation ITU-R BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," 2009.

[7] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of hdr images compressed with jpeg xt," in *2015 7th International Workshop on Quality of Multimedia Experience (QoMEX)*, Costa Navarino, Messinia, Greece, 2015, pp. 1–6.

[8] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference hdr image quality metrics," *Quality and User Experience*, vol. 2, no. 1, pp. 1–16, 2017.