Comparison of Human Opinion Scores in Subjective Video Quality Test

Yiwen Xu[1, a], Qi Li[1,b], Xia Li[1,c] and Ying Fang[1,d*]

[1]School of Physics and Information Engineering, Fuzhou University, Fuzhou, China

[a]xu_yiwen@fzu.edu.cn, [b]n151120062@fzu.edu.cn,

[c]n151127009@fzu.edu.cn, [d]fangying@fzu.edu.cn

Keywords: Video coding, high efficiency video coding (HEVC), H.265, video quality assessment (VQA), mean opinion score (MOS)

Abstract. The new MPEG-H/H.265 High Efficiency Video Coding(HEVC) standard greatly promotes the video compression efficiency and also creates new opportunities and challenges in perceptual video coding. To measure the perceptual quality, Mean Opinion Score (MOS) is obtained by averaging human scores in subjective test. But its value cannot necessarily represent exact difference between the human opinions. To solve this problem, interval MOS (iMOS) method, which employs the distribution information of human scores to measure the difference between MOS values with deviations, is introduced. Furthermore, the feasibility of using iMOS to evaluate compressed video quality is confirmed by experiments.

Introduction

During recent years, the increasing requirements of video services have greatly promoted the development of video coding standards, among which the most recent milestone is the new MPEG-H/H.265 High Efficient Video Coding (HEVC) technique [1]. Essentially in all popular video coding standards including HEVC and H.264 [2], the lossy video coding problem is formulated as a Rate-Distortion Optimization (RDO) problem to minimize the coding distortion subject to a constraint on Bit Rate (BR). Nevertheless, the current RDO approach has been criticized for its definition of distortion, where the traditional objective distortion/quality measures such as Sum of Absolute Difference (SAD), Mean Squared Error (MSE), and Peak-Signal-to-Noise-Ratio (PSNR), have been found to be poorly correlated with perceived video quality of Human Vision System (HVS).

To address this issue, perceptual video coding techniques have been developed where subjective distortion/quality measures are used to imitate human perception of video quality. As a criterion of these subjective measures, Mean Opinion Score (MOS) or its variation like differential MOS (DMOS) is obtained though subjective test and believed to reflect the human perceived quality [3,4,5]. Despite of that, it is impossible to integrate the subjective test into real-life video codecs. As an alternate, perceptual Video Quality Assessment (VQA) technique has undergone significant development aiming to design better visual quality metrics. Generally, the VQA approaches can be categorized into two groups, namely, vision modeling approaches and engineering approaches, among which the bottom-up engineering approaches have become popular in recent years [6].

In this paper, we attempt to address three major issues regarding the comparison of MOS, which is acquired by offline subjective test. We introduce how to compare MOS values by considering the distribution information of human scores, which reflects the deviations of human perceptions on the same sequence.

The Problem of MOS Value Comparison

In subjective test, the MOS values are obtained by rating scale methods, such as ITU-R quality, impairment and comparison scales [3]. In either case, the subjects are asked to classify the stimuli into a certain number of categories where each category is labeled with a scale. For example, ITU-R five grade quality scale includes “Excellent” (5), “Good” (4), “Fair” (3), “Poor” (2) and “Bad” (1).

After subjective test, the MOS value is obtained by averaging all scores of a stimulus given by different subjects.

In level of measurement [7], the subject scores are ordinal numbers, where the numbers reflect the rank orders of all stimuli and the addition operation is not applicable. The averaging procedure actually implies an assumption that the subject scores and MOS are interval numbers, where the differences between any two consecutive scales have identical impact on human vision (e.g., the difference between "Excellent" and "Good" is identical to that between "Fair" and "Poor"); however, it is not always the case for all subjects [8]. The human opinions on the differences between scales are hidden in the distribution of human scores, by exploiting which we can achieve an interval number of MOS.

Calculating Interval MOS Values

Given a group of stimuli, we can obtain an interval MOS (iMOS) values based on all subjective scores and the law of categorical judgement [9], which is based on the following assumptions. When a series of stimuli is presented to a subject, he or she can respond differently with respect to certain qualitative or quantitative attribute. This process is called a discriminal process and the assigned attribute defines a psychological continuum [10]. If we present the stimulus a large number of times, we can make a postulate that the associated discriminal processes form a normal distribution on the psychological continuum.

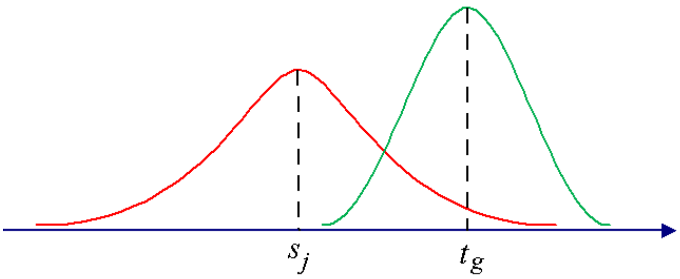


Fig. 1. Law of categorical judgement

The law of categorical judgement is described as follows. First, we divide the psychological continuum of the subject into a certain number of ordered categories, where the boundary of each category follows a normal distribution. For example, to sort n stimuli into $m+1$ categories, we assume that the iMOS and discriminal dispersion (*i.e.*, the standard deviation) of discriminal processes of stimuli $j(j=1,2,...,n)$ are s_j and σ_j respectively; and for each category boundary $g(g=1,2,...,m)$, the mean location and dispersion are t_g and σ_g, respectively. An example of s_j and t_g is shown in Fig. 1.

Second, we compare the distribution of discriminal processes of each stimulus with all boundaries and get

$$t_g - s_j = x_{jg}\sqrt{\sigma_j^2+\sigma_g^2-2r_{jg}\sigma_j\sigma_g}\ , \qquad (1)$$

where r_{jg} denotes the correlation between momentary positions of stimulus j and boundary g; x_{jg} represents the unit normal deviate corresponding to $P_{j<g}$, the probability of j is sorted below boundary g, as

$$x_{jg} = \sqrt{2}\mathrm{erf}^{-1}(2P_{j<g}-1)\ , \qquad (2)$$

where $\mathrm{erf}^{-1}(\cdot)$ represents the inverse error function. Under condition D of law of categorical judgement [9], we assume σ_j, σ_g and r_{jg} are constants for all g and j, and thus

$$t_g - s_j = cx_{jg}\ , \qquad (3)$$

where c represents a real constant; x_{jg} can be estimated with an observed value of $P_{j<g}$ and Eq. (2). Especially, in an interval scale, we can set $c = 1$ and $t_0 = 0$.

Third, by repeating Eq. (3) over all *j* and *g*, we can formulate a system of linear equations which is overdetermined with $m \times n$ - 1 variables and $m \times n$ equations. The iMOS values s_j, $(j = 1,2,...,n)$ are obtained by solving the equations.

There is another issue in this problem. Considering any two normal distributions in real number field have an overlapped region, we get $P_{j<g} \in (0,1)$ and thus x_{jg} and $t_g - s_{j`}$ are finite numbers. In practice, the probability $P_{j<k}$ is estimated with a finite number of subjective scores and thus it is possible that the estimated value $\hat{P}_{j<k} = 0$ or $\hat{P}_{j<k} = 1$. In such a case, x_{jg} is infinite and thus cannot be used in the equations. How to address this issue was elaborated in [8] where two conditions are required: firstly, there should be at least three grade scales are used; secondly, the subjective scores of each stimulus should be with a deviation larger than 0. The first condition can be easily fulfilled with ITU-R grade scales. The second condition may be fulfilled with a large number of subjects; if not, iMOS is not applicable here.

Compressed Video Quality Evaluation With IMOS

We use iMOS to evaluate compressed video quality when there are deviations in subjective scores. To show the relationship between iMOS and MOS in compressed videos, we build a database with four High Definition (HD) 1080p sequences: Cheetah, Crowd, Football and ParkScene. The Spatial Information (SI) and Temporal Information (TI) [4] values of these sequences are (74.98, 42.34), (64.91, 18.93), (59.76, 24.20) and (53.46, 15.94), respectively. Detailed information of subjective test is summarized as follows:

(1) Stimuli: All the four sequences are coded with x264 and five Constant Rate Factor (CRF) values: 15, 22, 28, 34 and 40, which results in 20 stimuli and all these stimuli are displayed in a random order;

(2) Duration: Each stimuli is with a 10-second duration and a frame rate of 24 frames per second;

(3) Device: We present all stimuli on TV with a viewing distance of 90 inches;

(4) Subjects: Totally 21 subjects performed the test;

(5) Grade scale: Single-stimulus test is employed with a continuous scale from 0 to 100 [3].

We calculate the MOS and iMOS values based on the database and illustrate the relationship between them in Fig. 2.

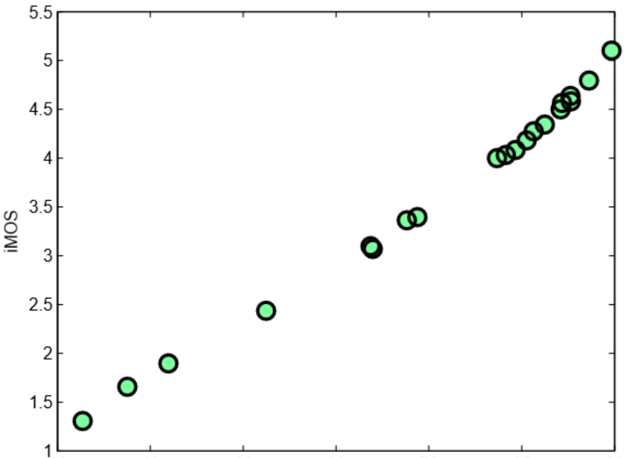
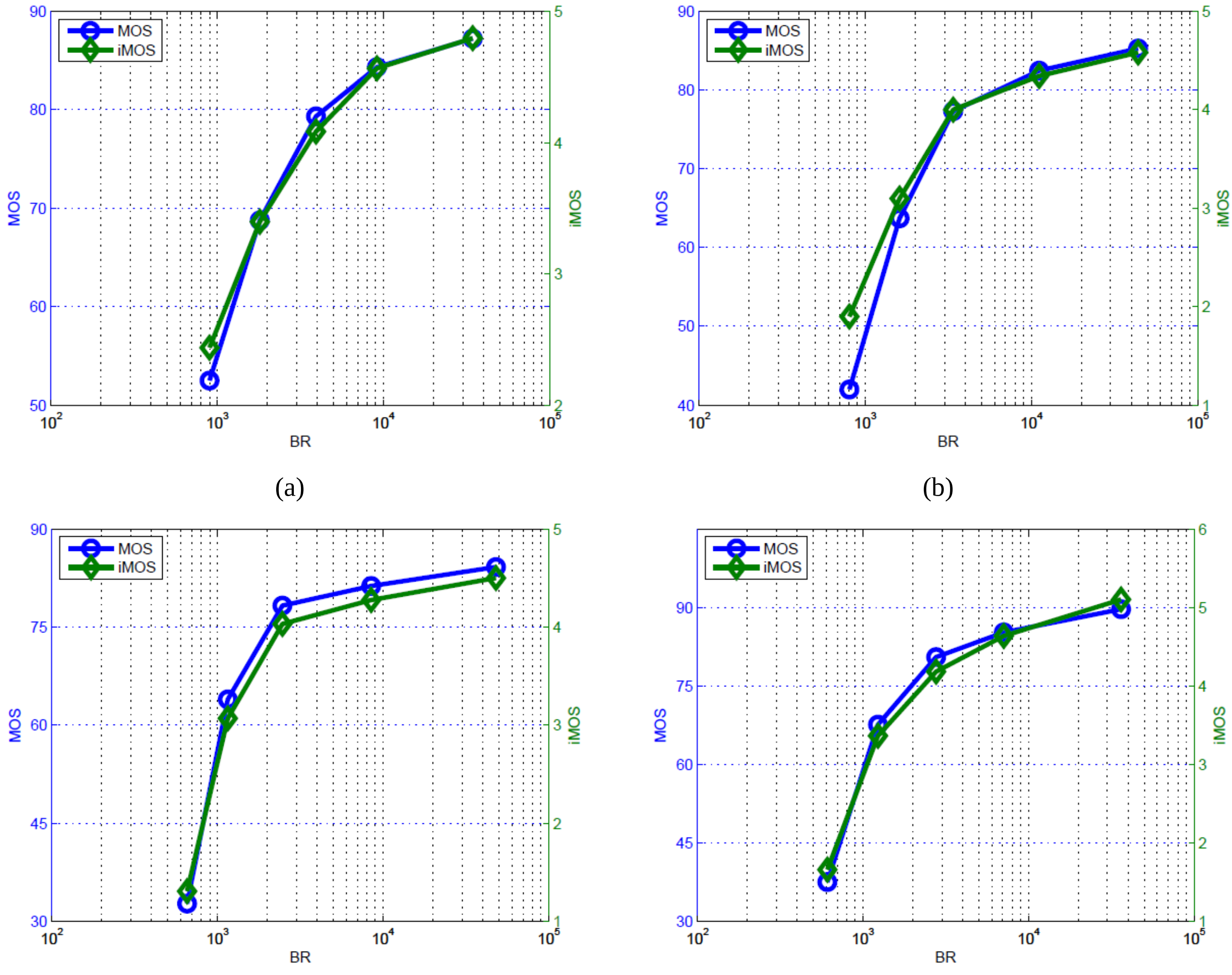


Fig. 2. The relationship between MOS and iMOS on our database

From Fig. 2, we notice that iMOS behaves an approximately linear relationship to MOS, with a Pearson Linear Correlation Coefficient (PLCC) of 0.9957. However, the slopes at the high and low ends are different from that in the middle range, which also justifies the above criticism that the differences between any two consecutive scales may not have identical impact on human vision. In [8], they had a similar conclusion on the user-level Quality of Service (QoS) considering mutually compensatory property of multimedia contents; while here we show the characteristics of iMOS values on compressed videos.

To further study the characteristics of iMOS values, we present the BR-iMOS curves of the four above sequences, as shown in Fig. 3. As a reference, the BR-MOS curve is also given in each sub-figure. From these figures we can notice that, the behavior of BR-iMOS is similar to BR-MOS: the quality value increases quickly and has a converging trend as BR increase. The increasing rate of BR-iMOS is similar to that of BR-MOS in middle rate range (*i.e.*, 10^2 ~10^3), while there exist differences at low and high rate ends.



(a)

(b)

(c)

(d)

Fig. 3. Typical examples of MOS and iMOS versus BR. (a) Cheetah; (b) Crowd; (c) Football; (d) ParkScene

Conclusions

In this work, we presented an effective method on comparison of MOS values with distributions. As a conclusion, iMOS can show subjective scores in an interval scale with simple computations. Hence, it would be a better alternate to represent human opinion when the distribution information of subjective scores is available.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61571129, 61671152), and Natural Science Foundation of Fujian Province (No. 2015J01250, 2016J01299),

References

[1] B. Bross, W.J. Han, G. J. Sullivan, J.R. Ohm, and T. Wiegand, “High efficiency video coding (HEVC) text specification draft 8, ” Doc. JCTVC-J1003, July 2012.

[2] “Advanced video coding for generic audiovisual services,” ISO/IEC 14496-10:2005(E) ITU-T Rec. H.264 (E), Mar. 2005.

[3] “Methodology for the subjective assessment of the quality of television pictures,” ITU-R Rec. BT.500-13, Jan. 2012.

[4] “Subjective video quality assessment methods for multimedia applications,” ITU-T Rec. P.910, Apr. 2008.

[5] “Methodology for the subjective assessment of video quality in multimedia applications,” ITU-R Rec. BT.1788, Jan. 2007.

[6] S. Winkler and P. Mohandas, “The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics,” IEEE Trans. Broadcast, 54(3), pp. 660-668, 2008.

[7] S. S. Stevens, “On the theory of scales of measurement,” Science, 103(2684), pp. 677-680, 1946.

[8] S. Tasaka and Y. Ito, “Psychometric analysis of the mutually compensatory property of multimedia QoS,” IEEE Int. Conf. Commun. 2003 (ICC’03), pp. 1880-1886, May 2013.

[9] W. S. Torgerson, “Theory and methods of scaling,” Wiley, New York, 1958.

[10] L.L. Thurstone, “A law of comparative judgment,” Psychological Review, 34(4), pp. 273-286, 1994.