SJTU 4K Video Subjective Quality Dataset for Content Adaptive Bit Rate Estimation without Encoding

Yutong Zhu[1], Li Song[1,2], Rong Xie[1,2], Wenjun Zhang[1,2]
[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2] Cooperative Medianet Innovation Center,
Shanghai, China
{zyt_irene, song_li, xierong, zhangwenjun}@sjtu.edu.cn

***Abstract*—Recently, parametric models of predicting video subjective quality by exploiting packet layer or bit stream layer information at decoder side has achieved significant progress. In contrast, there are few works on estimating subjective quality in the stage of video encoding, partially because of limited and outdated dataset. In this paper, we contribute a new 4K video dataset with full subjective scores (MOS) at different bitrates compressed by HEVC/H.265 standard. Furthermore, a new parametric model based on content clustering analysis is proposed and validated on our dataset as a benchmark for future research.**

Keywords—Objective evaluation techniques; Subjective evaluation techniques; Video Quality Assessment; Bit rate estimation; MOS; UHD; HEVC

I. INTRODUCTION

With the fast progress in video coding technology and rapid advancement in display devices, video, one efficient method of transporting information, has distributed over digital networks, broadcast channels and packaged media. Since the demand of both video content and quantity are already satisfied, people tend to pursue better video quality more than ever. In this context, video quality assessment (VAQ) has become a hot topic in both academia and industry.

Video quality assessment can roughly fall into two categories: Subjective and Objective. Subjective assessment requires severe restrictions on test conditions and time consuming, but it is still the most reliable method to evaluate the performance of an objective quality metric or algorithm [1] , as it reflects the subjects' visual perception directly. In contrast, objective assessment appears to be more flexible, less costly and can be used into real-time systems for optimization [2]. Up till now, a great amount of parametric models have been proposed to predict video quality. In [3], authors find out that video quality is basically affected by five dimensions: encoder type, video content, bit rate, frame rate and resolution. Work in [4] shows that bit rate is also an indispensable factor with video quality. In [5], authors propose a full-reference model with PSNR, frame rate and quantization artifacts. However, the model needs to encode the test sequences first. Work in [6] proposes a model using video content to predict PSNR. In [7], a video quality estimation model is proposed based on bit rate and frame rate. Video content is used as a feature for classification with an average level of estimation performance. Research in [8] studys several parametric models and reveals a fact that models considering video content tend to show a better performance.

Lots of work have been done with great progress at decoder side. However, there are few works on estimating video quality in the stage of video encoding, partially because of the limited and outdated dataset. Meanwhile, relative fields are still lack of available research-free database of 4K video sequences (videos with resolution of 3840×2160) with subjective assessment results. Hence, to fill the gap and estimate video quality at encoder side, a subjective video quality test is conducted with the dataset exposed contributively. Then, a simple benchmark modified from a no-reference model is provided to predict the bit rate when given certain video quality for a specific video sequence, and is validated on our 4K dataset as well.

The remainder of this paper is organized as follows. Section II describes the subjective assessment test mentioned above. Section III presents the performance of the subjective test. Section IV presents content based clustering analysis and the benchmark we modified. Section V provides some conclusions of our work.

II. SUBJECTIVE VIDEO QUALITY ASSESSMENT ON 4K SEQUENCES

So far the availability of high quality 4K uncompressed video sequences free for research purpose is still very limited, including SJTU 4K video sequences dataset[9], Elemental [10] and Ultra Video Group of Tampere University of Technology [11]. Furthermore, there are limited subjective datasets for video beyond HD and compressed by HEVC are available yet. This fact motivates our work in this paper.

A. Test Set-up

In our subjective test, the laboratory is set up according to

978-1-4673-9044-6/16/$31.00 ©2016 IEEE

[12]. The monitor is calibrated before the starting of the test. We select a professional high-performance 4K monitor SONY KD-55X9000A to display the compressed video sequences, driven by the graphic board of video server. A picture of test environment is shown in Fig. 1. The test involves two subjects rating the test video sequences, seated in the front of the monitor. The distance of the subjects from the monitor is about 1.5 times the height of the video monitor.

B. Test Methodology

As all the subjects of the test are filtered in advance to eliminate the influence from individuals, the *DSIS (Double Stimulus Impairment Scale)*, variant II, with a continuous impairment scale is selected. The DSIS method is required to use the five-grade impairment scale. During the test, the order of stimuli is random.

C. Source Sequences for Test

By choosing spatial and temporal complexity as reflection of video content, we calculate SI (Spatial Information) and TI (Temporal Information) indexes on the luminance component [13]. For each frame at time n, SI is defined as the maximum value of the standard deviations after sobel filtering, while TI is based on the differences between two consecutive frames on pixel level:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \quad (1)$$

$$TI = max_{time}\{std_{space}[F_n(i,j) - F_{n-1}(i,j)]\} \quad (2)$$

where F_n is the video frame at time n.

To make sure that the selected sequences span a wide rage of SI and TI, we choose 9 sequences from our 4K video database and another three 4K sequences from Elemental. The SI and TI indexes are displayed in Fig. 2. Each of the sequences is progressively scanned in 8 bit color depth with YUV 4:2:0 color sampling and 30fps. Among the twelve selected sequences, News and Tree Shade are set as training sequence and dummy sequence individually. All the first frame of the 4K raw sequences used in the test are displayed in Fig. 3.

D. Encoding Algorithm

All the source sequences are compressed by HEVC with random access configuration [14]. Due to the different spatial-temporal characteristics of video content, we select the bit rates for each sequence separately. Six compressed videos are selected for each source sequence by viewing compressed videos generated using a variety of bit rates and selecting a subset that spanned the desired range of visual quality. The complete bit rate sets can be found on our website: http://medialab.sjtu.edu.cn/resources/resources.html.

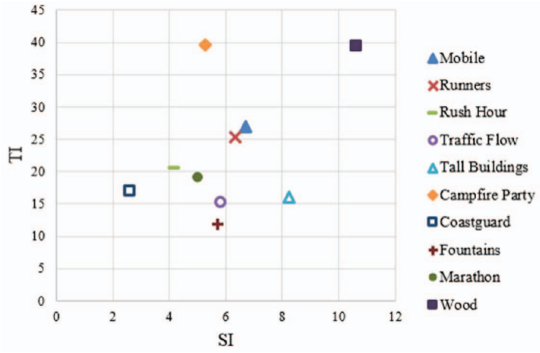

Fig. 2. Spatial and Temporal information indexes of the source

E. Test Subjects

Unlike a minimum of 15 subjects recommended in [12], the number becomes 24 in [15]. Rigorously, a total of 42 people are involved in our subjective test eventually. 18 of them are females and the age ranges from 20 to 31 years old. All of them are non-experts.

F. Data Processing

The raw subjective scores have been processed in order to obtain the final MOS value. The whole data processing procedure, including the data screening and analysis technique, is also strictly executed by steps listed in [12].

III. Subjective Test Performance Analysis

Given the limited space available, four representatives of the MOS values for different sequences with different bit rates is shown in Fig. 4. The comlete 4K video quality database is on our website as well. As a general comment, the MOS plots clearly show that the test has been properly designed for that the subjective scores span the most range of quality levels.

It can be seen that for different 4K video sources, after HEVC compression, the video quality appears different trends and MOS of most sequences are greater than 4 when the bit rate reaches 15Mbps. Therefore, many manufacturers come to an agreement that the average bit rate for 4K sequences in broadcasting is 15Mbps. However, this tacit standard seems to be a huge waste, as it can be observed from Fig. 4(a) that MOS gets to 4 with a bit rate of 12Mbps, while in Fig. 4(d) MOS already saturates with a bit rate of 3Mbps, which makes 15Mbps for 4K sequences quite a rough conclusion. Hence,

Fig. 1. Subjective test environment set up

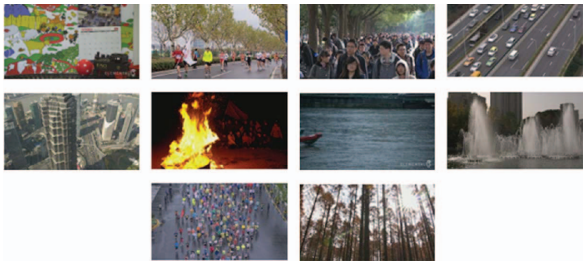

Fig. 3. The first frame of the 4K raw sequences used in the test

we propose a model to predict the exact bit rate required for each certain sequence at certain video quality.

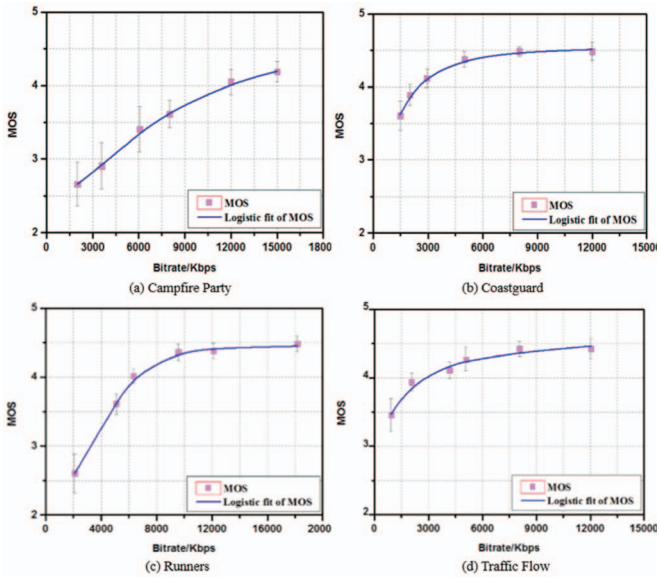

Fig. 4. The representatives of MOS with 95% confidence interval

IV. Content Adaptive Bit Rate Estimation Model

To achieve quality driven encoding goal, we propose a non-reference bit rate estimation model derived from objective video quality assessment. The entire process can be roughly divided into three steps: basic model selection, clustering according to content and final MOS-rate estimation model.

A. Basic Parametric Model Selection

Generally, MOS can be estimated as follows:

$$MOS = f(v_c, BR, f_r, r_s, e_c) \tag{3}$$

where f represents a relationship among the six variables. v_c represents video content, and BR means bit rate used in encoding process, while f_r, r_s and e_c represents frame rate, resolution and encoding type separately. In our paper, f_r equals to 30fps, while r_s is 4K and e_c is fixed as HEVC. Then, we can get the following formulas by simple mathematical derivation:

$$BR_p = g(v_c, MOS) \tag{4}$$

$$g = f^{-1} \tag{5}$$

where g is the inverse function of f, and BR_p is the predicted bit rate.

TABLE I. Silhouette Value of Clustering Analysis

Category	$K_{ca}=2$	$K_{ca}=3$	$K_{ca}=4$	$K_{ca}=5$
$Silh_{min}$	0.3905	0.1383	0.5069	0.5069
$Silh_{max}$	0.9381	0.9793	0.9677	1
$Silh_{mean}$	0.839	0.7643	0.7410	0.7717
$Silh_{dev}$	0.1726	0.2305	0.1620	0.1911

Considering our goal model to be video content based without encoding, hence, we choose the model in [16] and make some modification on it to reduce the model complexity. The original model uses video characteristics to estimate QP and then to predict MOS. Instead of estimating QP from reference sequences, we use the MOS data we collect above in a coefficients fitting method to predict bit rate. Combining (4) and the original model, we make a mathematical derivation. At last, we establish a modified model as (6) to (10) shows.

$$BR_p(v_c, MOS) = \left(\frac{\beta}{c_2}\right)^{\frac{\ln 10}{\alpha}} * v_c^{-\frac{c_6}{\alpha}} \tag{6}$$

$$v_c = TI * SI \tag{7}$$

$$\alpha(v_c) = c_1 + c_2 log(v_c) \tag{8}$$

$$\beta(v_c, MOS) = \frac{MOS}{\gamma(\gamma - MOS)} \tag{9}$$

$$\gamma(v_c) = c_4 + c_5 \log(v_c) \tag{10}$$

where c_1 to c_6 are model coefficients. The least squares method is used to fit the model.

B. Clustering According to Video Content

After selecting the basic model, we make bit rate estimation through the whole database, and come to a result that the *PCC (Pearson Correlation Coefficient)* and the *SCC (Spearman Correlation Coefficient)* between BR_p and real bit rate are 0.672 and 0.753, with the *RMSE* 1.174. To improve the accuracy, we make clustering analysis before prediction to group sequences with same characteristics into one category.

We cluster the sequences by TI and SI using K-means algorithm. We choose silhouette to evaluate the performance of clustering [17]. A high silhouette value indicates compact matching to its own category and poor matching to others. We calculate the silhouette for all the source sequences when clustering into different categories. Table I shows the minimum value $Silh_{min}$, maximum value $Silh_{max}$, mean value $Silh_{mean}$ and deviation value $Silh_{dev}$ of silhouette, while K_{ca} represents the number of categories.

According to Table I, there is a highest $Silh_{mean}$ and a medium $Silh_{dev}$ when K_{ca} equals to 2. However, when we cluster the sequences into 2 categories, the PCC and SCC between BR_p and real bit rate for one category are only 0.7740 and 0.8324, with the RMSE 0.9096. When K_{ca} equals to 3, $Silh_{min}$ falls down to 0.1383, which means poor performance and the configuration should have more categories. When K_{ca} equals to 5, the $Silh_{max}$ is as high as 1 with the fact that there is only one sequence in that category. By above analysis, we choose 4 categories (K_{ca}=4) as our final choice and sequences falling into different categories are summarized in Table II .

TABLE II. Results of Clustering Analysis

Category	Source Sequences
Category A	Mobile, Runners
Category B	Campfire Party, Wood
Category C	Fountains, Traffic Flow, Tall Buildings
Category D	Coastguard, Rush Hour, Marathon

C. Contend Adaptive MOS-Rate Estimation Model

By regressing model in (6) in least squares method, we obtain prediction results for all the sequences. The dispersion between BR_p and real bit rate is presented in Fig. 5. The PCC and SCC between BR_p and real bit rate of each category are calculated and displayed in Table III. RMSE is also presented.

TABLE III. PERFORMANCE COMPARISON FOR EACH CATEGORY

Category	PCC	SCC	RMSE	MOS
Category A	0.972	0.986	0.102	3.945
Category B	0.953	0.951	0.087	3.818
Category C	0.901	0.865	0.274	4.124
Category D	0.961	0.969	0.177	4.041
All Sequences	0.672	0.753	1.174	4.002

According to Table III, we can draw several conclusions. First, the proposed model show good performance on our dataset with a highest PCC of 0.972. However, further study and validation on large dataset are definitely needed. Second, the accuracy of prediction gets a 28.76% increase on PCC and 68.98% reduce on RMSE by clustering different videos into 4 categories. This fact shows that the methodology of clustering first before using parametric model is key to success of content adaptive prediction. Thirdly, different categories have diverse average MOS values as expected. Specifically, category B has the lowest MOS of 3.818 with the maximum TI and SI, while category C has the highest MOS of 4.124 with a minimum TI and medium SI. This phenomenon confirms that there exists a complex relationship between video quality and content.

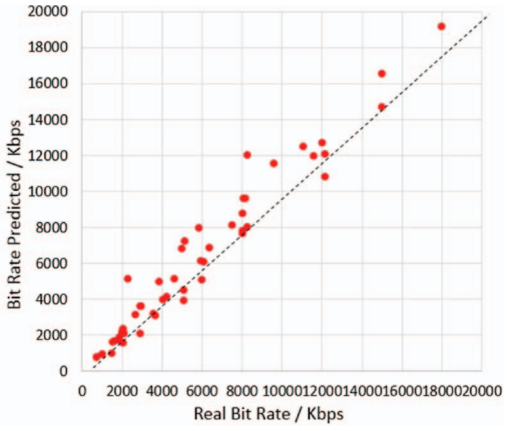

Fig. 5. The dispersion between real bit rate and predicted bit rate

CONCLUSION

In this paper we have investigated the problem of content adaptive bit rate estimation without encoding. First we present a publicly available dataset of subjective assessment results for 4K video sequences generated by HEVC encoder. Then we provide a simple but effective bit rate estimation model based on video content as a benchmark, which can predict the bit rate required for encoding processing at a certain video quality. More comprehensive work is under exploration.

ACKNOWLEDGMENT

This work was supported by NSFC (61521062, 61527804), the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

REFERENCES

[1] M. H. Pinson, L. Janowski and Z. Papir, "Video quality assessment: subjective testing of entertainment scenes," IEEE Trans. Signal Processing. vol. 32, no. 1, pp. 101-114, 2015.

[2] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution, " IEEE Trans. Broadcasting, vol. 61, no. 3, pp. 520-531, Sept 2015.

[3] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, "Cross-Dimensional perceptual quality assessment for low bit-rate vdeos, " IEEE Trans. on Multimedia, vol. 10, no. 7, pp. 1316-1324, 2008.

[4] G. Cermak, M. Pinson, S. Wolf. "The relationship among video quality, screen resolution, and bit rate," IEEE Trans. on Broadcasting, vol. 57, no. 2, pp. 258 - 262, 2011.

[5] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts, " IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 3, pp. 286-298, Mar 2011.

[6] L. Anegekuh, L. Sun, E. Ifeachor, "Encoded bitstream based video content type definition for HEVC video quality prediction," IEEE Internatonal Conference on Communications , pp. 1296-1301, Jun 2014.

[7] M. Ries, C. Crespi, O. Nemethova, and M. Rupp, "Content based video quality estimation for H.264/AVC video streaming," in Proc. WCNC, pp. 2669-2673, 2007.

[8] J. Joskowicz, R. Sotelo and J. C. L. Ardao, "Towards a general parametric model for perceptual video quality estimation," IEEE Trans. Broadcast., vol.59. no.4, pp.569-579, Dec 2013.

[9] L. Song, X. Tang, W. Zhang, X. Yang and P. Xia, "The SJTU 4K video sequences dataset," International Workshop on QoMEX, 2013.

[10] Elemental Technologies. [Online]. Available: http://www.elementaltechnologies.com/resources/4k-test-sequences.

[11] Ultra Video Group. [Online]. Available: http://ultravideo.cs.tut.fi/.

[12] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures, " International Telecommunication Union, 2012.

[13] ITU-T, "Subjective video quality assessment methods for multimedia applications", Recommendation ITU-R P 910, Sep 1999.

[14] JCT-VC of ISO/IEC MPEG and ITU-T VCEG. [Online]. Available: http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-11.0/

[15] M. H. Pinson, L. Janowski, R. Pépion, Q. Huyunh-The, Ch. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective test: an international study," IEEE J. Select. Topics Signal Processing, vol. 6, no. 6, pp. 640-651, Oct 2012.

[16] M. Takagi, H. Fujii and A. Shimizu, "Optimized spatial and temporal resolution based on subjective quality estimation without encoding," IEEE International Conference on Visual Communications and Image Processing, pp. 33 - 36, 2014.

[17] A. Khan, L. Sun and E. Ifeachor, "Content clustering based video quality prediction model for mpeg4 video streaming over wireless networkds," IEEE International Conference on Communications, pp. 1-5, 2009.