Which Metric Can Predict Coding Gain of H.265/HEVC over H.264/AVC?

Jianhua Xiao[1], Li Song[1,3], Zhengyi Luo[2], Rong Xie[1,3], Wenjun Zhang[1,3]

[1] *Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University*
[2] *School of Electronics and Information Engineering, Shanghai University of Electric Power*
[3] *Cooperative Medianet Innovation Center, Shanghai China*
xjh01091398@163.com, {song_li, rongxie, zhangwenjun}@sjtu.edu.cn, lzy@shiep.edu.cn

***Abstract*—Subjective evaluation conducted by JCT-VC (Joint Collaborative Team on Video Coding) members shows that H.265/HEVC achieves about 50% rate saving over H.264/AVC without sacrificing subjective quality. In this paper, we study 13 objective image and video quality assessment (IQA/VQA) metrics, including the recently proposed ones—FSIM [1], GMSD [2] and IWSSIM [3], in terms of coding gain prediction of HEVC over H.264. Experimental results on HEVC Class B and Class C test sequences show that most of the metrics underestimate the rate saving. Surprisingly, a relatively old metric—Noise Quality Measure (NQM) [4] index consists well with subjective evaluation. To verify the universality of the phenomenon, we carried out further tests on another ten video sequences with different levels of spatial and temporal complexities. The experimental results show that NQM still predicts rate saving more accurately than the other metrics.**

***Index Terms*—HEVC, objective quality assessment, BD-Rate**

I. INTRODUCTION

As HEVC was approved by JCT-VC, a great number of scholars have conducted researches to evaluate the performance of HEVC over previous standards. Subjective evaluation shows that H.265/HEVC achieves about 50% rate saving over H.264/AVC [5] [6].

Compared with subjective quality assessment, objective quality assessment is more convenient and less costly. A number of metrics have been proposed to evaluate the quality of images and video sequences. Meanwhile, objective image and video quality assessment (IQA/VQA) metrics have also been applied to performance evaluation of HEVC. Zhao et al. [7] evaluated the performance of HEVC using three objective metrics—PSNR, SSIM [8] and PQI [9]. They only offered the average BD-Rate [10] saving and didn't analyze the performance of the metrics further. Ohm et al. [6] compared the coding efficiency of HEVC with that of previous standards by means of subjective tests and the objective metric PSNR. Results show that PSNR doesn't always agree with subjective evaluation. Zeng et al. [11] studied the performance of VQA metrics by comparing the coding gain of HEVC over H.264, but it's not comprehensive to use only five metrics among numerous objective metrics.

From the above we know that it is highly desired to examine whether existing objective IQA/VQA metrics consist with the subjective evaluation of coding performance. In this paper, we first select as many as 13 objective quality metrics to evaluate the coding efficiency of HEVC. Then results of objective evaluation are compared with subjective results in [5] based on BD-rate. Moreover, we carry out further tests to evaluate the 13 metrics' performance on video sequences with different levels of spatial and temporal complexities. The provided results may help select suitable quality metrics for comparison or even development of video encoders. It is also expected for the results to be instructive for development of new quality assessment algorithms.

The rest of this paper is organized as follows. Section II introduces the metrics evaluated in our paper. The Bjøntegaard measurement method [12] is introduced to calculate BD-Rate in Section III. In Section IV, the coding configuration and experimental results are presented. The results are discussed in depth in Section V. Finally, conclusions are drawn in Section VI.

II. SELECTED OBJECTIVE QUALITY METRICS

Over these years, image and video quality assessment has always been a hot research topic. Most video quality assessment methods are extended from image quality assessment by evaluating video frame by frame, such as SSIM [8], VIF [13], VSNR [14]. They do not consider temporal information in video. Only a few metrics, e.g. MOVIE [15], take account of both spatial and temporal information but often have high computational complexities.

For the sake of comprehensiveness, we examine a wide variety of full reference (FR) IQA/VQA algorithms and

finally the following 13 metrics are chosen.

- *PSNR is widely used in image and video quality assessment. It is a simple function of the Mean Squared Error (MSE) between source and compressed video.*
- *The Feature Similarity (FSIM) [1] index for Image Quality Assessment is a new metric which uses phase congruency (PC) as the primary feature.*
- *GMSD [2] (Gradient Magnitude Similarity Deviation) computes the pixel-wise gradient magnitude similarity, which is then used to generate the final measure as the standard deviation of the Gradient Magnitude Similarity Deviation map.*
- *IW-SSIM [3] is improved from original MSSIM by introducing an information-content weighting (IW) based quality score pooling strategy.*
- *NQM [4] can be used to reveal nonlinear weighted signal to noise ratio for additive noise.*
- *SSIM [8] is a popular metric, which is designed initially for image quality assessment but has been extended to video now.*
- *VIF [13] stands for visual information fidelity, which is computed in wavelet domain.*
- *VIFP [13] computes the visual information fidelity in the pixel domain.*
- *VSNR [14] is a metric that computes the visual signal to noise ratio based on wavelet transform.*
- *Multiscale SSIM (MSSIM) [16] is designed based on SSIM but shows better consistency with subjective evaluation than PSNR, SSIM and so on.*
- *IFC [17] is not a distortion but fidelity criterion. It theoretically ranges from zero (no fidelity) to infinity (perfect fidelity within a nonzero multiplicative constant in the absence of noise).*
- *UQI [18] models any image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion.*
- *WSNR [19] computes signal to noise ratio with a weighted process, which also models HVS and considers the saliency map of image or video.*

Each metric index is computed on the luminance component frame by frame in this paper, and the final metric index is obtained by averaging the frame level quality.

III. Bit Rate Saving Calculation

We calculate the bit rate saving of H.265/HEVC over H.264/AVC by the Bjøntegaard measurement method [12] (as illustrated in Figure 1[12]). Firstly, the RD curve of each encoder is plotted by the following steps:

1. Each encoder compresses the video sequence with 4 different quantization parameters (QP).
2. The metric index for each compressed sequence is calculated.
3. A curve is fitted through the obtained 4 bitrate/index points.

It should be noted that bitrate is on logarithmic form for the reason that the difference between the curves is dominated by the high bitrates.

Secondly, the BD-rate predicted by each metric is calculated by two steps:

1. A third order polynomial is fitted through each set of data points.
2. The BD-Rate is the average bitrate difference over the integral interval [a, b].

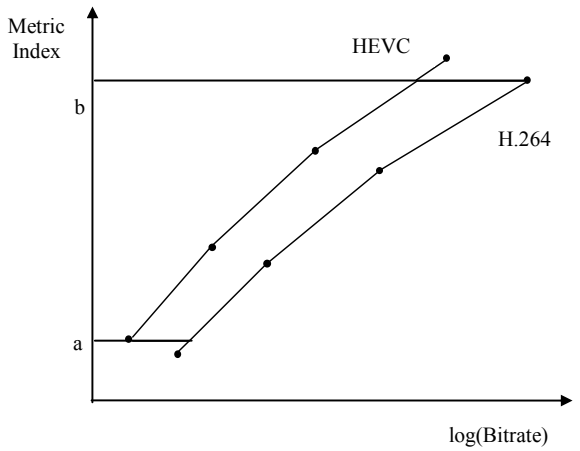

Figure 1. BD-Rate calculation

IV. Experiments And Analysis

In this paper, H.265/HEVC test model software HM16.0 and H.264/AVC test model software JM18.6 are used for experiments. For the convenience of comparison with subjective evaluation in [6], the same HEVC test sequences—5 class B sequences of 1080p resolution (1920x1080) and 4 class C sequences of WVGA resolution (854x480) are chosen. The random access high efficiency is chosen as the coding configuration of HM16.0, while for JM18.6 the coding configuration is set as HM-like for fair comparison. Rate control for both encoders is off. Quantization parameter (QP) values for HM16.0 are set as 31, 34, 37 and 40, while for JM18.6 QP values are set as 27, 30, 33 and 36.

The rate saving predicted by each metric over each video sequence is summarized in Table I. To measure the performance of each metric, the rate saving predicted by subjective scores (MOS) in [6] is also included in Table I. Many criteria, e.g. PLCC, MAE, RMSE, SRCC and KRCC, are available for evaluation of prediction accuracy. As RMSE is more intuitive as discussed in [20], RMSE of the BD-rate between the subjective and objective evaluation is calculated. Low RMSE values denote high prediction accuracy of rate saving.

As can be observed from Table I, most metrics underestimate the rate saving. With regard to the video

TABLE I. BD-RATE PREDICTED BY 13 IQA/VQA METRICS AND SUBJECTIVE SCORES (MOS) OVER VIDEO SEQUENCES CLASS B AND CLASS C

Objective Metrics	Video Sequences (Represented by the first three characters)									Average	RMSE
	Class B					Class C					
	Bas	BQT	Cac	Kim	Par	Bas	BQM	Par	Rac		
PSNR	-45.6%	-43.2%	-36.7%	-48.7%	-36.2%	-35.5%	-32.9%	-26.9%	-30.1%	-37.3%	13.15
SSIM	-48.5%	-40.5%	-36.8%	-49.9%	-36.9%	-35.8%	-33.7%	-27.2%	-31.0%	-37.8%	12.83
IWSSIM	-49.0%	-41.0%	-38.8%	-50.8%	-38.1%	-37.7%	-33.8%	-29.2%	-33.0%	-39.0%	11.90
GMSD	-46.9%	-39.9%	-37.5%	-49.2%	-37.3%	-35.6%	-32.9%	-27.7%	-31.6%	-37.6%	13.13
FSIM	-46.7%	-41.7%	-39.8%	-50.6%	-37.3%	-36.0%	-32.9%	-26.8%	-31.6%	-38.2%	12.47
MSSIM	-48.3%	-40.4%	-38.0%	-50.3%	-37.5%	-35.9%	-33.6%	-27.2%	-32.2%	-38.1%	12.55
UQI	-48.8%	-40.4%	-36.9%	-49.7%	-37.1%	-35.8%	-34.0%	-27.3%	-31.0%	-37.9%	12.75
VIF	-46.3%	-41.0%	-36.6%	-48.1%	-35.0%	-36.3%	-32.3%	-25.8%	-30.9%	-36.9%	13.53
VIFP	-47.4%	-41.8%	-37.0%	-49.0%	-35.6%	-36.0%	-33.2%	-26.6%	-31.2%	-37.5%	12.95
IFC	-46.4%	-41.2%	-36.8%	-47.8%	-34.7%	-36.5%	-32.5%	-25.6%	-30.8%	-36.9%	13.48
VSNR	-44.3%	-40.5%	-36.0%	-47.7%	-35.3%	-34.7%	-31.6%	-24.7%	-29.7%	-36.1%	14.31
WSNR	-51.4%	**-62.3%**	-40.5%	**-54.6%**	-47.8%	**-38.4%**	-39.2%	-35.7%	-34.9%	-45.0%	7.17
NQM	**-59.2%**	-61.5%	**-47.8%**	-63.9%	**-49.0%**	-36.8%	**-39.4%**	**-30.7%**	**-36.5%**	**-47.2%**	**5.19**
MOS	-66.1%	-63.1%	-50.2%	-55.2%	-49.7%	-44.9%	-41.6%	-29.6%	-42.7%	-49.2%	--

sequence "BQTerrace", rate saving may be underestimated by more than 20%. It can also be found that underestimation is more serious for sequences of high resolutions. In Table I, the best prediction of rate saving is highlighted in boldface. Apparently, the prediction of NQM and WSNR consists well with subjective evaluation. The classical metrics like PSNR and SSIM underestimate the rate saving by more than 10%, while the latest three metrics—IWSSIM, GMSD and FSIM obtain similar results with PSNR and SSIM. The conclusion can be further confirmed by the RMSE values. RMSE values for NQM and WSNR are 5.19 and 7.17 respectively, while RMSE values for the others range from 11.90 to 14.31.

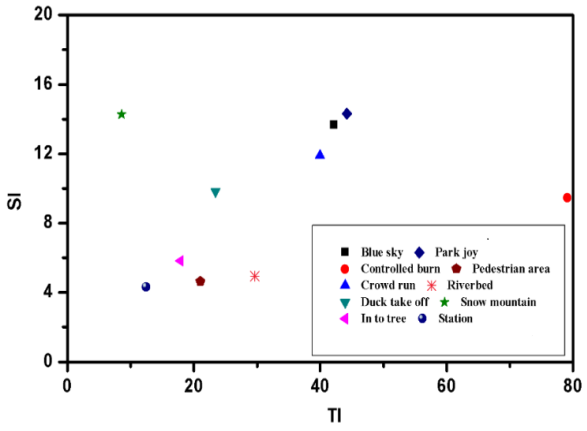

Figure 2. Spatial information and temporal information indexes of the selected video sequences

In order to verify the universality of this phenomenon, more experiments are conducted. Considering the possible influence of video content on quality evaluation, we choose another ten 1920x1080 sequences with different levels of spatial and temporal complexities, which are freely available from [21]. The analysis of video sequences' content complexity has been performed by computing the spatial information (SI) and temporal information (TI) indexes on the luminance component according to [22]. SI and TI indexes of video sequences are shown in Figure 2.

The coding configuration is set the same as before. The results are listed in Table II. As can be observed from Table II, the gap of predicted rate saving between NQM and other 11 metrics is more than 10%, which consists well with the previous results, while WSNR doesn't provide better performance than PSNR, which is far from its previous performance. We may draw the conclusion that WSNR doesn't provide stable performance. When the sequences have relatively high levels of spatial complexities, NQM still performs normally but the performance of the other metrics is unacceptable. Besides, we find that the 12 metrics tend to be more sensitive to sequences' spatial complexity than temporal complexity. For example, for the sequence "snow mountain" with high spatial complexity and the sequence "controlled burn" with high temporal complexity, the 12 metrics perform better over "controlled burn" than "snow mountain". Furthermore, the experimental results also show that the recently proposed three metrics—IWSSIM, GMSD and FSIM perform only slightly better than PSNR and SSIM.

V. DISCUSSIONS

As observed from above, we have the following findings. First, most metrics underestimate the rate saving of H.265/HEVC over H.264/AVC. What's more, most metrics, including the three recently proposed ones—IWSSIM, GMSD and FSIM, do not perform much better than the classic metric PSNR in spite of their relatively high

TABLE II. BD-RATE PREDICTED BY 13 IQA/VQA METRICS OVER ANOTHER TEN SEQUENCES WHICH HAVE DIFFERENT CONTENT COMPLEXITIES

Objective Metrics	Another Ten Video Sequences with Different Content Complexities										Average
	sky	burn	crowd	duck	tree	joy	area	river	snow	station	
PSNR	-42.1%	-27.5%	-16.4%	-31.3%	-48.9%	-17.7%	-45.0%	-34.4%	-16.3%	-91.6%	-37.1%
SSIM	-40.2%	-25.5%	-19.7%	-36.2%	-50.2%	-25.1%	-45.0%	-33.4%	-13.7%	-35.4%	-32.4%
IWSSIM	-44.8%	-26.9%	-19.1%	-38.7%	-57.5%	-28.8%	-46.1%	-33.0%	-24.2%	-54.0%	-37.3%
GMSD	-40.0%	-27.0%	-20.4%	-37.2%	-54.2%	-23.8%	-43.8%	-35.1%	-14.7%	-58.7%	-35.5%
FSIM	-42.9%	-28.0%	-20.3%	-40.2%	-58.1%	-24.7%	-45.6%	-35.7%	-19.1%	-56.5%	-37.1%
MSSIM	-43.4%	-26.0%	-21.1%	-40.4%	-56.3%	-27.1%	-46.5%	-34.8%	-16.3%	-34.1%	-34.6%
UQI	-39.2%	-29.8%	-24.1%	-41.7%	-53.4%	-35.7%	-50.0%	-37.0%	-12.9%	-80.4%	-40.4%
VIF	-39.9%	-22.0%	-16.0%	-30.9%	-46.1%	-17.5%	-43.3%	-31.4%	-15.5%	-49.6%	-31.2%
VIFP	-41.9%	-25.4%	-18.7%	-35.9%	-50.1%	-24.2%	-46.1%	-35.3%	-15.6%	-61.0%	-35.4%
IFC	-40.5%	-20.6%	-16.4%	-32.3%	-47.3%	-18.2%	-42.8%	-33.3%	-15.4%	-31.0%	-29.8%
VSNR	-39.1%	-21.3%	-16.0%	-32.2%	-50.6%	-17.8%	-43.2%	-32.1%	-17.7%	-51.3%	-32.1%
WSNR	-42.4%	-29.3%	-16.8%	-33.6%	-52.7%	-19.8%	-44.7%	-33.6%	-20.7%	-58.7%	-35.2%
NQM	-58.1%	-56.3%	-34.2%	-50.0%	-62.4%	-40.0%	-56.6%	-46.1%	-50.9%	-67.6%	**-52.2%**

computational costs. Secondly, some metrics, such as WSNR, may perform well over some specific video sequences but fail to provide good performance over others. Therefore, a more comprehensive video sequence database is necessary for VQA research and deserves more efforts. Thirdly, in this paper, NQM seems to be a premium metric to predict the rate saving of H.265/HEVC over H.264/AVC, especially for sequences with high spatial complexities. More tests are needed to verify our findings.

VI. CONCLUSIONS

This paper employs the rate saving of H.265/HEVC over H.264/AVC to evaluate the performance of 13 objective image and video quality assessment (IQA/VQA) metrics. The experimental results show that most metrics, including recently proposed ones, lack consistency with subjective evaluation, especially for video sequences with high levels of spatial complexities. In contrast, a relatively old metric—NQM gives relatively accurate and stable prediction. More comprehensive evaluation of NQM is underway.

ACKNOWLEDGMENT

This work was supported by NSFC (61221001, 61420106008), National Key Technology R&D Program of China (2013BAH53F04, 2012AA01170), the 111 Project (B07022) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

REFERENCES

[1] L. Zhang, X. Mou et at. "FSIM: a feature similarity index for image quality assessment," IEEE Trans. IP, vol. 20, pp. 2378-2386, 2011.

[2] W. Xue, L. Zhang and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," IEEE Trans. Image Process., vol.23, no.2, Feb. 2014.

[3] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," IEEE Trans. IP, vol. 20, pp. 1185-1198, 2011.

[4] N. Damera-Venkata, T. D. Kite, W. S. Geisler, et al. "Image quality assessment based on a degradation model," IEEE Trans. Image Process., vol.9, no.4, pp. 636–650, Sep. 2002.

[5] T. K. Tan, A. Fujibayashi, J. Takiue, "AHG8: Objective and subjective evaluation of HM5.0," JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/San jose, CA, 2012.

[6] J.-R. Ohm, G. Sullivan et al. "Comparison of the coding efficiency of video coding standards-including high efficiency video coding(HEVC)," IEEE Trans. Circuits and Systems for Video Technology, Dec. 2012.

[7] Y. Zhao, L. Yu, "Coding efficiency comparison between HM5.0 and JM16.2 based on PQI, PSNR and SSIM," JCTVC- H0063.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, et al. "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp. 600–612, Apr. 2004.

[9] Y. Zhao, L. Yu, Z. Chen, and C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," IEEE Trans. Circuits and Systems for Video Technology, vol. 21, no. 12, pp. 1890-1902, Dec. 2011.

[10] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," VCEG Contribution VCEG-M33, Austin, Apr. 2001.

[11] K. Zeng, A. Rehman, J. Wang, Z. Wang, "From H.264 to HEVC: Coding Gain Predicted by Objective Video Quality Assessment Models", VPQM. Scattsdale, AZ, USA, Jan.-Feb. 2013.

[12] G. Bjøntegaard, "Improvements of the BD-PSNR model," ITU-T SG16 Q.6 Document, VCEG-AI11,Berlin,Germany, July 2008.

[13] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process., vol.15, no.2, pp. 430–444, Feb. 2006.

[14] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," IEEE Trans. Image Process., vol.16, no.9, pp. 2284–2298, Sep. 2007.

[15] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," IEEE Trans. Image Processing, vol.19, no. 2, pp. 335–350, Feb. 2010.

[16] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in Proc. IEEE Asilomar Conf. Signals, Syst., Comput., Pacific Grove, CA, pp. 1398–1402, Nov. 2003.

[17] H. R. Sheikh, A. C. Bovik and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," IEEE Trans. Image Processing, vol.14, no.12, pp. 2117–2128, Dec. 2005.

[18] Z. Wang, A. C. Bovik. "A universal image quality index," IEEE Signal Process. Lett, vol.9, no.3, pp. 81–84, Mar. 2002.

[19] T. Mitsa, K. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms", IEEE International Conference on Acoustic, Speech, and Signal processing, Vol. 5, pp. 301- 304, 1993.

[20] H. R. Sheikh, M. F. Sabir and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Processing, vol.15, no.11, pp. 3441–3452, Nov. 2006.

[21] M. Gaubatz, Metrix Mux Visual Quality Assessment Package, [Online]. Available: http://foulard.ece.cornell.Edu/gaubatz/metrix_mux/.

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