2013 *International Workshop on Video Processing and Quality Metrics for Consumer Electronics* (VPQM2013) Scottsdale, AZ, USA, Jan.-Feb. 2013.

FROM H.264 TO HEVC: CODING GAIN PREDICTED BY OBJECTIVE VIDEO QUALITY ASSESSMENT MODELS

Kai Zeng, Abdul Rehman, Jiheng Wang and Zhou Wang

Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
Emails: {kzeng, abdul.rehman, jiheng.wang, zhou.wang}@uwaterloo.ca

ABSTRACT

Significant progress has been made recently towards the next generation video coding standard by the Joint Collaborative Team on Video Coding (JCT-VC). Recently reported preliminary subjective tests, conducted by JCT-VC members, show that the test model of High Efficiency Video Coding (HEVC) draft codec HM5.0 achieves an average of more than 50% rate savings over H.264 JM18.3 codec without sacrificing subjective quality. Here we study the performance of well-known objective video quality assessment (VQA) models and find that state-of-the-art models, including the Structural Similarity (SSIM), the Multi-Scale SSIM index (MS-SSIM), the Video Quality Metric (VQM), and the MOtion-based Video Integrity Evaluation index (MOVIE), all provide significantly better predictions of subjective video quality than peak signal-to-noise ratio (PSNR). Surprisingly, compared with subjective evaluation scores, all objective VQA models systematically underestimate the coding gain of HEVC-HM5.0 upon H.264-JM18.3. We carried out further subjective tests to study this somewhat unexpected phenomenon by comparing JM18.3 and HM5.0 coded videos in terms of frame-level and sequence-level quality, as well as flickering and ghosting effects. The results provide new insights for the future development of subjective/objective VQA and perceptually-tuned video coding methods.

1. INTRODUCTION

Since the official joint Call for Proposals (CfP) [1] on the next generation video compression standard was announced in January 2010 by ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG), the Joint Collaborative Team on Video Coding (JCT-VC) has made significant progress in developing the test model, known as High Efficiency Video Coding (HEVC), which targets at reducing 50% bit-rate of the MPEG4/H.264 AVC standard while maintaining the same level of subjective quality. Recently, a preliminary subjective test was conducted by JCT-VC members to quantify the rate-distortion (RD) gain of the HEVC draft codec HM5.0 against a similarly-configured H.264/AVC JM18.3 codec [2]. The results show that an average RD-gain of 57.1% is achieved based on the subjective test data in the form of Mean Opinion Scores (MOSs). A more detailed objective and subjective evaluation of HM5.0 was reported in [3], which again suggested that HM5.0 has achieved the target of 50% RD gain over H.264/AVC and the actual savings can be even higher. Although these subjective tests and evaluations were on random access coding configuration only and more comprehensive tests are still yet to be conducted, it is speculated that similar improvement may be achieved in other test conditions, and thus HEVC is very promising at achieving its initial RD performance target.

While subjective quality assessment is essential in fully validating the performance of video codecs, it is also highly desirable to know how the existing objective image and video quality assessment (IQA/VQA) models predict the subjective test results and the coding performance. In the past decades, objective IQA/VQA has been an active research topic, which aims to automatically predict perceived image and video quality of human subjects. They are useful in real world applications to control and maintain the quality of image/video processing and communication systems on the fly, where subjective quality assessment is often too slow and costly. They may also be embedded into the design and optimization of novel algorithms and systems to improve perceived image/video quality. Compared with IQA, VQA is a much more challenging problem because of the additional complications due to temporal distortions and our limited understanding about motion perception and temporal visual pooling. Traditionally, peak signal-to-noise ratio (PSNR) has been used as the "default" criterion in the video coding community in the design, validation and comparison of video codecs. Although PSNR is widely criticized for its poor correlation with perceived image quality and many perceptual objective IQA/VQA models have been proposed in the literature [4], currently PSNR is still the primary objective quality reference in codec development (such as HEVC) mostly by convention.

Given the subjective test data in the form of MOSs collected by JCT-VC members that compare H.264-JM18.3 and HEVC-HM5.0 [2], here we reexamine well-known ob-

Table 1. Performance Comparison of PSNR, VQM, MOVIE, SSIM and MS-SSIM

VQA Model	PLCC	MAE	RMS	SRCC	KRCC	Computational Complexity (normalized)	RD-gain (Class B)	RD-gain (Class C)	RD-gain (Average)
PSNR	0.5408	1.1318	1.4768	0.5828	0.3987	1	-45.0%	-34.1%	-39.6%
VQM [5]	0.8302	**0.7771**	0.9768	0.8360	0.6243	1083	-43.1%	-31.9%	-38.6%
MOVIE [9]	0.7164	0.9711	1.2249	0.6897	0.4720	7229	-36.4%	-25.1%	-33.8%
SSIM [6]	0.8422	0.8102	0.9467	0.8344	0.6279	5.874	-45.5%	-32.8%	-39.2%
MS-SSIM [8]	**0.8526**	0.7802	**0.9174**	**0.8409**	**0.6350**	11.36	-46.8%	-34.6%	-40.7%
MOS	-	-	-	-	-	-	-66.9%	-47.2%	-57.1%

jective VQA algorithms emerged in the past decade by observing how well they predict the subjective scores of compressed video sequences and how well they predict the RD-gain between HEVC-HM5.0 and H.264-JM18.3. Moreover, we carry out further subjective tests to exploit the relationship between frame-level and sequence-level subjective quality, and to investigate special temporal coding artifacts created by standard video codecs. This study may help the video coding community select useful VQA models in their future validation and comparison of novel video codecs, may provide new insights about the perceptual aspects of H.264 and HEVC coding schemes and how they may be further improved, and may also help VQA researchers discover the problems in the current subjective testing methodologies and objective VQA models and find ways to improve them.

2. TEST OF OBJECTIVE VIDEO QUALITY ASSESSMENT MODELS

Five existing objective VQA models are being examined, which include PSNR, the video quality metric (VQM) [5], the structural similarity index (SSIM) [6, 7] (As in [6], a preprocessing step of spatial downsampling by a factor of 2 is applied to each frame before the SSIM index is computed), the Multi-Scale SSIM index (MS-SSIM) [8], and the MOtion-based Video Integrity Evaluation index (MOVIE) [9]. All five models are well-known in the IQA/VQA and video coding communities. In the subjective data given in [2], a total of 72 HM5.0 and JM18.3 compressed video sequences were tested, which were generated from 9 original source video sequences, including 5 Class B sequences of 1080p resolution (1920×1080) and 4 Class C sequences of WVGA resolution (854×480). The encoding configuration of HM5.0 was set as random-access high-efficiency (RA-HE), and for fair comparison, the JM18.3 configuration was adjusted accordingly to best match that of HM5.0. No rate control scheme has been applied to either JM18.3 or HM5.0 encoding. The specific details of coding configurations can be found in [2]. The subjective test results were recorded in the form of MOS of each test video sequence.

Seven criteria were used to evaluate each objective VQA model. These include (1) Pearson linear correlation coefficient (PLCC), (2) mean absolute error (MAE), and (3) root mean square (RMS), which are computed after a nonlinear modified logistic fitting [10] between the MOS values and the scores given by the objective model. Two rank-order based evaluation measures, namely (4) Spearman rank-order correlation coefficient (SRCC) and (5) Kendall rank-order correlation coefficient (KRCC), between objective and subjective quality scores are also computed, which are independent of any fitting function that attempts to align the scores. The premium performance of objective quality models is represented by higher PLCC, SRCC, and KRCC, and lower MAE and RMS values. Since speed is often a major concern in real-world applications of VQA models, we also compared (6) the computational complexities of the VQA models, which are reported as their relative computation time normalized by the computation time of PSNR (this should be considered as only a crude estimate of the computational complexities of the VQA models because no algorithm and/or code optimization has been conducted to accelerate the speed). Finally, (7) the RD-gain of HM5.0 over JM18.3 is estimated for each source video sequence by comparing the RD curves of HM5.0 and JM18.3, where R denotes bit-rate and D denotes the distortion measure based on the specific VQA model and each RD curve is created by piecewise linear interpolation of the rate and distortion values of four coded video sequences generated by the same coding scheme [2]. The average RD-gain of HM5.0 over JM18.3 is then computed as the average of the RD-gains of all source videos.

The average performance of the objective models over all test video sequences are summarized in Table 1, where the best performances are highlighted with bold face. The scatter plots of objective scores versus MOSs are shown in Fig. 1. From Table 1 and Fig. 1, it can be observed that all four state-of-the-art VQA models clearly outperform PSNR in terms of PLCC, MAE, MSE, SRCC and KRCC, where on average MS-SSIM obtains slightly better results than the other three. On the other hand, VQM and MOVIE are ex-

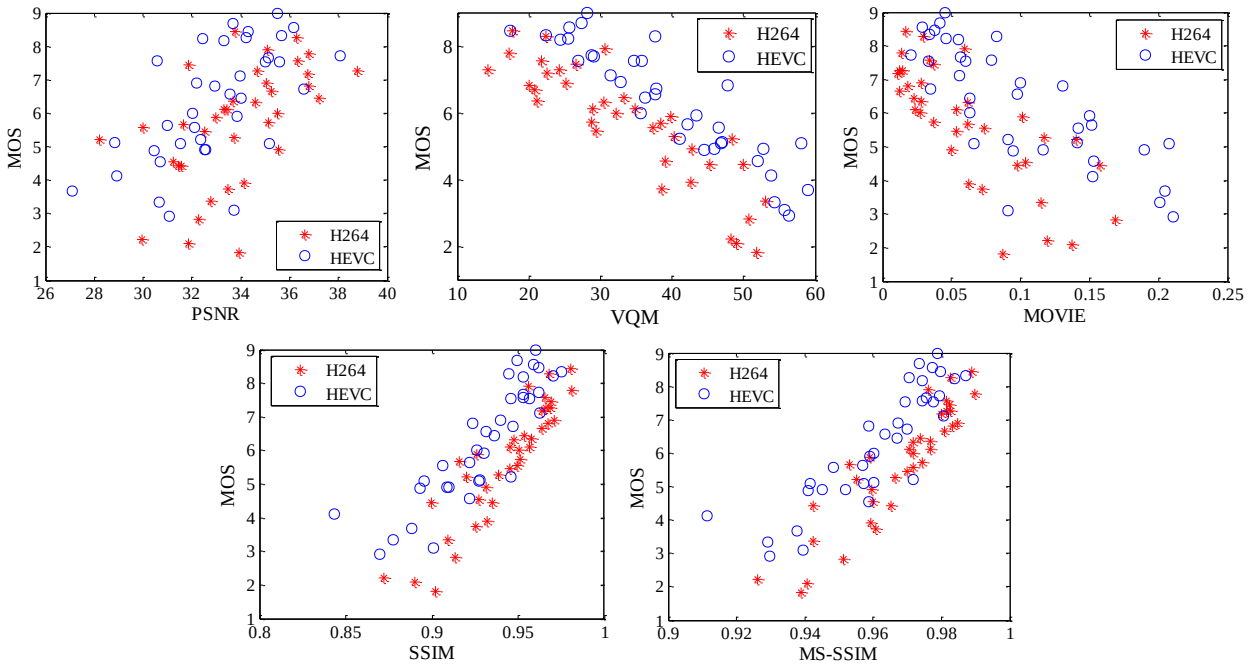

Fig. 1. Scatter plots of VQA measure vs. MOS.

tremely expensive in computational cost, while SSIM and MS-SSIM achieves a much better balance between quality prediction accuracy and computational complexity. Table 2 reports the paired statistical significance comparison using the approach introduced in [11], where a symbol "1" denotes the objective model of the row is statistically better than that of the column, "0" denotes that the column model is better than the row model, and "–" denotes that the two objective models are statistically indistinguishable.

Table 2. Statistical significance test for PSNR, MOVIE, VQM, SSIM and MS-SSIM

	PSNR	MOVIE	VQM	SSIM	MS-SSIM
PSNR	–	–	0	0	0
MOVIE	–	–	–	0	0
VQM	1	–	–	–	–
SSIM	1	1	–	–	–
MS-SSIM	1	1	–	–	–

Perhaps the most surprising results here is in the RD-gain columns in Table 1 – the five objective VQA models predict the average RD-gain of HM5.0 against JM18.3 to be between 33.8% to 40.7% , which largely underestimates the 57.1% gain obtained from subjective scores. Similar behaviors are also observed for individual test classes. This suggests that all objective VQA models are systematically in favor of H.264 JM18.3 while human subjects tend to prefer HEVC HM5.0. This can also be seen in Fig. 1, where in all scatter plots, the clusters of HM5.0 and JM18.3 coded video sequences are visually separated (though with overlaps), and HM5.0 sequences tends to have higher MOS values. Fig. 2 provides an example using 1080p "Parkscene" seuqnece, where we can observe how subjective and objective video quality measures change as a function of bit rate. Again, it can be seen that the gap between the HM5.0 and JM18.3 MOS-rate curves is significantly larger than those of the PSNR-rate and (MS-SSIM)-rate curves. Similar phenomena had been observed partially in previous studies. In [3], it was reported that PSNR accounts for 39% rate savings of HM5.0 over JM18.3, as compared to more than 50% by human subjective scores. Similar results are also found in [12]. In [13], the coding performance of HM5.0 and JM16.2 was compared under the RA-HE test conditions over 15 test sequences in terms of perceptual quality index (PQI) [14], PSNR and SSIM [6], and the results showed that the predicted RD-gain by all VQA models are almost the same.

3. SUBJECTIVE STUDY OF SPATIAL AND TEMPORAL VIDEO QUALITY

To better understand the significant bias of objective VQA models towards H.264-JM18.3 as opposed to HEVC-HM5.0, we carried out a series of subjective experiments to inspect the quality of coded video sequences at both frame and sequence levels. Ten compressed sequences (5 by JM18.3

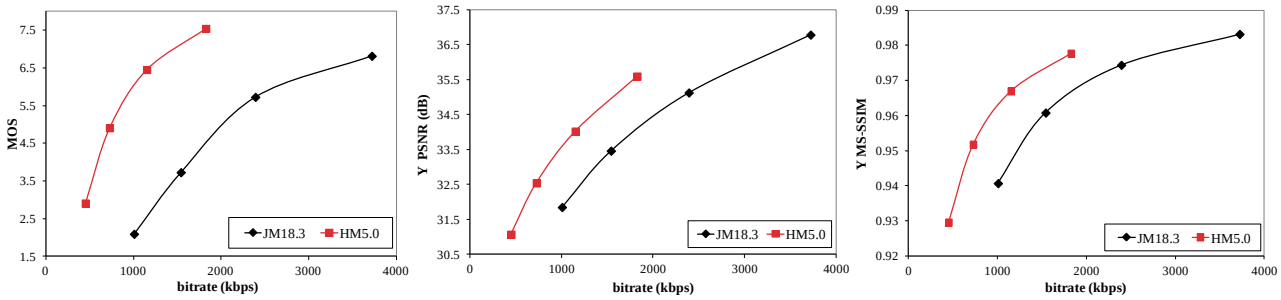

Fig. 2. Rate-quality comparison of JM18.3 and HM5.0 compressed 1080p "ParkScene" sequence, where the quality measures are MOS (left), PSNR (middle) and MS-SSIM (right), respectively. The RD-gain of HM5.0 upon JM18.3 computed using MOS, PSNR, and MS-SSIM are -63.6%, -36.8%, and -39.4%, respectively.

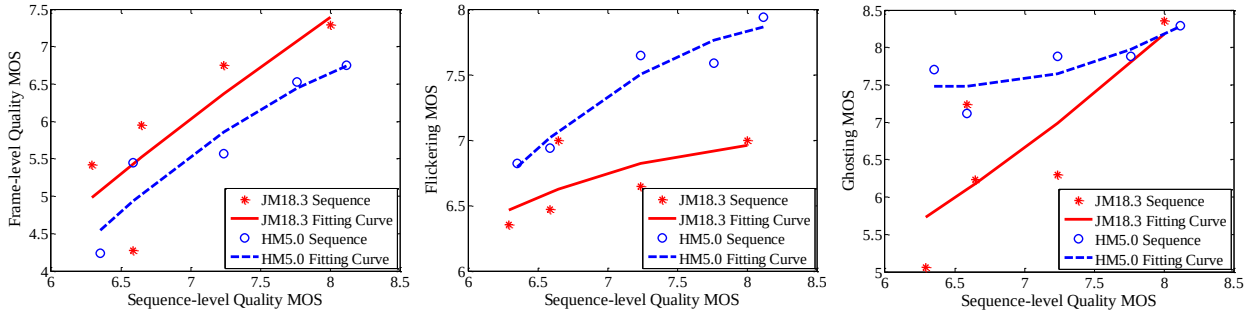

Fig. 3. Relationship between subjective test results for JM18.3 and HM5.0 coded sequences. Left: sequence-level MOS vs. average frame-level MOS; middle: sequence-level MOS vs. flickering MOS; right: sequence-level MOS vs. ghosting MOS.

and 5 by HM5.0) were selected and 5 frames were chosen randomly from each sequence, resulting in totally 50 still image frames. 17 naïve observers participated in the subjective assessment session. The test method conforms with ITU-T BT.500 [16]. Absolute categorical rating (ACR) was adopted to collect the mean opinion score (MOS) which is the average of subjective opinion from all observers. Four tests have been carried out. The first test is to assess frame-level image quality, where the subjects give scores regarding the quality of the 50 individual still image frames. The second test is on sequence level, where the subjects report a single score for each test video sequence. In the third and the fourth tests, the subjects are asked to evaluate the flickering and ghosting effects of the test video sequences, where flickering refers to the discontinuities of local average luminance over time, and ghosting refers to the traces of video content in previous frames that are remained in the current frame (often created by the Skip mode in the video codec).

From our subjective test, we have the following observations. First, there are significant conflicts between frame-level and sequence-level quality assessment. This can be seen from the left plot in Fig. 3, where frame-level MOSs (computed by averaging all still frame MOS values of a sequence) and sequence-level MOSs obtained in our subjective experiment do not correlate well with each other. In addition, there is a clear tendency that HM5.0 coded videos obtain higher sequence-level MOSs and lower frame-level MOSs in comparison with JM18.3. A visual example is shown in Fig. 4, which shows a still frame extracted from a JM18.3 and an HM5.0 coded "Horse" sequences. On a high quality monitor, the JM18.3 frame appears to better preserve the image details and thus has better quality. The same phenomenon has been observed in all frames throughout the whole video sequences. By contrast, the sequence-level MOS of the HM5.0 video is significantly higher than that of the JM18.3 video. This observation, combined with the fact that frame-based objective VQA measures often well predicts frame-level MOS (in our experiment, the SRCC between still frame MOS and MS-SSIM is 0.8627), provides an explanation for why objective VQA tends to underestimate sequence-level subjective quality. Second, significant annoying temporal artifacts may appear in coded video sequences that may dominate subjective evaluation of video quality. We have included flickering and ghosting assessment in our subjective tests. The scatter plots of sequence-level MOS versus flickering and ghosting are shown in the middle and right plots of Fig. 3, respectively, where higher flickering or ghosting MOS indicates less flickering or ghosting effect. From these plots, we observe that JM18.3 coded sequences have clearly stronger flickering and ghosting effects than HM5.0 sequences. This is in clear contrast to the left plot in Fig. 3 and provides strong support of the

H.264/AVC HEVC

Fig. 4. An example of visual comparison between H.264-JM18.3 and HEVC-HM5.0 coded videos. Left: H.264 frame, PSNR = 28.36dB, SSIM = 0.8012, MS-SSIM = 0.8601. Right: HEVC frame, PSNR = 27.64dB, SSIM = 0.7437, MS-SSIM = 0.8259. When comparing individual frames, H.264 frame appears to have clearly better visual quality, but when the video is played at normal speed, the H.264 video receives a significantly lower quality score likely due to strong temporal artifacts.

conjecture that compared with frame-level quality, temporal artifacts contribute strongly to the overall sequence-level quality. Third, there is significant spatial and temporal quality non-uniformity of coded video sequences. Such non-uniformity is partially predicted by the objective VQA models (for example, using the SSIM maps) and is more evident in JM18.3 coded video sequences.

The observations above give us useful insights to address several issues in subjective tests. First, the past experience of the subjects and the context of the subjective experiment need to be better taken into account. Second, questions may be asked to the subjects about what strategies they use to make an overall decision on an entire video sequence that has significant quality non-uniformity over space and/or time. Third, it is desired to record eye movement in the subjective experiments. The importance is not only to detect the regions of interest (ROIs) in the video content, but also to study whether compression artifacts change eye fixations and how the context (e.g., tasks given to the subjects) affects visual attention – are the subjects trying to understand the story of the video content or to detect the distortion artifacts? Previous studies suggest that compression artifacts generally have little impact on visual attention [15], but is this still true when extremely annoying artifacts occur?

4. FURTHER DISCUSSIONS

The observations in the current study raise new questions that need to be answered in the development of objective VQA models. First, there is a strong need to develop novel approaches to capture specific temporal artifacts (such as flickering and ghosting) in compressed video. PSNR, SSIM and MS-SSIM are completely IQA methods where no inter-frame interactions are considered. It is not surprising that temporal artifacts are missing in these models. However, both VQM and MOVIE consider temporal features, but are still not fully successful in capturing and penalizing the temporal artifacts. Second, many VQA models such as SSIM and MS-SSIM generates useful quality maps that indicate local quality variations over space and time. In the case of significant spatial and temporal non-uniformity in these quality maps, how to pool the maps into a single quality score of the entire video is not a fully resolved problem. There has been attempts to use non-linear model and temporal hysteresis for temporal pooling [17, 18]. However, our current test shown in Table 3 indicates that they only lead to small improvement over MS-SSIM and the large gap between subjective and objective RD-gain predictions still exists. Third, it would be useful to incorporate visual attention models. These attention models may be saliency predictors based on both low-level and high-level vision features, and may also be based on detections of severe visual artifacts.

Meanwhile, what we learned from this study may help us improve the design and implementation of video coding technologies. It is useful to be aware of and to avoid certain temporal artifacts such as flickering and ghosting effects, which may vastly change subjects' opinions about the quality of the entire video sequence. Many of these artifacts occur when quantization parameters are not carefully chosen and when Skip mode is selected in low- to mid-energy regions with slow motion. Moreover, rate control and rate-distortion optimization (RDO) schemes may be adjusted not only to achieve the best average quality over the whole video sequence, but also to reduce significant quality fluctuations across both space and time.

In conclusion, our study shows that advanced VQA models clearly outperform PSNR in predicting the quality scores given by human subjects. This suggests that the video coding community and the standard development body may consider replacing PSNR with perceptually more meaning-

Table 3. The impact of temporal pooling strategies on MS-SSIM method

VQA Model	PLCC	MAE	RMS	SRCC	KRCC	RD-gain (Average)
MS-SSIM [8]	0.8526	0.7802	0.9174	0.8409	0.6350	-40.7%
MS-SSIM with min temporal pooling [17]	0.8670	0.6859	0.8749	0.8645	0.6663	-43.2%
MS-SSIM with temporal hysteresis pooling [18]	0.8544	0.7498	0.9123	0.8467	0.6400	-42.4%
MOS	-	-	-	-	-	-57.1%

ful VQA models in not only the testing but also the development phases of novel video codecs. This could lead to substantial changes in the structural design and system optimization of the next generation video codecs. In terms of RD-gain predictions, however, none of the objective VQA models aligns well with the subjective test results. We conjecture that this may be due to one (or the combination) of several issues, including ambiguities in subjective testing methodologies, limitations of the current VQA models in capturing specific types of temporal artifacts (such as flickering and ghosting), and the lack of good spatiotemporal pooling strategies and both saliency and artifacts based visual attention models. The current discussions are non-conclusive but may inspire future improvement in both VQA and video coding methodologies.

5. REFERENCES

[1] ITU-T Q6/16 Visual Coding and ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio, "Joint Call for Proposals on Video Compression Technology," MPEG Document N11113, Jan. 2010, Kyoto, JP.

[2] V. Baroncini, J. R. Ohm, G. J. Sullivan, "Report on preliminary subjective testing of HEVC compression capability," *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, Feb., 2012.

[3] 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/WG11*, San José, CA, 2012.

[4] Z. Wang and A. Bovik, "Mean squared error: love it or leave it? - a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98–117, Jan. 2009.

[5] M. H. Pinson, "A new standardized method for objectively measuring video quality," *IEEE Tans. Broadcasting,* vol. 50, no. 3, pp. 312-322, Sept. 2004.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.,* vol. 13, no. 4, pp. 600-612, Apr. 2004.

[7] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, pp. 121–132, Feb. 2004.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Asilomar Conf. Signals, Systems and Computers,* Nov. 2003.

[9] 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.

[10] 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 Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[11] 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. 3440–3451, Nov. 2006.

[12] B. Li, G. J. Sullivan, J. Xu, "Comparison of compression performance of HEVC working draft 5 with AVC high profile," *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, 1-10 Feb., 2012.

[13] Y. Zhao and L. Yu, "Coding efficiency comparison between HM5.0 and JM16.2 based on PQI, PSNR and SSIM," *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, 1-10 Feb., 2012.

[14] 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.

[15] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 597-609, Sep. 2010.

[16] ITU-R BT.500-12, "Recommendation: Methodology for the subjective assessment of the quality of television pictures", Nov. 1993.

[17] C. Keimel, and K. Diepold, "Improving the prediction accuracy of PSNR by simple temporal pooling," 5th *Int. Workshop on Video Proc. and Quality Metrics for Consumer Electronics*, vol. 2009, 2010.

[18] K. Seshadrinathan, A.C. Bovik, "Temporal hysteresis model of time varying subjective video quality," *IEEE Inter. Conf. Acoust., Speech & Signal Proc.* pp.1153-1156, May 2011.