EXTENDING THE VALIDITY SCOPE OF ITU-T P.1202.2

Lark Kwon Choi[1], Yiting Liao[2], Barry O'Mahony[2], Jeffrey R. Foerster[2], and Alan C. Bovik[1]

[1]Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA
[2]Intel Labs, Intel Corporation, Hillsboro, OR, USA
larkkwonchoi@utexas.edu, {yiting.liao, barry.omahony, jeffrey.r.foerster}@intel.com,
bovik@ece.utexas.edu

ABSTRACT

We extend the validity scope of ITU-T Recommendation (Rec.) P.1202.2, *Parametric Non-Intrusive Bitstream Assessment of Video Media Streaming Quality - Higher Resolution Application Area* on the Intel and Technische Universität München (TUM) video quality assessment (VQA) databases (DB). To find a new use case for the Rec. P.1202.2 mode 1 compression artifact VQA model on dynamic adaptive streaming over HTTP (DASH), we investigated its possibility under a wide range of dataset classifications including different content types (spatial and temporal complexity), encoding profiles (Main and High), and device sizes (on the Intel VQA DB: HDTV, TFT tablet, AMOLED phone, Retina* tablet, and Retina phone; on the TUM VQA DB: monitor, HDTV, and projector). Results show that the Rec. P.1202.2 mode 1 compression artifact VQA model tends to overestimate low quality compressed videos and can be improved by taking into account quality variations on different display devices or under different encoding profiles. Hence, we propose some guidelines to calibrate the Rec. P.1202.2 mode 1 compression artifact VQA model.

1. INTRODUCTION

Fast, reliable, and accurate monitoring of perceptual video quality has become more pressing as streaming video services such as Netflix*, Hulu*, and Amazon* instant video proliferate; live online video services including Skype* and Google+ Hangouts* are expanding rapidly as well [1]. Despite intensive research on VQA models during the past two decades, it still remains, especially in real time video applications, a challenging problem. This is principally due to two issues: limited or no availability of reference videos and computational complexity.

With the availability of reference video datasets, research on objective VQA algorithms has advanced from full-reference metrics such as VQM [2] and MOVIE [3] and reduced-reference models such as V-RRED [4], to no-reference (NR) methods like DIIVINE [5], BRISQUE [6], and V-BLIINDS [7]. Although these VQA models excellently capture perceptual video quality, the encoded bitstream needs to be decoded before the models can be applied. Even if a VQA metric can be simple, a decoding process increases the overall computational complexity.

To predict video quality without complete decoding or pixel reconstruction from the bitstream, bitstream-based VQA algorithms have been a focus of research and standardization activities. Eden estimated the PSNR of H.264 HDTV videos from bitstream features [8]. Yang *et al.* proposed a temporal pooling of frame quality obtained from spatial and temporal complexities of each frame [9]. Keimel *et al.* suggested a data analysis approach with partial least squares regression for visual quality [10]. Staelens *et al.* constructed white box models by using genetic programming based symbolic regression [11].

Study Group 12 of the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) has recently approved ITU-T Rec. P.1202.2, *Parametric Non-Intrusive Bitstream Assessment of Video Media Streaming Quality - Higher Resolution Application Area* [12]. The Rec. P.1202.2 is a very useful VQA model for IPTV usages since it takes into account multiple artifacts in those scenarios, and it is a NR metric. In addition, the Rec. P.1202.2 can be implemented in real-time applications thanks to its simplicity. Its intended scope includes fixed-rate IPTV services in the range of 0.5 to 30 Mbps, evaluated with display characteristics as specified in ITU-R BT.500-11 [13]. The text describes performance results: the overall Pearson's linear correlation coefficient (LCC) is 0.938, with a root mean square error of 0.357 on 3069 videos with compression, slicing, and freezing artifacts.

Based on the advantages of Rec. P.1202.2, we tried to find a new use case for the Rec. P.1202.2 mode 1 compression artifact VQA model pursuing HTTP based adaptive streaming algorithms such as Dynamic Adaptive Streaming over HTTP (DASH). Although the Rec. P.1202.2 algorithm can provide a final video quality score

* All brands and names are property of their respective owners.

on a multiple distorted video with compression, slicing, and freezing artifacts, since DASH uses TCP to ensure an error-free transmission, in this paper, we consider only the compression artifact VQA model in the Rec. P.1202.2 mode 1. Using a NR VQA metric focused on compression artifacts will help to determine the lowest bit-rate which could be streamed while providing a good quality of experience. In a DASH scenario, frame freezes can still occur due to running out of content in the buffer, and this should also be taken into account but is not discussed here.

The ITU-T VQA DB used to generate performance results is not publicly available, and the specific performance on compression artifacts is not broken out. This makes it difficult for us to utilize the Rec. P.1202.2 for HTTP based adaptive streaming models. Hence, to study the applicable use cases of the Rec. P.1202.2 mode 1 compression VQA model in a DASH scenario, we tested the compression artifact VQA model separately and extended its validity scope on the Intel and TUM VQA databases. The investigation of the algorithm performance on a wide range of dataset classifications including different device sizes, encoding profiles, and content types show that it is possible to extend the main concept of the Rec. P.1202.2 mode 1 compression artifact VQA model to design a NR metric for HTTP streaming while the algorithm needs more tuning to meet the performance requirement in practice. For example, the Rec. P.1202.2 mode 1 compression artifact VQA model tends to overestimate low quality compressed videos and does not consider quality disparity under encoding profiles; it is also evident that alternate display devices were not within the scope of its Terms of Reference. The analysis of the model performance on the Intel and TUM databases leads us to propose some guidelines to improve the Rec. P.1202.2 mode 1 compression artifact VQA model.

The remainder of this paper is organized as follows. Section II summarizes the P.1202.2 mode 1 compression artifact VQA model. The detail analysis of the algorithm performance on the Intel and TUM VQA databases are shown in Section III. Section IV proposes some guidelines to calibrate the Rec. P.1202.2 mode 1 compression artifact VQA model. Concluding remarks are drawn in Section V.

2. ITU-T REC. P.1202.2

ITU-T Rec. P.1202.2 recommended objective models for non-intrusive monitoring of the video quality of IP-based video services based on packet-header and bitstream information. It specifies the VQA model algorithm for the higher resolution application area, which includes services such as IPTV [12]. Rec. P.1202.2 consists of two modes: mode 1, where the video bitstream is parsed and not decoded into pixels, and mode 2, where the video bitstream is fully decoded into pixels for analyzing. Both modes, as output, provide an estimate of the video quality in terms of the 5-point absolute category rating (ACR) mean opinion score (MOS) scale defined in ITU-T Rec. P.910 [14]. Each mode consists of compression, slicing, freezing, and combination artifacts modules. We introduce only a relevant portion of the ITU-T Rec. P.1202.2 mode 1 compression artifact VQA model to find a use case on a DASH scenario. For more information regarding other modules of the Rec. P.1202.2, please refer to [12].

The Rec. P.1202.2 mode 1 takes an H.264/AVC encoded video bitstream as input, extracts basic parameters, aggregates them into module parameters, then estimates MOS for the video sequence. From bitstream at picture level, the correctly decoded slice quantization parameter (*QP*) of the k^{th} slice, $slice\_QP_k$, the number of bytes of the k^{th} slice, $slice\_size_k$, and the number of pixels of the k^{th} slice, $slice\_pixel_k$, are extracted as basic parameters, where k is from 1 to the number of slice(s) of the *error free* Intra frame. The word *error free* means *no packet losses occurred*. In a compression artifact module, for each *error free* Intra frame, the *frame content complexity* is computed by averaging the value of *slice content complexity* of all slices in its frame. The k^{th} *slice content complexity* is calculated as below:

$$slice\,content\,complexity_k = a \times \frac{slice\_size_k}{slice\_pixel_k} + b, \qquad (1)$$

where a and b are arrays holding coefficient values for each $slice\_QP_k$ at each video resolution, respectively [12]. Next, *video content complexity* is obtained by averaging the value of *frame content complexity* of the *error free* Intra frame of the sequence. In addition, *video_QP* is computed by averaging all *slice_QP* values of *error free* Intra frames as follows:

$$video\_QP = \frac{The\,sum\,of\;slice\_QP}{The\,total\,number\,of\;slice}. \qquad (2)$$

Finally, the estimated MOS, *compression quality value*, is obtained by using *video_QP* and *normalized video content complexity* as below:

$$normalized\,video\,content\,complexity = \min\left(1, \sqrt{\frac{video\,content\,complexity}{60}}\right), \qquad (3)$$

$$compression\,quality\,value = c_1 + \frac{c_2}{c_3 + \left(\frac{video\_QP}{c_4 - c_5 \times normalized\,video\,content\,complexity}\right)^{c_6}}, \qquad (4)$$

where $c_{1\sim6}$ are coefficients defined in [12] for different video resolutions. The text does not describe how specific coefficients are obtained. When artifact is only caused by video compression, the *compression quality value* is output directly as the overall video quality score.

3. ALGORITHM PERFORMANCE

3.1. VQA databases

3.1.1. The Intel VQA database

The Intel VQA DB consists of fourteen source videos with a wide range of spatiotemporal complexity. They are 4:2:0 format, 1080p (1920 × 1080) at 25 or 30fps, and 10 ~ 15 sec duration, except Aspen Leaves (4s). The source videos are encoded from 110kbps at 448 × 252 to 6 Mbps at 1080p based on assumed realistic video content and display devices. Eighty compressed videos were displayed on a 42" HDTV; 96 compressed videos were displayed on four mobile devices (a 10.1" TFT tablet, a 9.7" AMOLED phone, a 4.8" Retina tablet, and a 3.5" Retina phone). About 30 subjects for each device rated the videos using the single-stimulus continuous quality evaluation (SSCQE) [13] method. Details are described in [15].

3.1.2. The TUM VQA database

The TUM VQA DB is composed of two different datasets: a 1080p 25fps dataset and a 1080p 50fps dataset. In the first dataset four source videos were encoded at four different bitrates between 5 ~ 30 Mbps and at two (Main and High) encoding profiles, yielding 48 data points. A total of 19 subjects participated in the test on a 24" LCD monitor using the double stimulus unknown reference (DSUR) method [16]. The second dataset contains five source sequences, compressed with H.264/AVC at 2 ~ 40 Mbps and at a High encoding profile resulting in 20 different data points. A 23"/24" LCD monitor, a 56" LCD TV, and a 2.8 meter projector were used for 19 subjects to rate videos using the single stimulus multimedia (SSMM) method [16]. For details, please refer to [16].

3.2. Evaluation of algorithm performance

3.2.1. Methodology

Peak Signal-to-Noise ratio (PSNR), Structural Similarity Index (SSIM) [17], Multi-Scale SSIM (MS-SSIM) [18], and the P.1202.2 mode 1 compression artifact VQA model were evaluated against the human subjective scores using the Pearson's linear correlation coefficient (LCC), the Spearman rank order correlation coefficient (SROCC), and the mean absolute error (MAE) on the Intel VQA DB. PSNR, SSIM, and MS-SSIM were applied on the luminance images of a frame-by-frame basis, and the final scores obtained for the video were the time-average of the frame-level quality scores. LCC evaluates the linear correlation and SROCC shows the monotonicity between the objective and subjective scores, while MAE represents prediction accuracy. MAE of PSNR, SSIM, MS-SSIM were obtained after non-linear regression [19]. On the TUM VQA DB we tested only H.264/AVC results (a total of 92 data points).

Since we are looking for a model that can be used in a real-time application in a DASH scenario and that has a good trade-off between performance and complexity, we compared the Rec. P.1202.2 mode 1 compression artifact VQA model with PSNR, SSIM, and MS-SSIM rather than other ITU standards such as VQM [2]. In a NR context using bitstream information only, since Keimel *et al.* already presented the high performance of their metric on the TUM DB [10], and since we are pursuing a new use case on the Rec. P.1202.2 mode 1 compression artifact model, not judging the performance of NR metrics, we investigated the performance of the Rec. P.1202.2 mode 1 compression artifact VQA model extensively under a wide range of dataset classifications including different device sizes, content complexities, and encoding profiles.

3.2.2. Results on the Intel VQA database

The Rec. P.1202.2 mode 1 compression artifact model shows the worst correlation to the subjective scores among the tested metrics as shown in Table 1 and does not monotonically correspond to MOS. For example, for the algorithm score of 4.2 in Figure 1, MOS varies widely from 1.5 to 5. Similarly, most of the predicted scores are above 4, while the actual MOS are in the range of 1.5 ~ 4. Although the P.1202.2 mode 1 compression artifact VQA model predicts high quality compressed videos well, it tends to overestimate low quality compressed videos.

Table 2 tabulates LCC, SROCC, and MAE between the algorithm scores and MOS for each device type. LCC is 0.5521 for all data points, while LCC using a device-based grouping is 0.5757, 0.7060, 0.7635, 0.6406, and 0.7813 for HDTV, TFT tablet, AMOLED phone, Retina tablet, and Retina phone, respectively. This implies that device-based classification can improve the prediction accuracy of the P.1202.2 mode 1 compression artifact model. Moreover, MAE results support that prediction accuracy increases as display size decreases. Similar results also can be found in Figure 2.

To understand content and device specific performance of the P.1202.2 mode 1 compression model we compared LCC, SROCC, and MAE for each content and device combination. As examples, Table 3 shows results between the algorithm scores and MOS on the high complexity content, "Aspen" and on the low complexity content, "Frontend" for each device. Content complexities are computed using ITU-T Rec. P.1202.2 [12] and ITU-T Rec. P.910 [14]. Results demonstrate that the P.1202.2 mode 1 compression artifact module almost predicts all the video bitstreams as good quality (4.2~4.4) for "Frontend," while it estimates a wider range for "Aspen" as shown in Figure 3. This implies that the current model prediction formula may require further tuning to improve prediction accuracy for low complexity contents.

Table 1. LCC, SROCC, and MAE between the algorithm scores and MOS on the Intel VQA DB.

	PSNR	SSIM	MS-SSIM	P.1202.2
LCC	0.6197	0.6423	0.7265	0.5521
SROCC	0.6253	0.6849	0.7548	0.6109
MAE	0.5102	0.5031	0.4278	0.7036

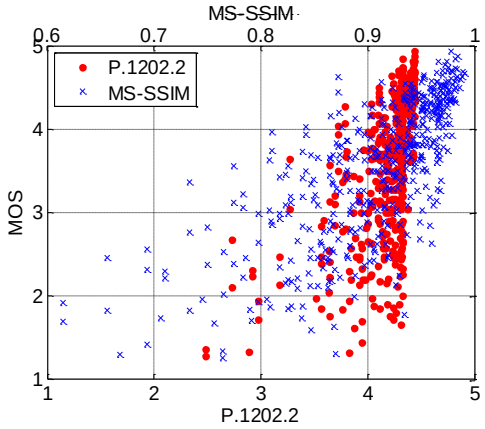


Figure 1. Scatter plot of P.1202.2 mode 1 compression module and MS-SSIM predictions against MOS on the Intel VQA DB.

Table 2. LCC, SROCC, and MAE between the algorithm scores and MOS for each device on the Intel VQA DB.

	HDTV	TFT-T	AMOLED-P	Retina-T	Retina-P
LCC	0.5757	0.7060	0.7635	0.6406	0.7813
SROCC	0.6472	0.7782	0.7619	0.7420	0.7313
MAE	1.0523	0.8485	0.6071	0.6772	0.3891

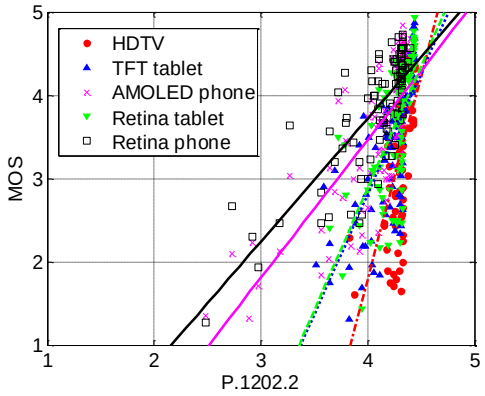


Figure 2. Scatter plot of P.1202.2 mode 1 compression module predictions against MOS for each device on the Intel VQA DB.

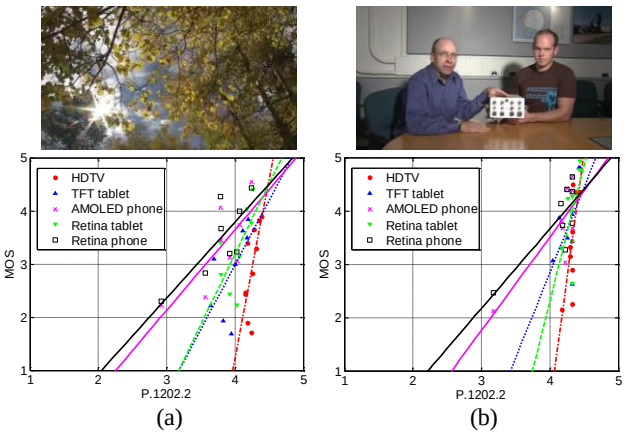
Table 3. LCC, SROCC, MAE between the algorithm scores and MOS for each device on (a) high complexity content, "Aspen" and (b) on low complexity content, "Frontend."

	HDTV	TFT-T	AMOLED-P	Retina-T	Retina-P
LCC	0.7076	0.6314	0.7718	0.6460	0.7889
SROCC	0.6930	0.6905	0.6667	0.6429	0.6905
MAE	1.3161	1.0867	0.6097	0.7936	0.4664

(a)

	HDTV	TFT-T	AMOLED-P	Retina-T	Retina-P
LCC	0.7646	0.5583	0.8195	0.5775	0.8153
SROCC	0.7178	0.4671	0.4910	0.4762	0.5476
MAE	0.8711	0.7183	0.5378	0.5834	0.4008

(b)



(a) (b)

Figure 3. Example frames and scatter plots with least-square liner fit of P.1202.2 mode 1 compression module scores against MOS for each device on (a) the high complexity content, "Aspen" and on (b) the low complexity content, "Frontend."

For all data points on 720p (720 × 1280) and 1080p video sequences, LCC, SROCC, and MAE between the algorithm scores and subjective MOS is 0.5334, 0.6173, and 0.3252, respectively. Device-based or content- and device-specific classifications yield similar results with above analyses.

3.2.3. Results on the TUM VQA database

Table 4 tabulates LCC, SROCC, and MAE between algorithm scores and MOS for each video frame rate and encoding profile combination as well as for all data points. LCC is 0.7294 for all data points, while LCC using frame rate and encoding profile combination is 0.8777, 0.8044, and 0.7743 for (25fps, Main), (25fps, High), and (50fps, High), respectively. In addition, most of algorithm scores are above 4 while the actual (subjective) MOS are in the range of 1.5 ~ 4.8 in Figure 4. The Rec. P.1202.2 mode 1 compression artifact model seems to do not take into account quality variations under different frame rate and encoding profiles.

Since a device-specific subjective study was performed only on the second dataset, we analyzed corresponding 50fps and High encoding profile data. LCC and SROCC are similar between algorithm scores and MOS for LCD monitor, LCD TV, and projector, while MAE increases for larger screen devices as shown in Table 5. When we compare all device-specific results on the Intel and TUM VQA DB, the P.1202.2 mode 1 compression model shows high prediction accuracy for the smaller screen devices. Figure 5 plots the algorithm scores against MOS for each device along with the best least-squares linear fit.

Regarding device and content specific combinations, as examples, Table 6 tabulates LCC, SROCC, and MAE between the algorithm scores and the subjective MOS on the high complexity content, "CrowdRun" and on the low

Table 4. LCC, SROCC, and MAE between the algorithm scores and MOS on the TUM VQA DB.

(fps,profile)	(25fps, Main)	(25fps, High)	(50fps, High)	All
LCC	0.8777	0.8044	0.7743	0.7294
SROCC	0.7835	0.8434	0.8315	0.7378
MAE	0.8191	0.3343	0.9980	0.8515

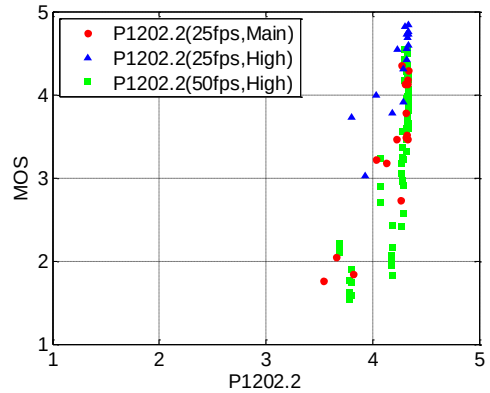


Figure 4. Scatter plot of P.1202.2 mode 1 compression module predictions against MOS for each frame rate and encoding profile on the TUM VQA DB.

Table 5. LCC, SROCC, and MAE between the algorithm scores and MOS for each device on the TUM VQA database.

	LCD Monitor	LCD TV	Projector
LCC	0.7995	0.7643	0.7895
SROCC	0.7865	0.8526	0.8650
MAE	0.8427	1.1328	1.0185

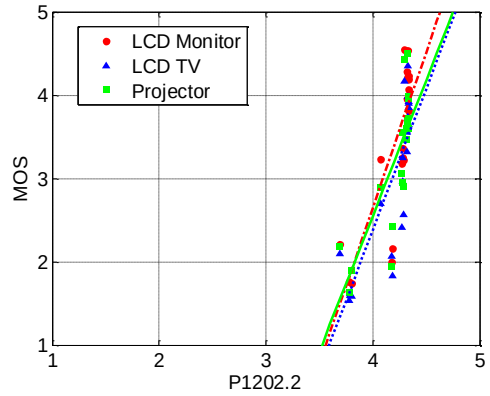


Figure 5. Scatter plot of P.1202.2 mode 1 compression module predictions against MOS for each device on the TUM VQA DB.

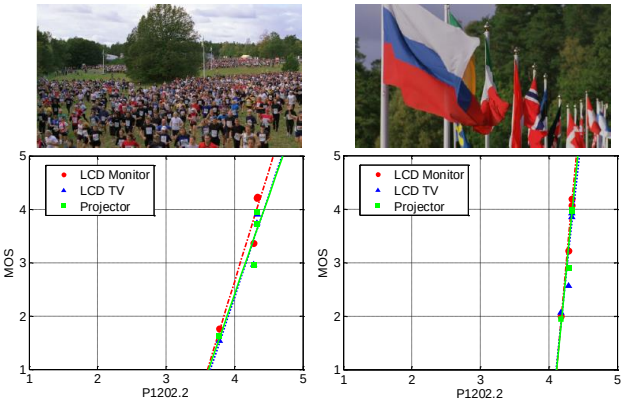
Table 6. LCC and MAE between the algorithm scores and MOS for each device on (a) the high complexity content, "CrowdRun" and on (b) the low complexity content, "FlagShoot."

	LCD Monitor	LCD TV	Projector
LCC	0.9623	0.9524	0.9438
SROCC	1.0000	1.0000	1.0000
MAE	0.7888	1.1448	1.1158

(a)

	LCD Monitor	LCD TV	Projector
LCC	0.9890	0.9114	0.9723
SROCC	0.8000	1.0000	0.8000
MAE	0.9168	1.1818	1.0858

(b)



(a) (b)

Figure 6. Example frames and scatter plots with least-square liner fit of P.1202.2 mode 1 compression module scores against MOS for each device on (a) the high complexity content, "CrowdRun" and on (b) the low complexity content, "FlagShoot."

complexity content, "FlagShoot" for each device. Although LCC and SROCC are high (over 0.91), MAE is also high due to the overestimation of low quality compressed videos. For other tested contents, results are similar. The algorithm predicts similar video quality on the "FlagShoot" rather than on the "FlagShoot" as shown in Figure 6.

3.2.4. Discussion of Results

Although the detailed performance depends on the tested VQA DB, overall results imply that the P.1202.2 mode 1 compression model tends to overestimate low quality compressed videos and can be improved by taking into account quality variations under different display devices, content complexities, frame rates, or encoding profiles. Regarding interdependencies between display devices and content complexities, although both factors are important, display devices seem to lead the algorithm performance on the tested VQA DB as can be seen in Tables 3 and 6. For better validation, the statistical analysis of variable significance and interdependencies is planned.

The discrepancy of the performance between the reported results on the Rec. P.1202.2 (e.g., LCC: 0.938) and the validated results (e.g., LCC: 0.5521 and 0.7294 on the Intel and TUM VQA DB, respectively) may result from the tested model and databases. The Rec. P.1202.2 shows only LCC obtained using a mode 1 combination model with compression, slicing, and freezing artifacts, while we specified a compression model only. The ITU-T VQA DB and the tested DB in this paper can have different quality ranges for compression artifacts. For example, the lowest quality in the compression DB can be of high quality in the ITU-T VQA DB. Other transmission impairments may dominate human scores when measured with compression artifacts, while our tests considered only compression artifacts. To fully understand the model performance cross experiment calibration is required.

Another speculation of the discrepancy of the performance can be display screen sizes. In ITU-T environments, the Rec. P.1202.2 mode 1 compression model was evaluated with display characteristics as specified in ITU-R BT-500 (e.g., the TV in the home environment), while a variety of display sizes from mobile form factors to a projector were used in the Intel and TUM subjective studies.

4. SUGGESTIONS

Based on the results shown in Section 3, we propose some guidelines to improve the P.1202.2 mode 1 compression artifact model. First, the coefficients $c_{1\sim6}$ in the model may need to be tuned to predict a wider range of MOS. As shown in Figures 1 and 4, the current model generally predicts a high MOS (> 4) for all H.264/AVC compressed videos in the tested VQA DB, while the actual MOS can be as low as 1.3. This misestimation happens frequently for low bitrate videos. Hence, calibrating the coefficients for those videos can help to improve prediction accuracy. This may require incorporating a greater degree of perceptual relevance into the current algorithm.

Secondly, the model may consider the impact of display device on video quality in the prediction process. With the growing capability of handheld devices, high-resolution video content targets not only big screen monitors but also small form factors such as tablets and smartphones across the compute continuum [20]. The subjective results have shown that for the same video played on different devices, human perceived quality is better for smaller devices. Therefore, we propose to add new sets of coefficients for laptop, tablet, and smartphone to more accurately predict perceived video quality for smaller form factors.

Thirdly, the model can be improved by adding a frame rate factor and coding profile relevant coefficients in the prediction formula. Two videos with different frame rates can be the same QP, bitrate, and resolution yielding the same P.1202.2 predicted quality, but actual perceived quality may be far apart because of the video frame rate. Low frame rate videos may cause choppy motion artifacts and degrade video quality. Since H.264/AVC provides different coding profiles that affect coding efficiency, including coding profile based coefficients would help to better account for the coding effects on video quality.

5. CONCLUSION AND FUTURE WORK

We investigated the performance of the P.1202.2 mode 1 compression artifact model on the Intel and TUM VQA databases. These two databases cover a variety of video contents with various levels of compression artifacts, and were displayed on different devices. Results show that the current P.1202.2 mode 1 compression artifact model can be further improved by tuning coefficients to enlarge the range of predicted video quality, and by considering more impact factors including display devices, content types, video frame rates, and coding profiles. For better model identification, extended cross validation is expected.

6. REFERENCES

[1] A. C. Bovik, "Automatic prediction of perceptual image and video quality," in *Proc. IEEE,* vol. 101, no. 9, Sep. 2013.

[2] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312-322, Sep. 2004.

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

[4] R. Soundararajan, and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684-694, Apr. 2013.

[5] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Scene Statistics to Perceptual Quality," *IEEE Trans. Image Process,* vol. 20, no. 12, pp. 3350-3364, Dec. 2011.

[6] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. Image Process,* vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[7] M. Saad and A. C. Bovik, "Blind quality assessment of videos using a model of natural scene statistics and motion coherency," Invited Paper, in *Proc. Ann Asilomar Conf Signals, Syst, Comput.,* 2012.

[8] A. Eden, "No-reference estimation of the coding PSNR for H.264-coded sequences," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 667-674, May 2007.

[9] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp.1544-1554, Nov. 2010.

[10] C. Keimel, M. Klimpke, J. Habigt and K. Diepold, "No-reference video quality metric for HDTV based on H.264/AVC bitstream features," in *Proc. ICIP*, 2011.

[11] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, P. Demeester, "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Trans. Circuits Syst. Video Technol*., vol. 23, no. 8, pp.1322-1333, Aug. 2013.

[12] ITU-T Rec. P.1202.2, "Parametric non-intrusive bitstream assessment of video media streaming quality - higher resolution application area," May 2013.

[13] ITU-R Rec. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[14] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," 1999.

[15] Y. Liao, A. Younkin, J. Foerster, and P. Corriveau, "Achieving high QoE across the compute continuum: How compression, content, and devices interact," in *Proc. VPQM,* 2013.

[16] C. Keimel, A. Redl, and K. Diepold, "The TUM High Definition Video Data Sets," *in Proc. QoMEX,* 2012.

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

[18] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[19] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.

[20] L. K. Choi, Y. Liao, and A. C. Bovik, "Video QoE models for the compute continuum," *IEEE MMTC E-LETTER*, vol. 8, no. 5, pp. 26-29, Sep. 2013.