AVC, HEVC, VP9, AVS2 or AV1? — A Comparative Study of State-of-the-art Video Encoders on 4K Videos

Zhuoran Li, Zhengfang Duanmu, Wentao Liu, and Zhou Wang

University of Waterloo, Waterloo, ON N2L 3G1, Canada
{z777li,zduanmu,w238liu,zhou.wang}@uwaterloo.ca

Abstract. 4K, ultra high-definition (UHD), and higher resolution video contents have become increasingly popular recently. The largely increased data rate casts great challenges to video compression and communication technologies. Emerging video coding methods are claimed to achieve superior performance for high-resolution video content, but thorough and independent validations are lacking. In this study, we carry out an independent and so far the most comprehensive subjective testing and performance evaluation on videos of diverse resolutions, bit rates and content variations, and compressed by popular and emerging video coding methods including H.264/AVC, H.265/HEVC, VP9, AVS2 and AV1. Our statistical analysis derived from a total of more than 36,000 raw subjective ratings on 1,200 test videos suggests that significant improvement in terms of rate-quality performance against the AVC encoder has been achieved by state-of-the-art encoders, and such improvement is increasingly manifest with the increase of resolution. Furthermore, we evaluate state-of-the-art objective video quality assessment models, and our results show that the SSIMplus measure performs the best in predicting 4K subjective video quality. The database will be made available online to the public to facilitate future video encoding and video quality research.

Keywords: Video compression, quality-of-experience, subjective quality assessment, objective quality assessment, 4K video, ultra-high-definition (UHD), video coding standard

1 Introduction

4K, ultra high-definition (UHD), and higher resolution video contents have enjoyed a remarkable growth in recent years. 4K/UHD (4096×2160 or 3840×2160) video increases the resolution by a factor of four from full-HD (FHD, 1920×1080) and offers significantly increased sharpness and fine details. 4K/UHD video displays are believed to deliver better quality-of-experience (QoE) to viewers and are becoming widely available on the consumer market.

While 4K/UHD videos raise the potentials for better user QoE, their higher data rates cast great challenges to video distributions, for which video compression technologies are crucial in controlling the bandwidth of video so as to fit the distribution pipeline. The currently most widely used video coding technologies based on H.264 Advanced Video Coding (AVC) standards hardly meet the requirement. To this end, several modern video encoders including H.265 High Efficiency Video Coding (HEVC) [24], AOMedia Video 1 (AV1) [2], and Audio Video Coding Standard (AVS2) [20] are deliberately optimized for compressing content of 4K and higher resolutions. With many video encoders at hand, it becomes pivotal to compare their performance, so as to choose the best algorithms and find the direction for further advancement. Because the human visual system (HVS) is the ultimate receiver in most applications, subjective evaluation is a straightforward and reliable approach to evaluate the quality of videos. Although expensive and time consuming [26], a comprehensive subjective study has several benefits. First, it provides useful data to study human behaviors in evaluating perceived quality of encoded videos. Second, it supplies a test set to evaluate and compare the relative performance of classical and modern video encoding algorithms. Third, it is useful to validate and compare the performance of existing objective video quality assessment (VQA) models in predicting the perceptual quality of encoded videos. This will in turn provide insights on potential ways to improve them.

Several recent subjective studies have been conducted to evaluate the encoder performance on 4K video compression [3, 7, 10, 23]. It is generally observed that the latest video encoders can deliver 4K contents with better viewer QoE, although the test only covers a small number of contents. In addition, most of the work covers FHD and 4K for HEVC and AVC encoders only. In [27], HEVC encoder is evaluated by using 10 contents under 4K resolution. In [6], the performance of HEVC, AVC, and VP9 [9] at FHD and 4K are compared on 10 contents, from which it is shown that HEVC and VP9 achieve better bitrate reduction than AVC at the same quality level. The performance of the emerging next-generation encoders, AV1 and AVS2, on 4K videos has not been systematically evaluated. In summary, all of the aforementioned studies suffer from the following problems: (1) the test dataset is limited in size; (2) the types of encoders do not fully reflect the state-of-the-art; and (3) the spatial resolutions do not cover commonly used display sizes. Moreover, many tests have been conducted by the developers or participants of the coding standards. Independent datasets and test results commonly available to the public is lacking.

In this work, we conduct subjective evaluation of popular and emerging video encoders on 4K content. Our contributions are threefold. First, we carry out an independent and so far the most comprehensive subjective experiment to evaluate the performance of modern video encoders including AVC [25], VP9 [9], AV1 [1], AVS2 [21] and HEVC [18]. Second, we applied statistical analysis on the subjective data and observe some significant trends. Third, we use the database to evaluate objective VQA models to compare their prediction accuracy and complexity. The database will be made available online for future research.

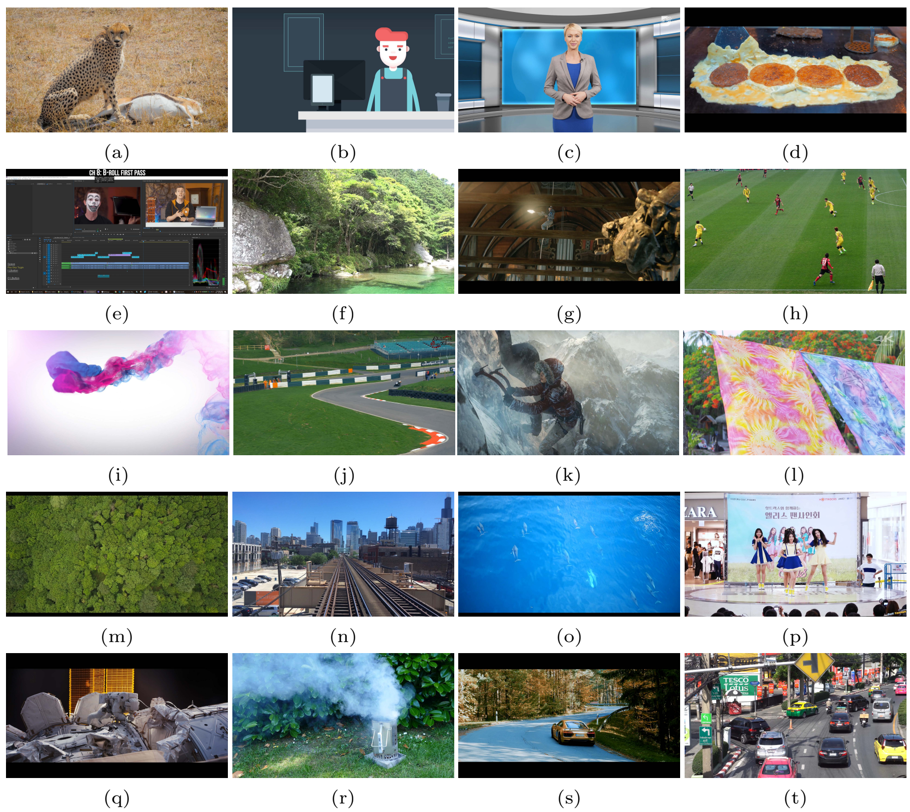

Fig. 1. Snapshots of source video sequences. (a) Safari. (b) 2D cartoon. (c) News. (d) Teppanyaki. (e) Screen recording. (f) Botanical garden. (g) Tears of steel. (h) Soccer game. (i) Animation. (j) Motor racing. (k) Climbing. (l) Colorfulness. (m) Forest. (n) Lightrail. (o) Dolphins. (p) Dance. (q) Spaceman. (r) Barbecue. (s) Supercar. (t) Traffic.

2 Video Database Construction and Subjective Experiment

The video database is created from 20 pristine high-quality videos of UHD resolution (3840×2160, progressive) selected to cover diverse content types, including humans, plants, natural scenes, architectures and computer-synthesized sceneries. All videos have the length of 10 seconds [8]. The detailed specifications are listed in Table 1 and the screenshots are shown in Fig. 1. Spatial information (SI) and temporal information (TI) [12] that roughly reflect the complexity of the video content are also given in Table 1, which suggests that the video sequences are of diverse spatio-temporal complexity and widely span the SI-TI space. Using the aforementioned video sequences as the source, each video is encoded with AVC, VP9, AV1, AVS2 and HEVC encoders with progressive scan at three spatial resolutions (3840×2160, 1920×1080, and 960×540) and four distortion levels. The detailed encoding configurations are as follows:

- HEVC: We employ x265 [18] with main profile for HEVC encoding. The GOP size is set to 60. Rate control mode is selected to be constant rate factor (CRF). Videos are encoded in "veryslow" speed setting.
- AVC: The x264 [25] with high profile of level 5 is used for AVC encoding. Other settings such as GOP size, rate control mode and speed setting are the same as those of the HEVC configurations.
- VP9: The libvpx software [9] is used for VP9 encoding. The encoding parameters, such as GOP size, rate control mode, etc., are set to be as similar as possible to HEVC. The parameter selection is based on [15].
- AV1: The AV1 reference software aomenc [1] is used for AV1 encoding. The encoding parameters are set to be as similar as possible to HEVC. The parameter selection is based on [15].
- AVS2: The libxavs2 [21] is used for AVS2 encoding. The encoding parameters, such as GOP size and speed setting are set to be as similar as possible to HEVC. The parameter selection is based on the configuration file "encoder_ra.cfg" that comes with AVS2 source code [21].

A small-scale internal subjective test is conducted and the encoding bitrates are adjusted to ensure that the neighboring distortion levels are perceptually distinguishable. Eventually, we obtain 1,200 videos encoded by 5 encoders in 3 resolutions at 4 distortion levels.

Our subjective experiment generally follows the single stimulus methodology as suggested by the ITU-T recommendation P.910 [12]. The experiment setup is normal indoor home settings with ordinary illumination level and no reflecting ceiling walls or floors. All videos are displayed at 3840×2160 resolution on a 28 inch 4K LED monitor with Truecolor (32bit) at 60Hz. The monitor is calibrated to meet the ITU-T BT.500 recommendations [11]. Videos are displayed in random order using a customized graphical user interface from which individual subjects' opinion scores are recorded.

Table 1. Spatial Information(SI), Temporal Information (TI), Framerate (FPS), and Description of Source Videos

Name	FPS	SI	TI	Description
Safari	24	26	41	Animal, smooth motion
2D carton	25	38	55	Animation, camera motion
News	25	32	45	Human, static
Teppanyaki	24	33	32	Food, average motion
Screen recording	30	82	12	Screen content, partial motion
Botanical garden	30	112	10	Natural scene, static
Tears of steel	24	28	61	Movie, high motion
Soccer game	30	54	24	Sports, high motion
Animation	30	55	32	Animation, high motion
Motor racing	24	57	37	Sports, camera motion
Climbing	30	38	73	Game, high motion
Colorfulness	30	23	65	Texture, smooth motion
Forest	24	46	24	Natural scene, camera motion
Lightrail	30	79	32	Architecture, camera motion
Dolphins	25	54	23	Animal, smooth motion
Dance	30	73	32	Human, high motion
Spaceman	24	51	2	Human, static
Barbecue	25	100	11	Natural scene, smooth motion
Supercar	25	80	22	Sports, average motion
Traffic	30	89	24	Architecture, high motion

A total of 66 naïve subjects, including thirty nine males and twenty seven females aged between 18 and 35, participated in the subjective test. Visual acuity and color vision are confirmed with each subject before the subjective test. To familiarize the subjects with the testing environment, a training session is performed before the formal experiment, in which 3 videos different from those in the formal experiment are rendered. The same methods are used to generate the videos used in the training and testing sessions. Therefore, before the testing session, subjects knew what distortion types would be expected. Subjects were instructed with sample videos to judge the overall video quality based on the distortion level. Due to the limited subjective experiment capacity, we employed the following strategy. Each subject is assigned 10 contents in a circular fashion. Specifically, if subject i is assigned contents 1 to 10, then subject $i+1$ watch contents 2 to 11. Each video is assessed for at least 30 times and more than 36,000 subjective ratings are collected in total. For each subject, the whole study takes about 3 hours, which is divided into 6 sessions with five 5-minute breaks in-between to minimize the influence of fatigue effect.

We employ 100-point continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) for three advantages: broader range, finer distinctions between ratings, and demonstrated prior efficacy [16]. After converting the subjective scores to Z-scores per session to account for any differences in the use of the quality scale between sessions, we proceed to an outlier removal process suggested in [11]. No outlier detection is conducted participant-wise. After outlier removal, Z-scores are linearly re-scaled to lie in the range of [0, 100]. The final quality score for each individual video is computed as the average of the re-scaled Z-scores, namely the mean opinion score (MOS), from all valid subjects. Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) between the score given by each subject and MOS are calculated. The average PLCC and SRCC across all subjects are 0.79 and 0.78, with standard deviation (STD) of 0.09 and 0.08, respectively, suggesting that there is considerable agreement among different subjects on the perceived quality of the test video sequences.

3 Evaluation of Video Encoders

We use the MOS of the test videos described in the previous section to evaluate and compare the performance of the encoders. It is worth noting that the performance comparison is based on the encoder configuration provided earlier, where all encoders are set to configurations equivalent to the 'veryslow' setting of the HEVC encoders.

Sample rate-distortion (RD) curves for individual test videos are given in Fig. 2. From the RD curves of all content, we have several observations. First, AVC under-performs all the other four encoders in most cases, which justifies the performance improvement of the newly developed video encoders in recent years. Second, the performance difference between different encoders, exhibited as the gaps between the RD curves, become increasingly manifest with the increase of resolution from 540p to 1080p, and then to 2160p. This validates the coding gain obtained by the advanced technologies specifically designed for high resolution videos in the newly developed encoders. This observation also justifies the necessity of cross-resolution subjective and objective video quality assessment because the visibility of coding artifacts changes from low to high resolution content. Third, we observe that AV1 achieves the highest bitrate savings for high motion content. This may be explained by the advancement of AV1 motion prediction schemes which utilizes warped motion, global motion tools and more reference frames [17].

In addition to the qualitative analysis, we also compute the average bitrate saving [4, 5] of each encoder over another. The result is shown in Table 2, from which we can observe that on average AV1 outperforms the other encoders with a sizable margin. However, it is worth noting that the RD performance gain by AV1 is highly content dependent and that AV1's performance is achieved on the condition of its much higher complexity compared with all other encoders.

The time complexity performance test is done on a Ubuntu 16.04 system with Intel E5-1620 CPU and 32GB RAM. As shown in Table 3, we can see that AV1 consumes over 500 times of AVC's computational time, which takes the least amount of encoding time. The results suggest that state-of-the-art AVC implementations are still highly competitive choices for time critical tasks, while the encoding speed of AV1 may hinder it from many practical applications. It is

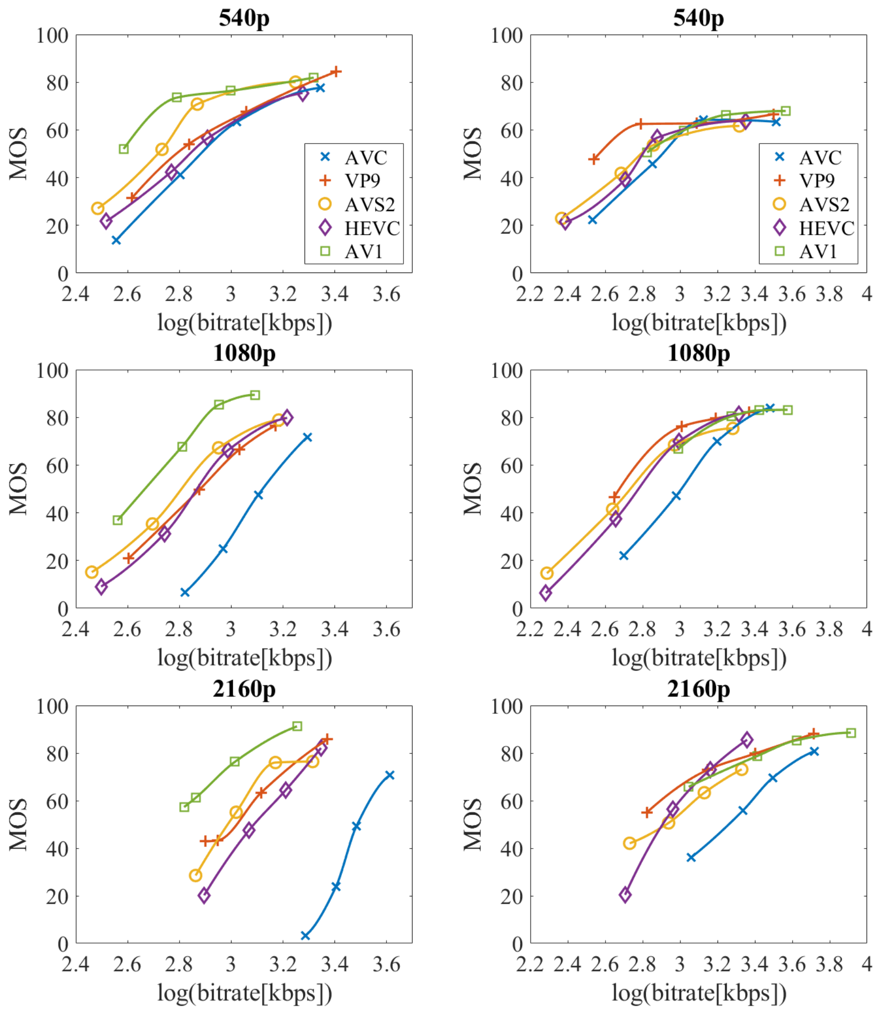

Fig. 2. RD curves of AVC, VP9, HEVC, AVS2 and AV1 encoders for 540p, 1080p and 2160p resolutions for Tears of steel (left) and Barbecue (right).

Table 2. Column BD-Rate Saving vs. Row (negative percentages suggest savings)

540p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-22.7%	0	-	-	-
AVS2	-20.3%	-4.7%	0	-	-
VP9	-28.9%	-20.5%	-25.7%	0	-
AV1	-34.4%	-23.3%	-17.6%	-4.5%	0

1080p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-42.2%	0	-	-	-
AVS2	-45.8%	-9.8%	0	-	-
VP9	-47.5%	-18.5%	-18.1%	0	-
AV1	-48.7%	-20.1%	-21.4%	-3.5%	0

2160p	AVC	HEVC	AVS2	VP9	AV1
AVC	0	-	-	-	-
HEVC	-61.2%	0	-	-	-
AVS2	-63.5%	-9.7%	0	-	-
VP9	-62.2%	-8.7%	-5.3%	0	-
AV1	-63.2%	-9.5%	-15.0%	-16.4%	0

worth mentioning that AV1 is still under development and the current version has not been fully optimized for multi-thread encoding. VP9 and HEVC show comparable time complexity, while AVS2 doubles their encoding time. They compromise between compression performance and speed.

Table 3. Encoder Relative Complexity vs. AVC at 3 Resolutions

	AVC	HEVC	AV1	VP9	AVS2
2160p	1	4.2810	590.74	5.2856	9.8568
1080P	1	4.7314	546.19	6.6286	10.0401
540P	1	5.2805	806.15	5.2572	11.7716

4 Performance of Objective Quality Assessment Methods

We use four representative objective VQA models including PSNR, VQM [19], VMAF [13] (version v0.6.1), VMAF-4K [14] (version v0.6.1), and SSIMplus [22] to test their generalizability on novel video encoders. The implementations of the VQA models are obtained from the original authors. Only SSIMplus supports

Table 4. Performance Comparison of VQA Models Using MOS as Ground-truth

VQA Models		All		540p		1080p		2160p	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR		0.4197	0.4162	0.3993	0.4143	0.4155	0.3858	0.3259	0.3252
VMAF-4K (v0.6.1)		0.5505	0.5530	0.5102	0.4726	0.5784	0.5371	0.4601	0.4414
VQM		0.6154	0.6282	0.5165	0.5357	0.6659	0.6722	0.5831	0.6163
VMAF (v0.6.1)		0.7371	0.7387	0.7247	**0.7018**	0.7909	0.7646	0.6335	0.6521
SSIMplus		**0.7930**	**0.7757**	**0.7604**	0.6874	**0.8662**	**0.8265**	**0.7469**	**0.7523**
Individual Subjects	Average	0.7917	0.7819	0.7229	0.7007	0.8287	0.8079	0.7770	0.7584
	STD	± 0.0068	± 0.0081	± 0.0108	± 0.0132	± 0.0066	± 0.0077	± 0.0078	± 0.0098

direct cross-resolution video quality evaluation. For the other VQA models, all representations are upsampled to 3840×2160 using bilinear filter and the VQA is performed on the up-sampled videos. PLCC and SRCC are employed to evaluate the performance of objective VQA models in terms of their effectiveness in predicting MOS.

Table 4 summarizes the overall performance of the VQA models and the breakdown results for three resolutions, where the top VQA models for each evaluation criterion are highlighted in bold. Overall, SSIMplus is the best performing VQA model in most cases. Specifically, the performance gaps between SSIMplus and the other models increases with resolution, and the gap is the largest at 4K/UHD resolution. This justifies the effectiveness of the HVS-based resolution adaptation mechanism underlying the SSIMplus approach [22]. PSNR, the traditional quality model, is the weakest in the current test, which is likely due to its ignorance of any HVS properties. For VQM, we observe major inconsistent scoring across different video content, suggesting that there is space for improvement in VQM in terms of content-adaptation. Our results also show that the VMAF model tends to overestimate the quality scores of AVC videos. This may be because VMAF is a learning-based approach and was originally trained using H.264/AVC compressed videos, but the statistical properties of the artifacts in the newly developed encoding methods are different. For example, HEVC and AV1 encoders produce less blockiness and smoother transition between frames. These features may not be properly captured by VMAF. Somewhat surprisingly, for the VMAF-4K model, which was claimed to be better suited for 4K TV device, the correlation is significantly lower than VMAF even when tested with the 4K dataset. This might be because the VMAF-4K model was trained to cover a very wide range of video resolutions, for which the size of its training dataset may not be sufficient [14]. Using MOS as the ground-truth, we can compare the scores given by any individual subject, and evaluate the performance of the subject in predicting MOS. The average of individual subjects' performance and the standard devidiation (STD) between all individual subjects are also given in Table 4. Such "average subject" performance gives us a baseline about the difficulty of the quality assessment task. Using this baseline, we observe that top VQA models such as SSIMplus performs closely to an average subject.

VQA measurement is often a computationally demanding task but real-world applications such as live QoE monitoring often desire video quality being evaluated in real-time. Fig. 3 compares VQA methods' prediction accuracy against speed on videos of 4K resolution, where the speed performance test for VQA methods is done on a Ubuntu 16.04 system with Intel E5-1620 CPU and 32GB RAM. It appears that the VMAF and VMAF-4K models are much faster than VQM while maintaining a similar level of quality prediction accuracy. Overall, the SSIMplus measure clearly offers the best compromise between speed and accuracy.

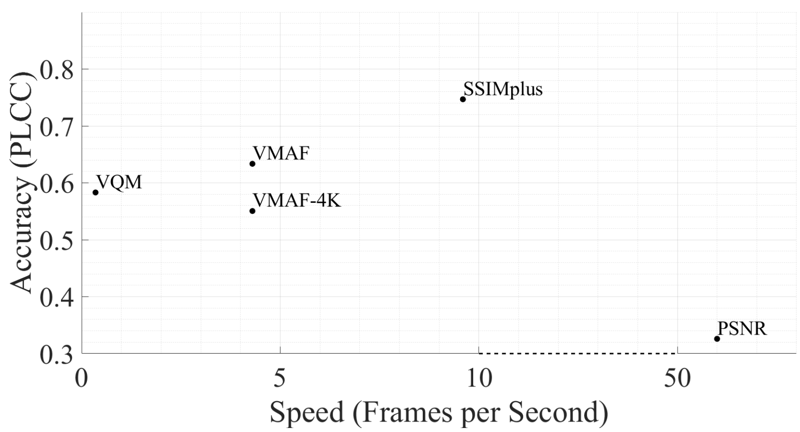

Fig. 3. Speed vs. prediction accuracy comparison of VQA models on 4K resolution videos

5 Conclusions and Discussion

We conduct an independent and so far the most comprehensive subjective evaluation and performance analysis, specifically on popular and emerging video encoders (AVC, HEVC, VP9, AVS2, and AV1) with video content of diverse resolutions and bitrates. The five video encoders are evaluated across 20 source 4K contents from the view points of content dependency and resolution adaptation. The subjective testing results are also used to test the performance of representative VQA models, among which the SSIMplus measure achieves the best compromise between accuracy and speed. The testing results will be made publicly available to facilitate future video coding and VQA research.

It is important to note that video coding standards define decoders only, and their encoder instantiations and configurations vary significantly from one to another. Due to the limited subjective experiment capacity and the large number of combinations of encoder configurations, absolute "fair" comparison of video decoders or coding standards is extremely difficult. Therefore, conclusions about the performance of video coding standards should be drawn with caution. The current study is valid for the given encoders with the specified encoding configurations only.

References

1. Alliance for Open Media: AV1 codec source code repository (Jun 2018), https://aomedia.googlesource.com/aom
2. Alliance for Open Media: The alliance for open media kickstarts video innovation era with "AV1" release (March 2018), https://aomedia.org/the-alliance-for-open-media-kickstarts-video-innovation-era-with-av1-release/
3. Bae, S.H., Kim, J., Kim, M., Cho, S., Choi, J.S.: Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services. IEEE Trans. Broadcasting **59**(2), 209–222 (2013)
4. Bjontegaard, G.: Calculation of average PSNR differences between RD-curves. In: ITU-T Q. 6/SG16, 33th VCEG Meeting (2001)
5. Bjontegaard, G.: Improvements of the BD-PSNR model, VCEG-AI11. In: ITU-T Q. 6/SG16, 34th VCEG Meeting (2008)
6. Cheon, M., Lee, J.S.: Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience. IEEE Trans. Circuits and Systems **28**(7), 1467–1480 (2018)
7. Deshpande, S.: Subjective and objective visual quality evaluation of 4K video using AVC and HEVC compression. In: SID Symposium Digest of Technical Papers. vol. 43, pp. 481–484 (2012)
8. Fröhlich, P., Egger, S., Schatz, R., Mühlegger, M., Masuch, K., Gardlo, B.: QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment? In: Proc. IEEE Int. Conf. on Quality of Multimedia Experience. pp. 242–247 (2012)
9. Google: libvpx (Jul 2018), https://chromium.googlesource.com/webm/libvpx.git
10. Hanhart, P., Rerabek, M., De Simone, F., Ebrahimi, T.: Subjective quality evaluation of the upcoming HEVC video compression standard. In: Applications of Digital Image Processing XXXV. vol. 8499, pp. 1–13 (2012)
11. ITU-R BT.500: Recommendation: Methodology for the subjective assessment of the quality of television pictures (Jan 2012)
12. ITU-R BT.910: Recommendation: Subjective video quality assessment methods for multimedia applications (Apr 2008)
13. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., Manohara, M.: Toward a practical perceptual video quality metric (Jun 2016), https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652
14. Li, Z., Vigier, T., Callet, P.L.: A vmaf model for 4k (Mar 2018), ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Madrid_Mar18/Meeting_Files/VQEG_SAM_2018_025_VMAF_4K.pdf
15. LIU, Y.: AV1 beats x264 and libvpx-vp9 in practical use case (Apr 2018), https://code.fb.com/video-engineering/av1-beats-x264-and-libvpx-vp9-in-practical-use-case/

16. Ma, K., Wu, Q., Wang, Z., Duanmu, Z., Yong, H., Li, H., Zhang, L.: Group MAD competition-A new methodology to compare objective image quality models. In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition. pp. 1664–1673 (2016)
17. Massimino, P.: AOM - AV1, How does it work? (Jul 2017), https://parisvideotech.com/wp-content/uploads/2017/07/AOM-AV1-Video-Tech-meet-up.pdf
18. MultiCoreWare Inc: x265 (Jul 2018), https://bitbucket.org/multicoreware/x265
19. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. Broadcasting **50**(3), 312–322 (2004)
20. PKU-VCL: AVS2 technology (2018), http://www.avs.org.cn/avs2/technology.asp
21. PKU-VCL: AVS2 codec source code repository (Jan 2018), https://github.com/pkuvcl/xavs2
22. Rehman, A., Zeng, K., Wang, Z.: Display device-adapted video quality-of-experience assessment. In: Human Vision and Electronic Imaging XX. vol. 9394, pp. 1–11 (2015)
23. Řeřábek, M., Ebrahimi, T.: Comparison of compression efficiency between HEVC/H.265 and VP9 based on subjective assessments. In: Applications Of Digital Image Processing Xxxvii. vol. 9217, pp. 1–13 (2014)
24. Tan, T., Mrak, M., Baroncini, V., Ramzan, N.: Report on HEVC compression performance verification testing. Joint Collab. Team Video Coding (JCT-VC) (2014)
25. VideoLAN: x264 (Jul 2018), http://git.videolan.org/git/x264
26. Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE signal processing magazine **26**(1), 98–117 (2009)
27. Zhu, Y., Song, L., Xie, R., Zhang, W.: SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding. In: IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. pp. 1–4 (2016)