Waterloo 4K Video Quality Assessment Database: A Comparative Study of Modern Video Encoders

Zhuoran Li, Zhengfang Duanmu, Wentao Liu and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo
200 University Ave W
Waterloo, ON, N2L 3G1, Canada
{z777li, zduanmu, w238liu, zhou.wang}@uwaterloo.ca

Abstract

We introduce a subject-rated 4K video quality database that contains perceptual video distortions from state-of-the-art encoders. The Waterloo 4K Video Quality database, so far the largest of its kind, consists a total of 1200 videos generated from 5 encoders including AV1, AVS2, HEVC, H.264, and VP9 at 3 spatial resolutions. According to the experimental results, the new generation of encoders provides significant average bitrate savings over H.264, with the best performance achieved by AV1. We evaluate the performance of 5 objective video quality assessment (VQA) models with regards to their efficacy in predicting subjective video quality. The subjective database is available online at `https://ece.uwaterloo.ca/~z777li/4kvqa/`.

1. Introduction

4K or UHD content is expected to deliver a better quality of experience that is widely adopted in consumer video services in recent years. By increasing the resolution by 4 times over the full HD (FHD) content, 4K is capable of providing more perceptual details and information. Compared with FHD standards, 4K not only sets a higher post for perceptual quality, but also raises a stricter requirement for video compression efficiency.

Since its standardization, 4K has attracted an increasing amount of attention from the video compression field. Although higher resolution delivers better video clarity, the increase in the number of pixels brings new challenge to video encoders under the limiting storage or bandwidth resource. To this end, several modern video encoders such as HEVC [1], AV1 [2], and AVS2 [3] are deliberately optimized for 4K content compression. With many video encoders at hand, it becomes pivotal to compare their performance, so as to find the best algorithms as well as further advancement direction. 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 [4], a comprehensive subjective 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 perceptual quality gain of 4K over FHD videos [5] [6] [7] [8]. Although conclusions are drawn in their paper that 4K contents deliver better quality-of-experience (QoE) against FHD, most of the work only covers a small number of contents. Moreover, in terms of resolutions and encoders, most of the work only covers FHD and 4K for HEVC and H.264 encoders. In [9], only HEVC encoder is evaluated by using 10 contents under 4K resolution. Cheon *et al.* compared the performance of HEVC, H.264, and VP9 at FHD and 4K on 10 contents [10], from which they conclude that the added value of 4K over FHD was more noticeable at high bitrate, which was more prominent for contents having high spatial complexity. However, the performance of next-generation encoders, AV1 and AVS2, on 4K videos has not been systematically evaluated. In summary, all of the above studies suffer from the following problems: (1) the dataset is limited in size; (2) the type of encoders do not fully reflect the state-of-the-art; and (3) the spatial resolutions do not cover enough commonly used display sizes.

In this work, we conduct subjective evaluation of traditional and state-of-the-art video encoders on 4K content. Our contributions are threefold. First, we construct so far the largest subject-rated 4K video database. The database contains 20 high quality sequences of diverse content types and 1200 compressed videos generated from 5 encoders including H.264 [11], VP9 [12], AV1 [13], AVS2 [14] and HEVC [15]. Second, we carry out a subjective experiment to evaluate the performance of video encoders. Our analysis illustrate that AV1 achieves quality gain at the cost of significantly longer encoding time. Third, we evaluate 5 objective VQA models. Existing objective VQA models exhibit moderate performance in predicting the perceptual distortion introduced by novel encoders at high resolutions.

2. 4K Video Database and Subjective Quality Experiment

Video Database Construction

A video database, named Waterloo 4K Video Quality database, of 20 pristine high-quality videos of size 3840 × 2160 are selected to cover diverse content, including humans, plants, natural scenes, architectures and computer-synthesized sceneries. All videos have the length of 10 s [16]. The detailed specifications are listed in Table 1 and screenshots are shown in Fig. 1. Spatial information (SI) and temporal information (TI) [17] that roughly reflect the complexity of 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 sequences as source, each video is encoded with H.264, VP9, AV1, AVS2 and HEVC at three spatial resolutions (3840×2160, 1920×1080, and 960×540) and four distortion levels. The specific encoders and their detailed configurations are shown in Table 6 and Table 5 in Appendix. A small-scale internal subjective test is conducted and the encoding bitrates are adjusted accordingly to guarantee that the neighboring distortion levels are perceptually distinguishable. Eventually, we obtain 1200 videos encoded by 5 encoders in 3 resolutions and 4 distortion levels.

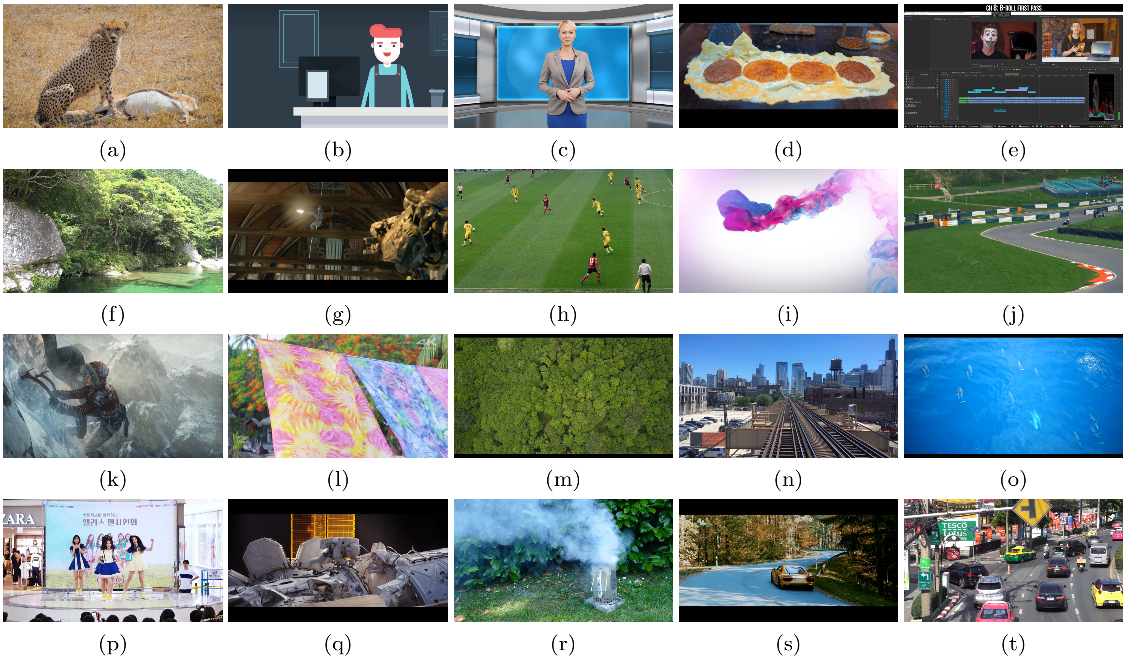


Figure 1: Snapshots of reference video sequences. (a) Safari. (b) 2D cartoon. (c) News. (d) Teppanyaki. (e) Screen recording. (f) Botanical garden. (g) Tears of steel. (h) Soccer game. (i) Animation. (j) Motor racing. (k) Climbing. (l) Colorfulness. (m) Forest. (n) Lightrail. (o) Dolphins. (p) Dance.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

Table 1: Spatial Information(SI), Temporal Information (TI), Framerate (FPS), and Description of Reference 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

Subjective Experiment Methodology

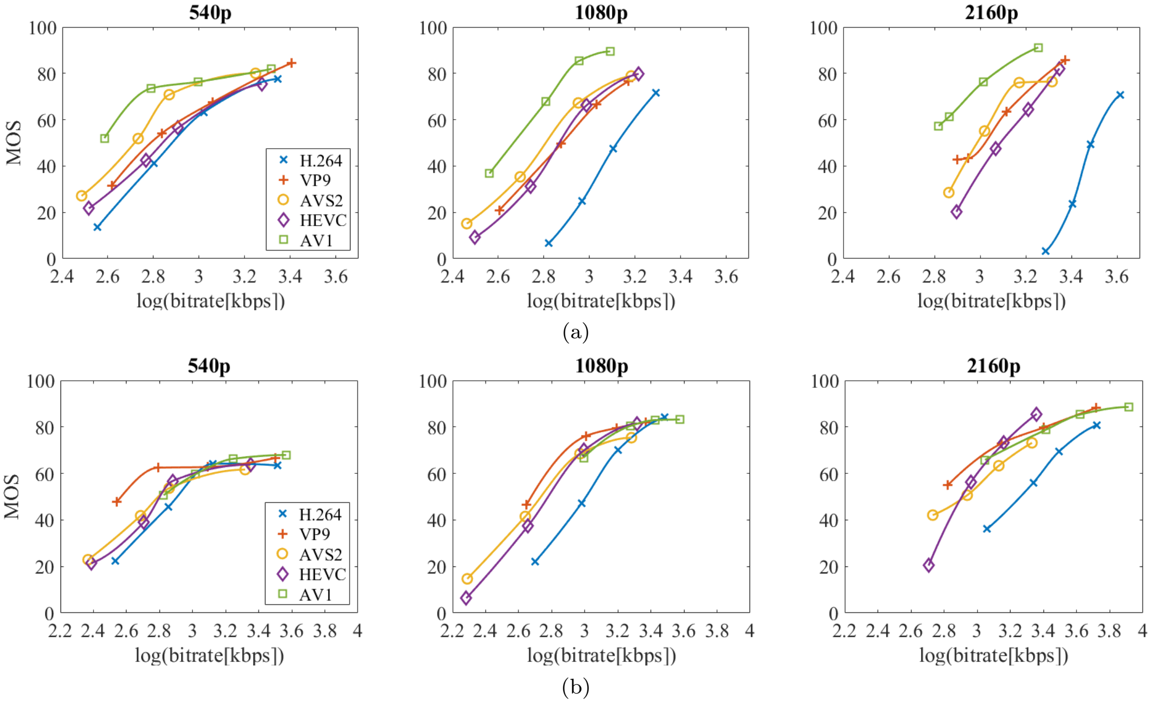
Our subjective experiment generally follow the single stimulus methodology as suggested by the ITU-T recommendation P.910 [17]. The experiment setup is normal indoor home settings with ordinary illumination level and no reflecting ceiling walls or floors. All videos are displayed at 3840x2160 resolution on a 28 inch 4K LED monitor with Truecolor (32bit) at 60Hz. The monitor is calibrated to meet the ITU-T BT.500 recommendations [18]. Videos are displayed in random order using a customized graphical user interface and individual subjects' opinion score are recorded.

A total of 66 naive subjects, including thirty nine males and twenty seven females aged between 18 and 35, participated the subjective test. Visual acuity and color vision are confirmed with each subject before the subjective test. A training session is performed before the formal experiment, in which 3 videos different from those in 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 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. 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.

In our experiment, we chose 100-point continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) due to three advantages: broader range, finer distinctions between ratings, and demonstrated prior efficacy [19]. After converting 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 [18]. No outlier detection is conducted participant-wise due to the fact that in our rotational experiment only three or four participants watched the same 600 videos. 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 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.

Subjective Data Analysis

We use the MOS value of the 5 encoders described in the previous section to evaluate and compare their performance. It is worth noting that the performance comparison is only based on the encoder configuration provided in Appendix, where all encoders are set to configuration equivalent to the 'veryslow' setting of HEVC encoders.



(a)

(b)

Figure 2: RD curves of all three resolutions for source video (a) Tears of steel. (b) Barbecue.

Sample rate-distortion (RD) curves are given in Fig. 2. From the RD curves of all content, we have several observations. First, H.264 under-performs all the other four encoders in most cases, which justifies the performance improvement of the latest video encoders in recent years. Second, under the resolution of 540p, the RD curves of all encoders are clustered together across almost all bitrates, while the encoder performance difference becoming more obvious when the resolution increase. This observation validates the necessity of 4K subjective video quality testing because compression distortion are more visible for the high resolution content. Third, we can observe that AV1 achieves the highest bitrate saving 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 [20].

In addition to the qualitative analysis, we also compute the bitrate saving [21] [22] of each encoder over another. The result is shown in Table 2, from which we can observe that AV1 outperforms the other encoders with a sizable margin. However, it is worth noting that AV1's performance is achieved on the condition of its much longer computation time compared with all other encoders.

The time complexity performance test is done on a Ubuntu 16.04 system with Intel E5-1620 CPU. As shown in Table 3, we can see the AV1 consumes over 500 times of H.264's computational time, which take the least amount of encoding time. The results suggest that state-of-the-art H.264 implementations are still highly competitive choices for time critical tasks, while the encoding speed of AV1 hinders it from practical applications. It is worth mentioning that AV1 is still under development and the current version has not been fully optimized for multi-thread encoding. VP9

Table 2: Column BD-Rate Saving vs. Row (Lower the Better)

540p	AVC	VP9	AVS2	HEVC	AV1
AVC	0	-	-	-	-
VP9	-28.9%	0	-	-	-
AVS2	-20.3%	34.5%	0	-	-
HEVC	-22.7%	24.2%	4.9%	0	-
AV1	-34.4%	-4.5%	-17.6%	-23.3%	0

1080p	AVC	VP9	AVS2	HEVC	AV1
AVC	0	-	-	-	-
VP9	-47.5%	0	-	-	-
AVS2	-45.8%	22.1%	0	-	-
HEVC	-42.2%	22.7%	10.8%	0	-
AV1	-48.7%	-3.5%	-21.4%	-20.1%	0

2160p	AVC	VP9	AVS2	HEVC	AV1
AVC	0	-	-	-	-
VP9	-62.2%	0	-	-	-
AVS2	-63.5%	5.5%	0	-	-
HEVC	-61.2%	9.5%	10.7%	0	-
AV1	-63.2%	-16.4%	-15.0%	-9.5%	0

Table 3: Encoder Relative Complexity vs. H.264 at 3 Resolutions

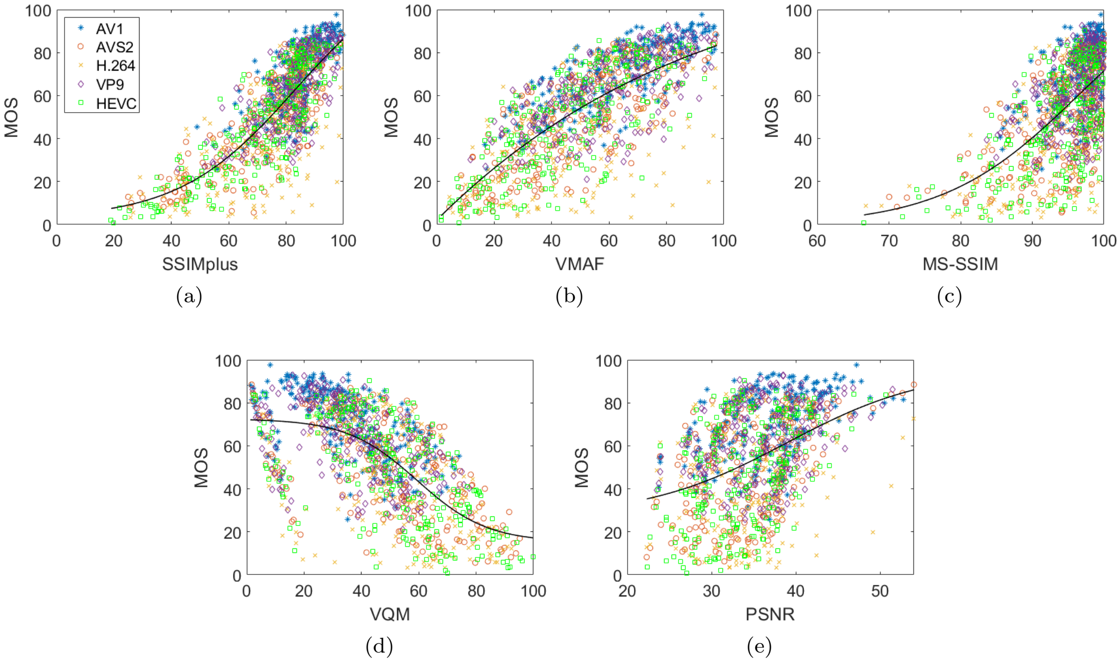
	H264	HEVC	AV1	VP9	AVS2
4K	1	4.2810	590.74	5.2856	9.8568
1080P	1	4.7314	546.19	6.6286	10.0401
540P	1	5.2805	806.15	5.2572	11.7716

and HEVC show comparable time complexity while the AVS2 double their encoding time. They compromise between compression performance and speed.

3. Objective Quality Assessment

We use 5 objective VQA models including PSNR, MS-SSIM [23], VQM [24], VMAF [25] and SSIMplus [26] to test their generalizability on novel video encoders. The implementations of the VQA models are obtained from the original authors. Since none of the VQA models except for SSIMplus supports cross-resolution video quality evaluation, for the other 4 VQA models, all representations are upsampled to 3840×2160 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. Scatter plots of objective scores vs. MOS for all the 5 QoE metrics on our dataset, along with the best fitting logistic functions, are shown in Fig. 3.

Table 4 summarizes their overall performance and the performance under the three resolutions, where the top VQA models for each evaluation criterion are highlighted in bold. It can be observed that in most cases SSIMplus is the best performing VQA



(a) (b) (c) (d) (e)

Figure 3: Video Quality Metric versus MOS. (a) SSIMplus. (b) VMAF. (c) MS-SSIM. (d) VQM. (e) PSNR.

Table 4: Performance Comparison of VQA Models

	Overall		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
SSIMplus	**0.7930**	**0.7757**	**0.7604**	0.6874	**0.8662**	**0.8265**	**0.7469**	**0.7523**
VMAF	0.7371	0.7387	0.7247	**0.7018**	0.7909	0.7646	0.6335	0.6521
VQM	0.6154	0.6282	0.5165	0.5357	0.6659	0.6722	0.5831	0.6163
MS-SSIM	0.5942	0.5555	0.5100	0.4549	0.6440	0.5936	0.5945	0.5574
Average Subject	0.7917	0.7819	0.7229	0.7007	0.8287	0.8079	0.7770	0.7584

model, in that it can capture video quality across different resolutions. For VQM, we can observe that the points are clustered for different contents, which indicate that VQM has large potential for improvement in terms of content-adaption. PSNR, the traditional quality model, is the weakest in the current test, which is likely due to its ignorance of any human visual system properties. Moreover, based on the scatter plot, we can see that both SSIMplus and VMAF tend to overestimate the quality score of H.264 videos. This perhaps indicates that with the better quality performance of modern video encoders, the quality standard of subjective participants also increase, while the VQA model, which trained or tuned on classical encoders, may fail to predict the quality score accurately. From our observation, modern video encoders, such as AV1 and HEVC, produce less blocky compression artifact and more smooth frame transition compared with H.264, which VQA models may paid less attention to. The average subject-wise correlation against MOS is also included in Table 4, from which we can see that even though current VQA models are not fully efficacious in predicting video QoE, top models such as SSIMplus and VMAF can achieve average human performance.

4. Conclusions and Discussion

We introduce the Waterloo 4K Video Quality database, which contains 1200 encoded videos that were derived from diverse source videos and 5 modern video encoders. We assessed 5 VQA models with statistical analysis. The database is made publicly available to facilitate future 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, "fair" comparison of video encoders is extremely difficult, if not impossible.

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. Moreover, state-of-the-art VQA models exhibit moderate correlations the the MOS, suggesting space for further improvement.

References

[1] T. Tan, M. Mrak, V. Baroncini, and N. Ramzan, "Report on HEVC compression performance verification testing," *Joint Collab. Team Video Coding (JCT-VC)*, 2014.

[2] Alliance for Open Media. (March. 2018) The alliance for open media kickstarts video innovation era with "AV1" release. [Online]. Available: https://aomedia.org/the-alliance-for-open-media-kickstarts-video-innovation-era-with-av1-release/

[3] PKU-VCL. (2018) AVS2 technology. [Online]. Available: http://www.avs.org.cn/avs2/technology.asp

[4] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[5] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in *Applications of Digital Image Processing XXXV*, vol. 8499, no. 84990V, 2012, pp. 1–13.

[6] S. Deshpande, "Subjective and objective visual quality evaluation of 4K video using AVC and HEVC compression," in *SID Symposium Digest of Technical Papers*, vol. 43, no. 1, 2012, pp. 481–484.

[7] S.-H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, "Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services," *IEEE Trans. Broadcasting*, vol. 59, no. 2, pp. 209–222, 2013.

[8] M. Řeřábek and T. Ebrahimi, "Comparison of compression efficiency between HEVC/H.265 and VP9 based on subjective assessments," in *Applications Of Digital Image Processing Xxxvii*, vol. 9217, no. 92170U, 2014, pp. 1–13.

[9] Y. Zhu, L. Song, R. Xie, and W. Zhang, "SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2016, pp. 1–4.

[10] M. Cheon and J.-S. Lee, “Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience,” *IEEE Trans. Circuits and Systems*, vol. 28, no. 7, pp. 1467–1480, 2018.

[11] FFmpeg team. (Jul. 2018) FFmpeg v.2.8.15. [Online]. Available: https://trac.ffmpeg.org/wiki/Encode/H.264

[12] ——. (Jul. 2018) FFmpeg v.2.8.15. [Online]. Available: https://trac.ffmpeg.org/wiki/Encode/VP9

[13] Alliance for Open Media. (Jun. 2018) AV1 codec source code repository. [Online]. Available: https://aomedia.googlesource.com/aom

[14] PKU-VCL. (Jan. 2018) AVS2 codec source code repository. [Online]. Available: https://github.com/pkuvcl/xavs2

[15] FFmpeg team. (Jul. 2018) FFmpeg v.2.8.15. [Online]. Available: https://trac.ffmpeg.org/wiki/Encode/H.265

[16] P. Fröhlich, S. Egger, R. Schatz, M. Mühlegger, K. Masuch, and B. Gardlo, “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*, 2012, pp. 242–247.

[17] ITU-R BT.910, “Recommendation: Subjective video quality assessment methods for multimedia applications,” Apr. 2008.

[18] ITU-R BT.500, “Recommendation: Methodology for the subjective assessment of the quality of television pictures,” Jan. 2012.

[19] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, “Group MAD competition-A new methodology to compare objective image quality models,” in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1664–1673.

[20] P. Massimino. (Jul. 2017) AOM - AV1, How does it work? [Online]. Available: https://parisvideotech.com/wp-content/uploads/2017/07/AOM-AV1-Video-Tech-meet-up.pdf

[21] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” in *ITU-T Q. 6/SG16, 33th VCEG Meeting*, 2001.

[22] ——, “Improvements of the BD-PSNR model, VCEG-AI11,” in *ITU-T Q. 6/SG16, 34th VCEG Meeting*, 2008.

[23] Z. Wang, E. Simoncelli, A. Bovik *et al.*, “Multi-scale structural similarity for image quality assessment,” in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003, pp. 1398–1402.

[24] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[25] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (2016, Jun.) Toward a practical perceptual video quality metric. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[26] A. Rehman, K. Zeng, and Z. Wang, “Display device-adapted video quality-of-experience assessment,” in *Human Vision and Electronic Imaging XX*, vol. 9394, no. 939406, 2015, pp. 1–11.

Appendix

Table 5: Encoder Version and Package

Encoder	Version	Package URL
AV1	v1.0.0	`https://aomedia.googlesource.com/aom/+/v1.0.0`
AVS2	v1.0	`https://github.com/pkuvcl/xavs2`
HEVC	ffmpeg v.2.8.15	`https://trac.ffmpeg.org/wiki/Encode/H.265`
H264	ffmpeg v.2.8.15	`https://trac.ffmpeg.org/wiki/Encode/H.264`
VP9	ffmpeg v.2.8.15	`https://trac.ffmpeg.org/wiki/Encode/VP9`

Table 6: Encoder Configurations

AV1	aomenc INPUT –width=WIDTH –height=HEIGHT –i420 -y –fps=FRAMERATE/1 –cpu-used=1 –threads=4 –profile=0 –lag-in-frames=19 –min-q=0 –max-q=63 –auto-alt-ref=1 –kf-max-dist=60 –kf-min-dist=60 –drop-frame=0 static-thresh=0 –bias-pct=50 –minsection-pct=0 –maxsection-pct=2000 –arnr-maxframes=7 –arnr-strength=5 –sharpness=0 –undershoot-pct=100 –overshoot-pct=100 –tile-columns=2 –frame-parallel=0 –test-decode=warn -v –end-usage=q –cq-level=BITRATE –webm -o OUTPUT
AVS2	xavs2 -f encoder_ra.cfg -p InputFile=INPUT –FramesToBeEncoded=FRAMERATE –FrameRate=FR –SourceWidth=WIDTH –SourceHeight=HEIGHT –InputSampleBitDepth=8 –SampleBitDepth=8 –TargetBitRate=BITRATE –OutputFile=OUTPUT
HEVC	ffmpeg -i INPUT -c:v libx265 -preset veryslow -s WIDTHxHEIGHT -crf CRF -x265-params “ref=5:keyint=60:min-keyint=60:scenecut=0” OUTPUT
H264	ffmpeg -i INPUT -c:v libx264 -preset veryslow -s WIDTHxHEIGHT -crf CRF -refs 5 -g 60 -keyint_min 60 -sc_threshold 1 -f mp4 OUTPUT
VP9	ffmpeg -i INPUT -c:v libvpx-vp9 -pass 1 -speed 1 -s WIDTHxHEIGHT -crf CRF -b:v 0 -tile-columns 0 -frame-parallel 0 -f webm /dev/null; ffmpeg -i INPUT -c:v libvpx-vp9 -pass 2 -speed 1 -s WIDTHxHEIGHT -crf CRF -b:v 0 -tile-columns 0 -frame-parallel 0 -auto-alt-ref 1 -lag-in-frames 25 -f webm OUTPUT