Katsenou, A., Zhang, A., Afonso, M., & Bull, D. (2019). A Subjective Comparison of AV1 and HEVC for Adaptive Video Streaming. In *2019 IEEE International Conference on Image Processing (ICIP 2019)* Institute of Electrical and Electronics Engineers (IEEE).
<https://doi.org/10.1109/ICIP.2019.8803523>

Peer reviewed version

Link to published version (if available):
[10.1109/ICIP.2019.8803523](https://doi.org/10.1109/ICIP.2019.8803523)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at <https://ieeexplore.ieee.org/document/8803523>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

A SUBJECTIVE COMPARISON OF AV1 AND HEVC FOR ADAPTIVE VIDEO STREAMING

Angeliki V. Katsenou, Fan Zhang, Mariana Afonso, and David R. Bull

Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK
{angeliki.katsenou, fan.zhang, mariana.afonso, dave.bull}@bristol.ac.uk

ABSTRACT

In this paper we compare the performance of two state-of-the-art competing codecs, AV1 and HEVC, in the context of adaptive streaming. We specifically consider a Dynamic Optimizer (DO) methodology that is content-aware and selects the resolution of the video sequence after constructing the convex hull of the Rate-Quality curves of all considered resolutions. We start with an objective evaluation of the Dynamic Optimizer, based on both PSNR and VMAF quality metrics. The Rate-VMAF curves show an average of 6.3% BD-Rate gain of AV1 over HEVC, while the Rate-PSNR curves show an average BD-Rate loss of 1.8%. We then report subjective tests which evaluate the perceived quality of the selected bitstreams generated by the two codecs. In this case it was found that, for most rate points, the difference in the perceived quality between HEVC and AV1 is not significant.

Index Terms— Subjective Evaluation, Video Compression, AV1, HEVC, Adaptive Video Streaming.

1. INTRODUCTION

The increasing demand for streamed video content over the Internet is addressed mainly by dynamically adaptive streaming using methods such as MPEG DASH. In adaptive streaming, providers create several versions for each video sequence (or segment thereof) using different encoding parameters, with the aim of satisfying the requirements and limitations of each end user. For example, the version of a video sequence streamed to a client's smart phone is likely to be at a different bit rate and resolution to that streamed to his/her smart TV. From a provider's perspective, the aim is to deliver the best video quality at the lowest possible bit rate [1]. To this end, the appropriate selection of the video coding parameters is an important but challenging task. For example, assuming a channel bandwidth of 1Mbps in a fixed bit rate encoding system, and a sequence encoded at 1280×720 , 1920×1080 and 3840×2160 , the question would be which resolution should be streamed: the version with the highest resolution that possibly carries coding (e.g. blocking) artifacts or a version at a lower resolution that might result in scaling artifacts (e.g. blurriness).

In the traditional “bit rate ladder”, normally used in fixed bit rate encoding systems, a static content-agnostic pairing of bit rates and resolutions is decided prior to encoding [2, 3]. Recent research [1, 4, 5] has however focused on the optimization of the traditional “bit rate ladder” approach. The Dynamic Optimizer approach employed

by Netflix (DO) [1, 4, 6], optimizes the Rate-Quality (RQ) characteristics by running several trial encodings of each video at different resolutions and different quantisation levels. This information is then used to construct the convex hull of the RQ curves (particularly, scaled Rate-PSNR in [1] and scaled Rate-VMAF in [4]) that reveals the optimal parameters for the available bit rate range. A similar approach is presented in [5], where the authors measure the actual usage of millions of video clips to create probability distributions of available bandwidth and viewport sizes. These probability distributions feed the optimization process to ensure that the video quality is preserved while reducing the required bit rate compared to existing techniques. The DO approach [4, 6] uses VMAF [7] as a criterion for resolution selection because VMAF has been shown to achieve higher correlation with perceived quality compared to MSE-based metrics.

Several authors [3, 6, 8, 9] have compared the performance of the recently released AV1 codec to HEVC [10] and VP9 for adaptive streaming. These comparisons were however based solely on objective metrics - none of them report subjective results. In contrast, recent subjective comparisons between AV1 and HEVC [11–13] do not consider the impact of the adaptive streaming environment.

In this paper, we compare the performance of DO with two codecs, AV1 (version 0.1.0-9647-ga6fa0877f) [14] and HEVC (HM16.18) [15], firstly based on objective quality metrics (PSNR and VMAF) and secondly by way of a subjective assessment. The aim of this study is twofold: (i) to assess the coding gains of the DO using the competitive AV1 and HEVC codecs, and (ii) to verify the effectiveness, in terms of perceived quality, of using VMAF as the criterion for adaptive streaming.

The structure of the paper is as follows: In Section 2 the testing methodology is described and the selected dataset is presented. In Section 3 the objective evaluation of the DO for AV1 and HEVC is provided. Next, follows the subjective evaluation in Section 3. Finally, conclusions and future work are outlined in Section 5.

2. TESTING METHODOLOGY

2.1. Dataset

Nine source sequences from five different UHD (3840×2160) 10 bit sources, Harmonic [16], BVI-Texture [18], Netflix [19], Huawei [17] and B<>com [17], were selected as shown in Fig. 1. The sequences were spatially downsampled to 1920×1080 (1080p), 1280×720 (720p), and 960×544 (544p) using Lanczos-3 filters [20]. The specific sequences were selected for two reasons. First, sequences with diverse spatial and temporal content were selected: three static sequences with no camera motion (S1, S2, S3), three sequences with dynamic textures (S4, S5, S6), and three sequences with complex camera motion (S7, S8, S9). Secondly, another reason for selecting some of these sequences (S2, S5, S7, S9) is that they have often been

The work presented was supported by the Leverhulme Early Career Fellowship and by the Engineering and Physical Sciences Research Council (EPSRC), EP/M000885/1.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Fig. 1: Sample frames from the test dataset.

Fig. 2: Scatter plot of SI and TI for the test dataset.

used for MPEG standardisation activities. The spatial and temporal information of the dataset (SI and TI) [21] exhibit good coverage as reflected in Fig. 2. High SI values correspond to diverse spatially content and are usually high for highly textured areas. High TI values typically characterize high levels of motion.

2.2. Test conditions

We consider two codecs: AV1 and HEVC, selected because they are generally considered to be competing technologies with similar performance [3, 6, 8, 11, 13, 22]. Each codec was configured such that the structural delay was limited to 16 frames (i.e. the group of pictures (GOP) length is equal to 16) with random access intervals set to 64 frames for all test videos. In order for the codec comparison results to have a practical impact in an adaptive streaming environment, Dynamic Optimization, which determines the optimal resolution at a given bitrate/quality, was applied. Therefore, multiple encoding at different quantization levels and resolutions, i.e. {1080p, 720p and 544p}, were produced and only the optimal rate points were considered for the final rate-distortion performance assessment.

Results were generated for the target bit rates listed in Table 2. Each codec produced bitstreams that did not exceed these target bit rate points. These were selected based on a preliminary encoding of the test sequences using AV1 (see Table 1 for the codec configuration parameters). This is because AV1 has a narrower range of possible bitrates for a given resolution. In order to achieve adequate coverage of the perceptual reconstruction qualities using VMAF, fixed target VMAF values within the range 20-90 were chosen with a significant perceptual difference between rate points (approximately 12 points VMAF difference). It was thus ensured that the first rate point (VMAF = 90) was not at a saturated level of quality.

Table 1: Codec Configurations.

Codec	Configuration
HM16.18 [17]	Default main10 profile with random access configuration, GOP-Size=16 and IntraPeriod=64.
AV1, version 0.1.0-9647-ga6fa0877f [14]	<pre> \$encoderDir -o \$BitstreamFN --verbose --passes=2 --limit=300 --i420 --profile=0 --cpu-used=1 --width=\$width --height=\$height --input-bit-depth=10 --fps=60000/1001 --kf-max-dist=64 --kf-min-dist=64 --arnr-maxframes=7 --arnr-strength=5 --lag-in-frames=16 --aq-mode=0 --bias-pct=100 --minsection-pct=1 --maxsection-pct=10000 --end-usage=q --cq-level=\$qp --min-q=0 --max-q=63 --auto-alt-ref=1 --max-gf-interval=16 --min-gf-interval=4 --frame-parallel=0 --threads=1 --tile-columns=0 --color-primaries=bt709 --ivf \${sequenceName} </pre>

Table 2 reports the selected resolutions for each sequence, for each codec and at each rate point after constructing the convex hull of the rate-VMAF curves across all considered resolutions. As can be seen, different combinations of resolutions are selected according to content. It is important to note that, for both codecs, low resolutions are selected for sequences with high texture and motion masking, namely for Calming Water, ToddlerFountain and RollerCoaster. On the other hand, for more static sequences such as Myanmar and CatRobot, higher resolutions occur more frequently in the RQ convex hull.

3. OBJECTIVE EVALUATION

In this section, we report an objective evaluation of the DO using AV1 and HEVC. We calculated the PSNR and the VMAF [7] values resulting in the RQ curves in Figs. 3 and 4, respectively. As can be seen, the difference between the RQ curves for the two different codecs is small, and varies according to content and quality metric used. For example, AV1 appears to perform better for sequences with high amount of motion, such as the CalmingWater, the ToddlerFountain and the RollerCoaster for both PSNR and VMAF.

In order to quantify the differences between the RQ curves for the two codecs, we computed the Bjøndegeard Delta bit Rate (BD-Rate) [23] for both Rate-PSNR and Rate-VMAF curves. We annotate the subfigures in Figs. 3 and 4 with the BD-Rate of AV1 over HM. Overall, we observe a similar performance from the two codecs with an average of 1.8% BD-Rate loss for the Rate-PSNR curves and 6.3% gain for the Rate-VMAF curves when we compare AV1 over HM.

4. SUBJECTIVE EVALUATION

4.1. Testing Setup and Methodology

The viewing room employed conformed to home environment conditions, as outlined in BT.500-13 [24]. A consumer display, a SONY KD65Z9D, with a peak luminance of 410 cd/m² (measured using a Konica Minolta CS-2000 spectroradiometer), 1920 × 1080 spatial resolution, BT.2020 colour space (full range), and a refresh rate of 60 fps was used. The viewing distance was set to 3 × H, where H denotes the height of the screen, as indicated in [25]. All special features of the display were deactivated (e.g. motion smoothing, cinematic profiles). Finally the decoded sequences were all upsampled to the highest resolution, i.e. 1080p, using the Lanczos-3 filter.

According to BT.500 [24], when it is not possible to provide the full range of quality, the double-stimulus method is recommended.

Table 2: Target bit rate points (in kbps) and resolution selections per sequence after applying the DO methodology for AV1 and HM.

No.	Sequences	Target bit rates [kbps]					HM					AV1				
		R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
S1	AirAcrobatic	300	550	900	1750	3400	720p	1080p	1080p	1080p	1080p	720p	720p	720p	720p	720p
S2	CatRobot	950	1500	2100	3100	5500	1080p	1080p	1080p	1080p	1080p	720p	720p	1080p	1080p	1080p
S3	Myanmar	2050	3500	5500	8100	10400	1080p	1080p	1080p	1080p	1080p	720p	1080p	1080p	1080p	1080p
S4	CalmingWater	1500	3000	6500	12400	20500	544p	544p	544p	544p	720p	544p	544p	544p	544p	720p
S5	ToddlerFountain	3000	6000	13000	20500	35000	544p	544p	544p	544p	720p	544p	544p	544p	544p	720p
S6	LampLeaves	5000	8000	14000	20500	33500	720p	720p	1080p	1080p	1080p	544p	544p	720p	720p	720p
S7	DaylightRoad	900	1200	1800	3000	5400	720p	1080p	1080p	1080p	1080p	720p	720p	1080p	1080p	1080p
S8	RedRock	450	700	1050	1700	2450	720p	720p	1080p	1080p	1080p	544p	720p	720p	720p	720p
S9	RollerCoaster	550	700	1475	2300	3550	544p	544p	720p	720p	1080p	544p	544p	720p	720p	720p

Fig. 3: PSNR-Rate curves for the compared codecs and all tested sequences. In each subfigure legend the BD-Rate gains of AV1 over HM are reported.

Based on the defined Double Stimulus Continuous Quality Scale (DSCQS) protocol from BT.500-13 [24], the presentation structure during our experiment was as follows: each encoded sequence was paired with its original version and was presented twice; the presentation order of the pairs was random and different per participant. The total duration of the experiment was 61min and it was split in two consecutive sessions with a 10 minute break in between.

20 observers (19 non-experts and 1 expert) participated in the experiment, aged from 20-40, and with a proportion of 7 females over 13 males. All participants had normal or corrected-to-normal colour vision (verified with a Snellen chart). Prior to the processing of the scores, we performed screening of the participants using the outlier rejection method that BT.500 recommends [24]. None of the participants were rejected.

4.2. Results and Discussion

Differential Mean Opinion Scores (DMOS) were calculated for each test condition and linearly scaled to the range 0-100 (Bad to Excellent). The first results of this study are illustrated in Fig. 5, where the 100-DMOS scores are plotted against the bit rates. It can be observed that the subjective results show similar RQ trends for both

Fig. 4: VMAF-Rate curves for the compared codecs across all tested sequences. In each subfigure legend, the BDRate gains of AV1 over HM are reported.

codecs.

After performing the significance test using one-way Analysis of Variance (ANOVA) between paired AV1 and HM sequences, only six rate points were indicated as significantly different. This is illustrated by the p -values for each one of the 45 encoded pairs of AV1/HEVC sequences as depicted in Fig. 6. In six rate points HM is significantly better than AV1 ($p < 0.05$): at R4 for AirAcrobatic; at R1-R4 for Myanmar; and at R2 for LampLeaves.

For the Myanmar sequence, coding artifacts on semantically important parts of the videos can be observed. An example of these artifacts is shown in Fig. 7. Although the general picture quality is good for both codecs, the heads of the monks walking are not properly reconstructed for AV1 (a combination of blockiness and basis pattern effect is observed). A similar but smoother effect is observed in the higher bitrate range. Given that this sequence does not contain high levels of motion (the only motion/change is due to the monks walking) these areas are the natural regions of foveation. These artifacts were highlighted by the participants at the end of their sessions as the reason for the low objectives scores for the AV1 videos.

For the other two cases of significant difference in opinion scores, we note that the selected resolution by the DO is lower for

Fig. 5: 100-DMOS and standard error versus bit rate (in kbps) for AV1 and HM using DO.

Fig. 6: p -values for the pairs of all rate points of the encoded test sequences (every other 5 bins correspond to the 5 rate points of each sequence). The red dashed line illustrates the p -value threshold of statistically significant difference.

AV1. So the DO decision to select the lower resolution is attributed to a marginally higher VMAF value that was not aligned to the perceived quality.

This experiment demonstrates that, although DO selects different resolutions for the two different codecs, their performances do not significantly differ. For the considered dataset, lower resolutions are generally selected for AV1 compared to HM. Despite this, for most rate points the subjective tests and the statistical analysis of the collected scores showed that the perceived quality is similar. This is also the case for sequences with high texture and motion masking (S4 Calming Water, S5 ToddlerFountain and S9 RollerCoaster).

Another important observation is that, after performing a one-way ANOVA between successive bit rate points for both codecs, there is no significant difference for all sequences at the last two rate points. For example, there is no significant difference for the last three rate points of S9 RollerCoaster in the subjective RQ curve of HM (same for the RQ of AV1). This contradicts the observed monotonicity of the objective RQ curves for the two higher rate points (for some sequences this expands to lower rate points as well). This is an important finding for further research on the sensitivity and saturation of PSNR, VMAF and other objective quality metrics at high

(a) AV1 encoded sample frame.

(b) AV1 encoded cropped patch.

(c) HM encoded sample frame.

(d) HM encoded cropped patch.

Fig. 7: Example frame (#187) in (a) using AV1 and in (c) using HM. (b) and (d) illustrate the same magnified cropped patches of the AV1 and HM compression artifacts on semantically important regions.

quality levels.

5. CONCLUSION

In this paper, we report a subjective evaluation of AV1 and HEVC for an adaptive streaming scenario using the DO methodology. The subjective results confirm that the two codecs do not significantly differ in most cases despite some indicated difference in objective assessments. Important observations related to content, resolution selection and the monotonicity of the objective quality metrics derive from both the objective and subjective RQ curves. These observations trigger future research in two main directions: (i) the optimisation of the adaptive streaming methodologies with respect to content features or a combination of objective metrics and (ii) the improvement of objective quality metrics to better reflect perceived quality and its saturation in visually lossless levels.

6. REFERENCES

- [1] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 1484–1488.
- [2] S. Lederer, C. Müller, and C. Timmerer, "Dynamic Adaptive Streaming over HTTP Dataset," in *Proceedings of the 3rd Multimedia Systems Conference*. 2012, pp. 89–94, ACM.
- [3] A. Zabrovskiy, C. Feldmann, and C. Timmerer, "A practical evaluation of video codecs for large-scale http adaptive streaming services," in *25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 998–1002.
- [4] Ioannis Katsavounidis, "Dynamic optimizer - a perceptual video encoding optimization framework," <https://medium.com/netflix-techblog/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f>.
- [5] C. Chen, Y. Lin, S. Benting, and A. Kokaram, "Optimized Transcoding for Large Scale Adaptive Streaming Using Playback Statistics," in *25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 3269–3273.
- [6] I. Katsavounidis and L. Guo, "Video codec comparison using the dynamic optimizer framework," in *Proc. SPIE*, Aug. 2018.
- [7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "The NETFLIX tech blog: Toward a practical perceptual video quality metric," <http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html>, note = [Online; accessed 2018-08-04].
- [8] L. Guo, J. De Cock, and A. Aaron, "Compression Performance Comparison of x264, x265, libvpx and aomenc for On-Demand Adaptive Streaming Applications," in *Picture Coding Symposium (PCS)*, June 2018, pp. 26–30.
- [9] J. Sole, L. Guo, A. Norkin, M. Afonso, K. Swanson, and A. Aaron, "The NETFLIX tech blog: Performance comparison of video coding standards: an adaptive streaming perspective," <https://medium.com/netflix-techblog/performance-comparison-of-video-coding-standards-an-adaptive-streaming-perspective-d45d0183ca95>, [Online; accessed 2018-12-13].
- [10] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.
- [11] P. Akyazi and T. Ebrahimi, "Comparison of Compression Efficiency between HEVC/H.265, VP9 and AV1 based on Subjective Quality Assessments," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018.
- [12] D. Grois, T. Nguyen, and D. Marpe, "Coding efficiency comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC encoders," in *Applications of Digital Image Processing XL, SPIE*, 2017.
- [13] A. Seixas Dias, S. Blasi, F. Rivera, E. Izquierdo, and M. Mrak, "An Overview of Recent Video Coding Developments in MPEG and AOMEDIA," in *IBC*, Sept. 2018.
- [14] "Alliance for Open Media (AOM) Git repository," <https://aomedia.googlesource.com/aom/>, note = [Online; accessed 2018-05-30].
- [15] HEVC Reference Software (HM) Repository, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/, [Online; accessed 2018-02-20].
- [16] Harmonic Inc 4K demo footage, <https://www.harmonicinc.com/4k-demo-footage-download/>, [Online; accessed 2017-05-01].
- [17] K. Sharman and K. Sühring, "Common Test Conditions for HM video coding experiments," Tech. Rep., document JCTVC-AC1100 of JCT-VC, Oct. 2017.
- [18] Miltiadis Alexios Papadopoulos, Fan Zhang, Dimitris Agrafiotis, and David Bull, "A video texture database for perceptual compression and quality assessment," in *IEEE Inter. Conf. on Image Processing*, 2015, pp. 2781–2785.
- [19] Netflix, "Video test media," <https://media.xiph.org/video/derf/>, [Online; accessed 2018-08-04].
- [20] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.
- [21] Stefan Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.
- [22] T. Nguyen and D. Marpe, "Future Video Coding Technologies: A Performance Evaluation of AV1, JEM, VP9, and HM," in *Picture Coding Symposium (PCS)*, June 2018, pp. 31–35.
- [23] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," Tech. Rep., 13th VCEGM33 Meeting, Austin, TX, 2001.
- [24] ITU-T, "Recommendation BT.500-13: Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., 2012.
- [25] ITU-T, "Recommendation P.910: Subjective video quality assessment methods for multimedia applications," Tech. Rep., 2008.