This WACV 2023 Workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Subjective and Objective Video Quality Assessment of High Dynamic Range Sports Content

Zaixi Shang, Yixu Chen, Yongjun Wu, Hai Wei, Sriram Sethuraman
Amazon Prime Video
Seattle, WA

Abstract

High Dynamic Range (HDR) video streaming has become more popular because of the faithful color and brightness presentation. However, the live streaming of HDR, especially of sports content, has unique challenges, as it was usually encoded and distributed in real-time without the post-production workflow. A set of unique problems that occurs only in live streaming, e.g. resolution and frame rate crossover, intra-frame pulsing video quality defects, complex relationship between rate-control mode and video quality, are more salient when the videos are streamed in HDR format. These issues are typically ignored by other subjective databases, disregard the fact that they have a significant impact on the perceived quality of the videos. In this paper, we present a large-scale HDR video quality dataset for sports content that includes the above mentioned important issues in live streaming, and a method of merging multiple datasets using anchor videos. We also benchmarked existing video quality metrics on the new dataset, particularly over the novel scopes included in the database, to evaluate the effectiveness and efficiency of the existing models. We found that despite the strong overall performance over the entire database, most of the tested models perform poorly when predicting human preference for various encoding parameters, such as frame rate and adaptive quantization.

1. Introduction

High dynamic range (HDR) imaging is a set of technologies that expand the dynamic range of brightness and colorfulness conveyed in the digital video signal and allow the artistic intent to be accurately rendered on different display devices. HDR as next generation imaging technology has prevailed among the streaming service provider and mobile devices.

The efficiency of video compression and robustness of the encoded video quality largely depends upon the quality metric used to perform optimal encoding parameters selection. Subjective quality measurement is not practical at scale, and hence, objective video quality metric (OVQM) that are highly correlated with subjective quality become critical. OVQMs help service providers ensure that delivered streams have a high perceived video quality, yet are optimal from a delivery cost perspective. OVQMs developed for standard dynamic range (SDR) rarely work well for HDR live sports video. Since the live sports HDR contents have unique challenges as it was usually encoded in single-pass fast encoding setting and distributed in real-time without the Video on Demand (VoD) post-production workflow. To build such OVQMs and improve existing OVQMs performance on sports contents, an HDR sports video quality metric dataset is necessary.

In this paper, we present a new HDR video quality dataset named HDR-Sports. We benchmark existing HDR and SDR video quality metrics on this dataset to evaluate the performance and efficiency of those OVQMs for live sports applications.

2. Related Work

Research on HDR video quality has increased during the last few years [5, 18, 19, 20, 23]. The majority of research is focused on VoD applications, but live sports broadcasting with HDR is still an unexplored field. As a result, the HDR quality metrics [18, 19, 20] do not generalize well without retraining on the live sports video quality assessment (VQA) dataset because of the following two reasons:

1. The live sports HDR workflow focused more on reflecting the high dynamic range true color instead of made it visually pleasing with artistic intents in the post-production workflow in VoD.

2. Live streaming sports video quality assessment also has its unique challenges as the high-motion sports content has to be encoded in small segments with faster single-pass encoding setting to reduce the distribution delay. Defects like intra-frame (I-frame) pulsing that rarely exist in VoD 2-pass slow encoding setting will affect the live video quality.

On the other hand, the existing efforts in live streaming VQA [9, 12, 17, 10, 25, 8, 29] are greatly biased towards SDR video format. Thus, an HDR dataset containing sports material is crucial for creating new video quality measurements or optimizing those that already exist. Additionally, a model may be trained for both VoD and live HDR applications by integrating the prior VoD HDR dataset [26] with VoD contents using the anchor video method [21].

3. Scopes and Novelties

This dataset tried to include multiple scopes into a single database design for the HDR live sports scenario. The following points are the principles we followed when designing this study.

1. Two different encoding modes Quality-Defined Variable Bitrate (QVBR) [3] and Constant Bitrate (CBR) are included in our encoding settings to study the impacts on video quality of the rate-control mode under the same bitrates.

2. I-frame pulsing issue is a specific video defect for live application which appears under fixed key-int [4] length setting under low bitrates conditions. This dataset includes different flicker adaptive quantization settings which tried to reduce the I-frame quantization difference with adjacent frames to alleviate the flicker or "pop" on I-frames. [2].

3. The bitrate-resolution ladders are designed in a way where some common bitrates existed across different resolutions to help us construct a marginal model to improve the resolution crossover prediction accuracy.

4. The chosen bitrates should cover a wide range of quality levels and are well separated with perceivable quality different within the same resolution.

5. Besides resolution crossover, frame rate crossover is also considered. All sports source videos are high frame rate (HFR). Because of the size limit of the study, we only focus on finding the frame rate crossover at 540p resolutions. At an estimated crossover quality level based on the past study [26], we include both HFR and standard frame rate (SFR). Above those quality levels, we only include HFR videos and vice versa.

6. Selected anchor videos from LIVE-HDR dataset [26] were also included in this study to combine our previous datasets with this dataset for joint training on different types of content.

4. DETAILS OF SUBJECTIVE STUDY

This section explains the designing the subjective study, the issue we are targeting to solve and the choices of encoding parameters setting.

4.1. Source Sequences

As shown in Table 1, we obtained 4 different game match source videos from content partners, including 15 videos from English Premier League (EPL), 8 videos from UEFA Champions League (UCL), 12 videos from National Football League (NFL), and 7 videos from Association of Tennis Professionals (ATP). The NFL videos are 10-bit SDR videos with BT.709 color gamut, 1920x1080 resolution, and the others are HDR10 videos, with PQ transfer function and BT.2020 color primaries and 3840x2160 resolution. All the videos are 50 frames per second. A total of 42 sports video clips with a duration of 6 to 9 seconds are cut from the sources with different scene types (close-up or long shots) and different-level of motions.

4.2. Subjective Testing Design

We recruited 140 participants to rate the quality of the video clips in a lab with living room ambient lighting. 207 30-minute sessions were conducted in parallel on 4 TVs following single-stimulus absolute category rating with hidden reference (ACR-HR) method [22]. Each viewer view 7 contents (including 1 anchor content from previous HDR databases) with about 154 videos (including anchor). The 42 sources are divided into 7 groups and each subject viewed 2 out of 7 groups.

3 LED lights for each TV were used to produce 200 lux incident illumination on the TV. The viewing distance is 1.5 times the height of the TV following the recommendation [13]. The 4 55-inch UHD HDR TVs used are two identical Samsung UN55RU8000F and two identical LG OLED55C9PUA.

After a training session introducing the user interface, the subjects would watch each short video clip and then rate the quality of the video on a slider bar verbally marked by "Bad", "Poor", "Fair", "Good" and "Excellent" from lowest to the highest quality. The scores were sampled as integers from [0, 100] without showing to the subjects. Each video was rated by around 30 subjects.

Over 32,000 human ratings from 140 subjects were collected during this study. The final mean opinion score (MOS) is generated using the SUREAL method [16] as described in Section 5.2.

4.3. Encoding Parameter Choices

To best represent live streaming application scenarios, we used the AWS Elemental Media Live encoder L882AE [1] with the software version 2.23.4 to generate the High

Sports	Number of clips	Resolution	Frame rate	Color space	Transfer function	Bit-depth
EPL	15	3840x2160	50	Rec.2020	ST2084 (PQ)	10
UCL	8	3840x2160	50	Rec.2020	ST2084 (PQ)	10
APT	7	3840x2160	50	Rec.2020	ST2084 (PQ)	10
NFL	12	1920x1080	50	Rec.709	BT.1886 (Gamma)	10

Table 1: Source video clips format

Efficiency Video Coding (HEVC) HDR encoding in real-time.

1002 processed video sequences (PVS) were generated from the 42 source sequences with various encoding setting under QVBR and CBR. With the additional 66 anchor videos from LIVE-HDR [26] and HDR-AQ dataset, there are 1068 total videos in HDR-Sports dataset, which is the largest in-lab VQA dataset to the best of our knowledge.

Following the principles in Section 3, we designed the following encoding parameter sets in Table 2 for CBR rate control encoding mode and Table 3 for QVBR mode. Among the 42 source videos, 7 are encoded with QVBR mode and the remaining 35 are CBR mode.

In Table 2, each row represents different bitrates that were used in Elemental CBR HEVC encoder. Each column represents the resolutions. Each entry is the extra encoding parameters that were applied to a specific bitrate resolution combination. The first setting before in each entry is frame rate, where for each video we could have HFR and SFR versions together or separately. HFR is used for bitrate higher than 750 Kbps and resolution larger than 540p. SFR is applied to lower bitrates and smaller resolutions. The second setting is the adaptive quantization (AQ) setting for I-frame pulsing issue. For each resolution, we select a single bitrate to have an extra PVS with flicker AQ enabled and study its impact on I-frame pulsing defects. "S&T" means spatial and temporal AQ is enabled, while "S&T/F" means an extra PVS with flicker AQ enabled is added.

In Table 3, each row represents different quality level settings in Elemental QVBR HEVC encoder. Each column represents the resolutions. Each entry has 3 extra encoding parameters. The 1st setting is the maximum bitrate setting in terms of Mbps for elemental QVBR mode. The 2nd and 3rd setting is the same as CBR setting in Table 2.

5. Processing of Subjective Scores

5.1. Internal Consistency

The internal consistency of the collected score is a good indicator of the quality of subjective scores as well as an upper limit of the possible best performance of a VQA model. We studied the internal consistency of the subjective scores as follows. The subjects were randomly divided into two equal-sized groups and the mean score of each video was obtained within the subjects in the group. The mean scores are obtained from both groups and we calculated the correlation between the two groups. We plotted the scatter plot of one random split in Figure 1. We repeated this process for 1000 splits and noted the Spearman correlation coefficient (SROCC) and Pearson correlation coefficient (PLCC) for each time. The median SROCC is 0.9672 and the median PLCC is 0.9666.

To find the consistency of the ratings of certain testing scopes, we further divide the videos into a few groups and calculated the internal consistency of each group. First, we consider the present of different frame rate. The first group of videos consist of all videos at the resolution of 396p and 540p since a portion of the videos has multiple frame rates. The second groups of video are the videos that have both flicker AQ version and no flicker AQ version, *i.e.*, the videos of 2160p resolution at 2.6Mbps, the videos of 1080p, 720p, 540p and 396p resolution at 750kbps. The third group of videos are the videos encoded using the QVBR mode. We show the scatter plot of one split in Figure 1 and the median SROCC and PLCC in Table 4. As noticed, both the internal consistency on all videos and the QVBR videos are high and the rest two groups are lower, since both groups include video that cover a wide quality range and the quality between each encoded video are perceptually separable. On the other hand, the video from Group 1 and Group 2, are dominantly videos of the same bitrate but encoded either at different frame rate and resolution, or with different AQ options. The perceptual differences are significantly less obvious than the average of the database, and it's more challenging to distinguish these videos. However, the correlation coefficients are still fairly high and this reflects the high quality of the subjects ratings from this study. It can also be noticed that the SROCC is lower than PLCC in Group 2. This reflects that the subjects don't have a clear preference within both AQ options. The quality scores are similar but these scores still faithfully reflect the absolute quality of these videos, so the SROCC is lower but the PLCC is still high.

5.2. SUREAL Score Calculation

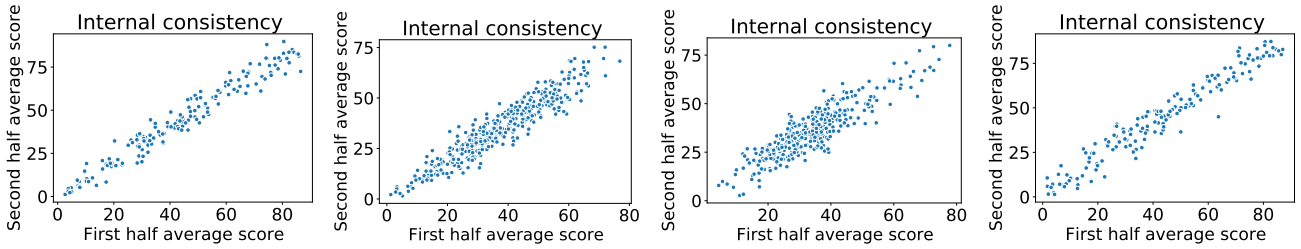
In order to provide clean and reliable training data, we use the SUREAL method to recover the quality score [15]. This method accounts for the noise and unreliability of human ratings and attributes the noise to subject bias, subject inconsistency, content ambiguity and outliers. By jointly

	3840x2160	1920x1080	1280x720	960x540	704x396
60 Mbps	HFR, S&T^	-	-	-	-
30 Mbps	-	HFR, S&T^	-	-	-
12 Mbps	HFR, S&T	HFR, S&T	-	-	-
6 Mbps	HFR, S&T	HFR, S&T	HFR, S&T	-	-
2.6 Mbps	HFR, S&T/F	HFR, S&T	HFR, S&T	-	-
1.8 Mbps	-	-	-	HFR/SFR, S&T	SFR, S&T
750 Kbps	-	HFR, S&T/F	HFR, S&T/F	HFR/SFR, S&T/F*	SFR, S&T/F
300 Kbps	-	-	HFR, S&T	SFR, S&T	SFR, S&T

Table 2: CBR mode encoding parameter settings. The row header is the bitrate used in CBR mode. The cells marked by ^ means that the setting is used to generate visually lossless encoding pseudo reference videos for smooth playback and full reference algorithm input. The cells marked by * means that the flicker AQ only applied to HFR setting but not the SFR one.

	3840x2160	1920x1080	1280x720	960x540	704x396
9.00	60, HFR, S&T^	30, HFR, S&T^	-	-	-
8.33	35, HFR, S&T	16, HFR, S&T	7.5, HFR, S&T	3, HFR/SFR, S&T	3, SFR, S&T
7.66	-	15, HFR, S&T	-	-	-
6.33	-	-	4, HFR, S&T	1.5, HFR/SFR, S&T	1.5, SFR, S&T
6.00	-	9, HFR, S&T	-	-	-
5.33	12, HFR, S&T	-	-	-	-
4.33	-	-	3, HFR, S&T	-	-
3.00	-	1.8, HFR, S&T/F	2, HFR, S&T/F	1, HFR/SFR, S&T/F*	1, SFR, S&T/F
2.00	5.6, HFR, S&T	-	-	-	-

Table 3: QVBR mode encoding parameter settings. The row header is the quality levels ranging from 1 to 10. The 1st number in each cell is the maximum bitrate in Mbps for QVBR mode. The cells marked by ^ means that the setting is used to generate visually lossless encoding pseudo reference videos for smooth playback and full reference algorithm input. The cells marked by * means that the flicker AQ only applied to HFR setting but not the SFR one.



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

Figure 1: The scatter plot of the average score from both half of subjects in one random split. (a) all the videos; (b)-(d) some groups of videos. (b) Group 1: 540p and 396p videos; (c) Group 2: Videos with Flicker AQ options; (d) Group 3: Videos encoded in QVBR mode.

estimating the subjective quality, bias, consistency of subjects and ambiguity of video contents, this method is able to fully exploit the subject information, and handles the problem of outliers without subject rejection using any heuristic methods. The SUREAL method finds a Maximum Likelihood (ML) estimate of the scores. Using this method, we represent the opinion scores $s_{i_k j}$ as random variables $S_{i_k r}$ as the summation of three parts

$$S_{i_j} = \psi_j + \Delta_i + \nu_i X, \tag{1}$$

where ψ_j is the true quality of video j, Δ_i represents the bias of subject i, the non-negative term ν_{i_k} represents the inconsistency of subject i, and $X \sim N(0,1)$ are i.i.d. Gaussian random variables. These unknown parameters are obtained using maximum likelihood estimation. We plotted the estimated subject biases and their inconsistencies in Figure 2. It may be observed that both the subject biases and inconsistencies are quite dispersed. Thus, by accounting for the noise and unreliability of each subject, the underline quality score is recovered. The histogram of the recovered

Video group	SROCC	PLCC
All videos	0.9672	0.9666
Group 1	0.9304	0.9332
Group 2	0.8677	0.9020
Group 3	0.9707	0.9732

Table 4: The internal correlation of each group of videos. Group 1: 540p and 396p videos; Group 2: Videos with Flicker AQ options; Group 3: Videos encoded in QVBR mode.

score is shown in Figure 3. The scores cover a wide range from 0 to 100, and a large portion of video scores fall in the range of 30-50. This is consistent with the design of the database.

5.3. Subjective Score Analysis

We plotted the average score of the videos having resolutions of 540p and 396p in Figure 4. This plot reveals that human viewers have a mixed preference between 540p and 396p, but this trend changes at 750 Kbps and starting from 750 Kbps the human scores of 540p videos are significantly higher than those of 396p. This is because the 300 Kbps is a very tight bitrate budget, so some 396p videos are preferred by viewers, especially when some contents are not severely affected by scaling artifact, such as smooth areas. However, as the bitrate increase, compression artifact is reduced and viewers prefer compression over more scaling. Similar trend is observed in frame rate. As the bitrate budget gets larger, the quality scores of high frame rate videos have a higher slope than standard frame rate, reflecting viewers' preference of smooth motion over less compression artifact.

As shown in Figure 5 and Figure 6, different videos have very different spatial information (SI), temporal information (TI) and rate quality characteristics. Taking 4 videos as examples, the video "epl2" and "UCL6" with the presence of high motion indicated by high TI, viewer prefer videos with lower resolution with less compression artifacts under similar bitrates. While for clips with lower TI, e.g. "epl18" and "tennis1", the scaling artifacts contributes more to the quality as bitrate increase and clear crossovers are present.

5.4. Combining Multiple Datasets Using Anchor Videos

Different VQA datasets are conducted under different environments. The MOS score collected can be affected by different ambient lights condition [26] and displays with different pixel densities [11]. What's more, the relative video quality in a subjective rating session also affects the rating due to short-term memory effect. Therefore directly combine the MOS scores from different VQA datasets is not a good practice. Therefore we followed the anchor video method in [21] to combine our previous HDR VQA datasets.

66 anchor videos including reference videos and the corresponding compressed videos from our previous datasets HDR-LIVE [26] and HDR-AQ (not publicly available) are included in this study. The set of anchor video includes a few clips with different coding complexity and all its encoding variants to cover a wide range of quality levels. Without losing generality, we mapped the subjective scores s of the anchor videos from the HDR-LIVE database and HDR-AQ database to the scores of these anchor videos in the new database o using the non-linear transform [24] in Equation (2). The parameters $\beta_i (i = 1, 2, 3, 4)$ of the non-linear mapping are obtained by minimizing the MSE between the original scores and mapped scores. The non-linear transform is given by

$$f(s) = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{s-\beta_3}{|\beta_4|}}} + \beta_2 \tag{2}$$

where The parameters are initialized with $\beta_1 = max(o)$, $\beta_2 = min(o)$, $\beta_3 = mean(s)$, $\beta_4 = std(s)/4$ during the minimization. The non-linear mapping is then applied to all videos to map all the MOS from the HDR-LIVE and HDR-AQ dataset.

As shown in Figure 7, the same anchor videos don't have the same scores when the subjective tests are conducted under their unique environments. The non-linear fitting could help reduce this bias among different datasets.

6. Evaluation of Objective VQA Algorithms

We evaluated several objective VQA algorithms on this database. As in [27], we use the SROCC, the PLCC, and the Root Mean Square Error (RMSE) as metrics to evaluate the performance of the objective VQA algorithms. Here we present the results of a few popular FR VQA models: peak signal-to-noise ratio (PSNR), SSIM, MS-SSIM [30], SpEEDQA [6], ST-RRED [28] and VMAF [7]. To account for the difference in resolution and frame rate, we upscaled all the videos to 4K using bicubic interpolation and 50 fps by duplicating the frames. The results of the objective VQA algorithms are obtained from the altered videos. The result is shown in Table 5.

6.1. Prediction Accuracy of frame rate crossover

We test whether the objective models are accurate on the frame rate crossovers using 540p videos. We used those videos encoded at 540p and both 750 kbps and 1800 kbps to determine whether the objective quality models are able to correctly predict human's preference between the videos of different frame rate. On these videos, we evaluate the percentage of the videos that the prediction aligns with human preference. For example, if both SUREAL score and VMAF prediction agrees video "epl2" at 540p, 750 kbps

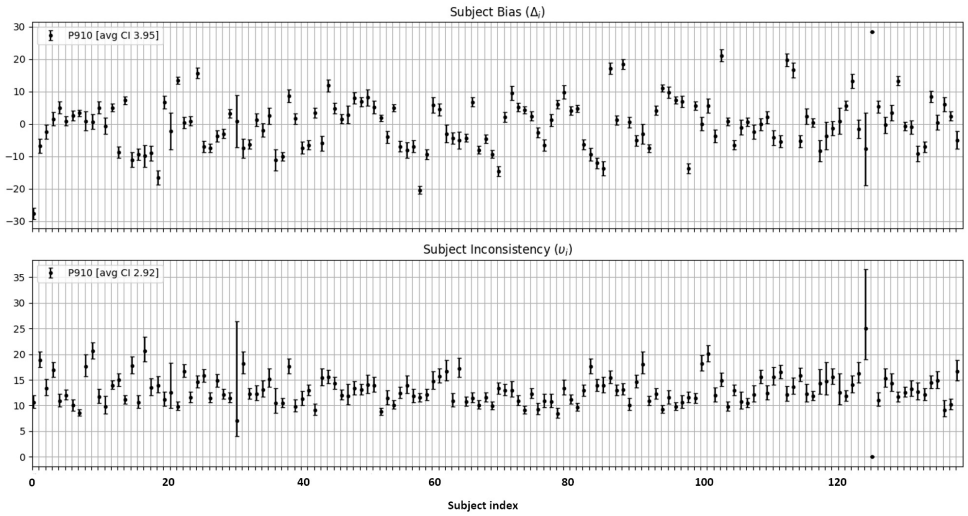


Figure 2: Histograms showing distributions of MOS, $ZMOS$, and SUREAL scores.

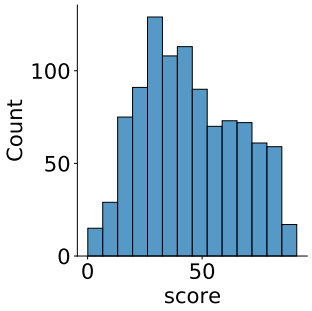


Figure 3: Histograms of recovered SUREAL scores.

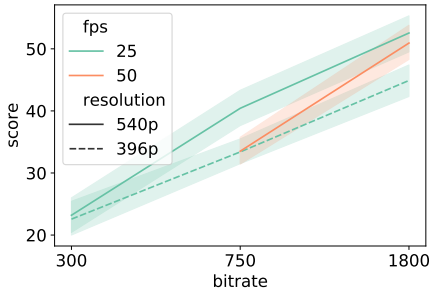


Figure 4: Scores of videos at 540p and 396p resolutions, by frame rate.

obtains a higher score at 50 fps than at 25 fps, VMAF gets one pair of videos correct. We show the results in Table 6. Both VMAF and VMAF_4K have the highest correction rate. Note SpEED-QA and PSNR, VMAF and VMAF_4K have the same percentage of correction due to the same number of correct judgement.

Model	SROCC	PLCC
SSIM	0.6062	0.7282
MS-SSIM	0.6205	0.7397
ST-RRED	0.8374	0.8330
SpEED-QA	0.6453	0.2978
PSNR	0.5943	0.6946
VMAF	**0.8742**	**0.8828**
VMAF_4K	0.8703	0.8555

Table 5: The correlation of the predicted score using evaluated VQA models against human score on the entire database. The top performing model is boldfaced.

Model	Percentage of correction
SSIM	54.28%
MS-SSIM	57.14%
ST-RRED	67.14%
SpEED-QA	58.57%
PSNR	58.57%
VMAF	85.71%
VMAF_4k	85.71%

Table 6: The percentage of contents of which the model prediction is aligned with human judgement for the frame rate crossover test.

6.2. Preference Prediction Accuracy on AQ Videos

We take the part of the videos that have two AQ options and evaluated the performance of the objective models using SROCC, PLCC. We also evaluated the prediction ac-

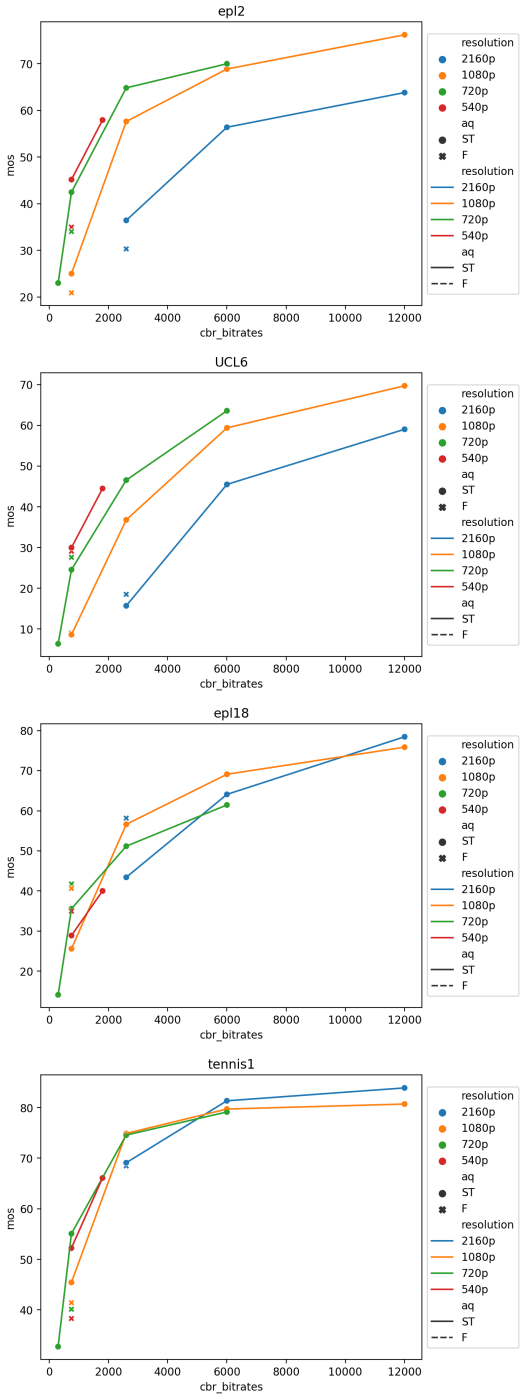


Figure 5: Some example quality-bitrate curves for CBR mode HFR videos in HDR-Sports dataset. Y-axis is the Sureal MOS and X-axis is the bitrates in CBR mode.

curacy of human preference on the videos that have scores are significantly difference between MAX AQ strength and low AQ strength. We report the accuracy of each algorithm in Table 7. Similarly, we also calculated the percentage of

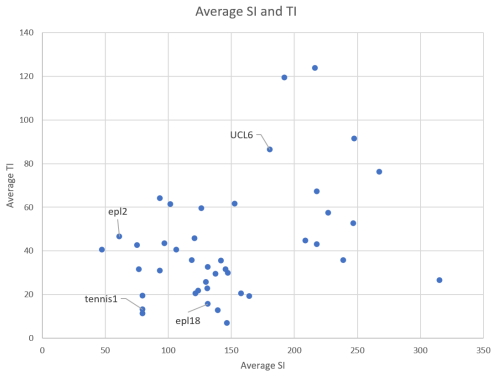


Figure 6: Average spatial information (SI) and temporal information (TI) of HDR-Sports dataset using method from ITU-P.910 [14]

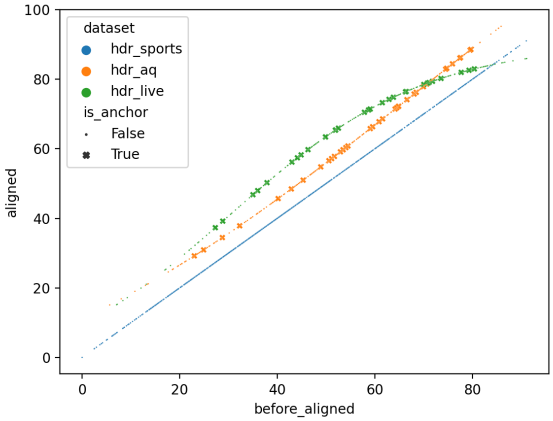


Figure 7: Non-linear fitting using anchor videos to combine different datasets. x axis is the score before alignment and y axis is the score after alignment. The cross markers represent the anchor videos, while the small dots represent non-anchor videos.

contents that the model correctly predicted the human preference over different AQ mode. The results are reported in Table 8. All the compared algorithms have really low percentage of correction, and in fact, a lot of the algorithms predicted the same small number of pairs of videos right and resulted in the same percentage shown in Table 8.

6.3. Prediction Accuracy of the QVBR Mode

We further evaluated the performance of the objective models on the videos that are encoded with the QVBR mode and the CBR model separately, shown in Table 9 and Table 10.

Model	SROCC	PLCC
SSIM	0.2785	0.5666
MS-SSIM	0.2474	0.5113
ST-RRED	0.4489	0.4271
SpEED-QA	0.2128	-0.0314
PSNR	0.2816	0.5365
VMAF	**0.7258**	**0.7469**
VMAF_4k	0.7153	0.7015

Table 7: The correlation of the predicted score using evaluated VQA models against human score obtained from the part of the videos that have two AQ options. The top performing model is boldfaced.

Model	Percentage of correction
SSIM	23.07%
MS-SSIM	23.07%
ST-RRED	21.53%
SpEED-QA	22.30%
PSNR	21.53%
VMAF	21.53%
VMAF_4k	21.53%

Table 8: The percentage of contents of which the model prediction is aligned with human judgement for the AQ test.

Model	SROCC	PLCC
SSIM	0.6549	0.7432
MS-SSIM	0.6837	0.6777
ST-RRED	0.9231	0.4360
SpEED-QA	0.6888	0.3355
PSNR	0.6535	0.7363
VMAF	**0.9452**	**0.7767**
VMAF_4K	0.9441	0.7767

Table 9: The correlation of the predicted score using evaluated VQA models against human score obtained from the part of the videos that encoded using the QVBR mode. The top performing model is boldfaced.

6.4. Results Analysis

From Table 5, it's easy to notice that the tested OVQMs obtained reasonably good results on the new database, with VMAF and ST-RRED leading the performance. On the other hand, VMAF performs significantly better than any other compared models. It prediction over the trade-off between better spatial quality versus smoother motion is mostly accurate. However, all the tested OVQMs were not accurate in predicting the preference over different AQ

Model	SROCC	PLCC
SSIM	0.5949	0.7284
MS-SSIM	0.6038	0.7328
ST-RRED	0.8261	0.8300
SpEED-QA	0.6341	0.2729
PSNR	0.5845	0.6961
VMAF	**0.8700**	**0.8785**
VMAF_4K	0.8663	0.8408

Table 10: The correlation of the predicted score using evaluated VQA models against human score obtained from the part of the videos that encoded using the CBR mode. The top performing model is boldfaced.

mode. One reason is that the predicting the preference over AQ model is a much more difficult question because it's not only about predicting an overall quality, but it involves the understanding of the complex relationship between scene types and local quality. Lastly, the correlation obtained from the QVBR encoded videos are higher than the correlations from the entire database, while VMAF is still leading the performance.

7. Conclusion

We studied different HDR live streaming scenarios and determined the most important encoding parameters that are typically overlooked by other subjective quality studies. We created the database that includes faithful representation of the selected scopes and conducted the largest in-lab subjective VQA study. Using gathered data, we identified the effect of each encoding parameters as well as encoding options. We also demonstrated the performance of the state-of-the-art OVQMs, on the entire new database, as well as on each specific portion of the data. This subjective and objective quality study reveals the most overlooked but critical problems in HDR video streaming. We anticipate this knowledge to be relevant for the design of new subjective VQA studies and objective VQA models.

References

[1] Aws elemental live l800 series. https://elemental-activations-hw-spec-sheets.s3.amazonaws.com/AWS+Elemental+Live+L882AE+Specification+Sheet.pdf. AWS Elemental Live L882AE Specification Sheet.

[2] Aws elemental live quantization controls. https://docs.aws.amazon.com/elemental-live/latest/ug/vq-quantization.html.

[3] Quality-defined variable bitrate. https://aws.amazon.com/media/tech/quality-defined-variable-bitrate-qvbr/.

[4] X265 command line options: keyint. https://x265.readthedocs.io/en/2.5/cli.html#cmdoption-keyint.

[5] Maryam Azimi, Amin Banitalebi-Dehkordi, Yuanyuan Dong, Mahsa T Pourazad, and Panos Nasiopoulos. Evaluating the performance of existing full-reference quality metrics on high dynamic range (hdr) video content. *arXiv preprint arXiv:1803.04815*, 2018.

[6] Christos G Bampis, Praful Gupta, Rajiv Soundararajan, and Alan C Bovik. Speed-qa: Spatial efficient entropic differencing for image and video quality. *IEEE signal processing letters*, 24(9):1333–1337, 2017.

[7] Netflix Technology Blog. Toward a practical perceptual video quality metric, Apr 2017.

[8] Joshua P. Ebenezer, Yixu Chen, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Subjective and objective quality assessment of high-motion sports videos at low-bitrates. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 521–525, 2022.

[9] Joshua P. Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, and Alan C. Bovik. No-reference video quality assessment using space-time chips. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2020.

[10] Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE Transactions on Image Processing*, 30:8059–8074, 2021.

[11] Abubkr Elmnsi, Niemah Osman, and Is-Haka Mkwawa. Mobile devices pixel density and video quality. In *2017 4th NAFOSTED Conference on Information and Computer Science*, pages 325–330, 2017.

[12] Xiaojun Hei, Yong Liu, and Keith W. Ross. Inferring network-wide quality in p2p live streaming systems. *IEEE Journal on Selected Areas in Communications*, 25(9):1640–1654, 2007.

[13] Recommendatios ITU-R. Recommendation 500-10; methodology for the subjective assessment of the quality of television pictures. *ITU-R Rec. BT. 500-10*, 2000.

[14] P.910 ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications. 07/2022.

[15] Zhi Li, Christos G Bampis, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments. *Electronic Imaging*, 2020(11):131–1, 2020.

[16] Zhi Li, Christos G. Bampis, Lukáš Krasula, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments, 2020.

[17] Zhi Li, Ali C. Begen, Joshua Gahm, Yufeng Shan, Bruce Osler, and David Oran. Streaming video over http with consistent quality. In *Proceedings of the 5th ACM Multimedia Systems Conference*, MMSys '14, page 248–258, New York, NY, USA, 2014. Association for Computing Machinery.

[18] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4), jul 2011.

[19] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Hdr-vqm: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015.

[20] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Study of high dynamic range video quality assessment. In Andrew G. Tescher, editor, *Applications of Digital Image Processing XXXVIII*, volume 9599, page 95990V. International Society for Optics and Photonics, SPIE, 2015.

[21] Alexander Raake, Silvio Borer, Shahid M Satti, Jörgen Gustafsson, Rakesh Rao Ramachandra Rao, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, et al. Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of uhd/4k: Itu-t p. 1204. *IEEE Access*, 8:193020–193049, 2020.

[22] ITU-T Recommendation. Itu-t p. 910. *Subjective video quality assessment methods for multimedia applications*, 2008.

[23] Martin Rerabek, Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi. Subjective and objective evaluation of hdr video compression. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, number CONF, 2015.

[24] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010.

[25] Zaixi Shang, Joshua P. Ebenezer, Alan C. Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Assessment of subjective and objective quality of live streaming sports videos. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021.

[26] Zaixi Shang, Joshua P Ebenezer, Alan C Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Subjective assessment of high dynamic range videos under different ambient conditions. *arXiv preprint arXiv:2209.10005*, 2022.

[27] Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31:1027–1041, 2021.

[28] Rajiv Soundararajan and Alan C Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2012.

[29] Maria Torres Vega, Decebal Constantin Mocanu, Jeroen Famaey, Stavros Stavrou, and Antonio Liotta. Deep learning for quality assessment in live video streaming. *IEEE Signal Processing Letters*, 24(6):736–740, 2017.

[30] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.