Towards Better Quality Assessment of High-Quality Videos

Suiyi Ling
suiyi.ling@capacites.fr
CAPACITÉS SAS
Nantes, France

Yoann Baveye
yoann.baveye@capacites.fr
CAPACITÉS SAS
Nantes, France

Deepthi Nandakumar
nandakd@amazon.com
Amazon Video
Bangalore, India

Sriram Sethuraman
sssethur@amazon.com
Amazon Video
Bangalore, India

Patrick Le Callet
patrick.lecallet@univ-nantes.fr
LS2N lab, University of Nantes
Nantes, France

ABSTRACT

In recent times, video content encoded at High-Definition (HD) and Ultra-High-Definition (UHD) resolution dominates internet traffic. The significantly increased data rate and growing expectations of video quality from users create great challenges in video compression and quality assessment, especially for higher-resolution, higher-quality content. The development of robust video quality assessment metrics relies on the collection of subjective ground truths. As high-quality video content is more ambiguous and difficult for a human observer to rate, a more distinguishable subjective protocol/methodology should be considered. In this study, towards better quality assessment of high-quality videos, a subjective study was conducted focusing on high-quality HD and UHD content with the Degradation Category Rating (DCR) protocol. Commonly used video quality metrics were benchmarked in two quality ranges.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing**;

KEYWORDS

datasets, HD, UHD, video quality assessment

ACM Reference Format:
Suiyi Ling, Yoann Baveye, Deepthi Nandakumar, Sriram Sethuraman, and Patrick Le Callet. 2020. Towards Better Quality Assessment of High-Quality Videos. In *1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications (QoEVMA'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3423328.3423496

1 INTRODUCTION

Nowadays, video streaming, especially the High-Definition and Ultra-High-Definition video contents, accounts for the maximal percentage of Internet downstream traffic [26]. At the same time, the expectation of Quality-of-Experience (QoE) of users grows significantly along with the rapid development of display manufacturers, high-quality video producers (film industry) and video streaming services. Video quality assessment (VQA) metrics are aimed at predicting the perceived quality, ideally, by mimicking the human visual system. Such metrics are imperative for achieving higher compression and for monitoring the delivered video quality. According to one of the most recent relevant studies [21], correlation between the subjective scores and the objective scores predicted by commonly used video quality metrics is significantly poorer in the high quality range (HD) than the ones in the lower quality range (SD). Thus, a reliable VQA metric is highly desirable for the high quality range to tailor the compression level to meet the growing user expectation.

The robustness of objective VQA metrics is grounded in subjective experiments, where ground-truth quality scores are collected from human observers. The quality (*e.g.*, whether there are noisy ratings) and the distinguishability (*e.g.* whether there are enough significant pairs) of the collected subjective data is crucial for the eventual development of quality metrics [14, 15]. It is emphasized in [21] that the subjective scores obtained with Absolute Category Rating (ACR) protocol do not consistently distinguish between quality levels in the high-quality region, even when significant quality differences were expected. In order to develop quality metrics with better distinguishability in high quality range, a more discriminative experimental protocol should be considered when collecting subjective data.

Based on the discussion above, towards better quantification of the perceived quality of high-quality videos, we conducted subjective studies on high-quality HD and UHD contents. To enhance the distinguishability of the obtained subjective data in the high-quality range, the Degradation Category Rating (DCR) methodology was utilized. When benchmarking the objective quality metrics in the low and high quality range, in addition to the different types of correlation against the subjective metrics, the "Krasula" framework is employed to compare the area-under-the-curve (AUC) for the "different-or-similar" and "better-or-worse" discrimination tasks.

2 RELATED WORK

In the last decades, many subjective studies were conducted for better quality assessment of UHD contents. One of the very first subjective test considering UHD contents was presented in [2] for the investigation of using high efficiency video coding (HEVC) for 4K-UHD TV broadcasting services. Yet, this study is limited as only

QoEVMA'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8158-1/20/10…$15.00
https://doi.org/10.1145/3423328.3423496

three bitrates were tested. The study summarized in [3] compared the perceived quality of HD and UHD at the same bitrate, encoded with HEVC. Another subjective experiment for 4k UHD content was shown in [27], but only 6 contents were considered. Recently, in order to benchmark the performances of the popular and emerging coding techniques, *e.g.,* the H.264 Advanced Video Coding (AVC), HEVC, VP9, Audio Video Coding Standard (AVS2) and AOMedia Video 1 (AV1), a large scale subjective 4k dataset was collected. However, these studies did not lay specific emphasis to the high-quality range. In [21], a more recent study was conducted to better measure and optimize video quality for HD, especially for high quality ranges. Nevertheless, in this study (1) the quality range was divided based on the *encoding resolution*, *i.e.,* the frame sizes, instead of using the Mean Opinion Scores (MOS); (2) UHD contents were not taken considered; (3) only the correlation coefficients between subjective scores and predicted objective scores were considered when benchmarking the quality metrics, which does not lend itself to a deeper analysis; (4) Absolute Category Rating was employed for the subjective ratings, while more discriminating subjective protocols need to be considered as discussed earlier.

3 SUBJECTIVE EXPERIMENT SETUP

3.1 Content selection

Content selection is a crucial step in a test design [24]. The contents used in this study were selected from a candidate set that contains 229 uncompressed 1 minute-long videos from a top streaming service provider. These videos had different source resolutions (from UHD to 640x480) and frame-rates. From these clips, 10-seconds long single-scene UHD and HD videos were selected.

In a nutshell, to select representative source videos that cover a wide-range of content characteristics and ambiguity behaviors, we first select representative features based on their correlation with the ambiguity level of the contents. Contents are then clustered using the selected features so that sequences belonging to the same cluster have similar characteristics. With the representative clusters, representative samples could be further selected from each cluster [13]. Details are provided in the following sub-sections.

3.1.1 Content features. In this study, 4 types of content features, and a total of 21 dimensions are considered [18]. The details are given below (**dim** denotes the dimension of the feature):

(1) Spatial complexity (3 **dim**): Spatial information (SI) increases with the detail or sharpness visible within each frame, *e.g.*, high contrast edges, fine detail and textures. A video with low spatial information will have large areas with similar pixel values. The computation of this feature is based on applying the "Sobel" filter on each frame of the video (on the luminance component). Afterwards, the standard deviation of the response of each frame (Y channel) is calculated as the spatial complexity of this frame [8]. Except for computing the maximum values across frames as employed in [8], the minimum and mean values are also considered in this study.
(2) Temporal complexity (3 **dim**): Temporal information (TI) increases in proportion to individual pixels that change value from one frame to the next. Temporal information does not correspond to moving objects, but to changing pixels. The computation of TI is based on the difference between the pixel values (of the luminance plane) of successive frames. As detailed in [8], to get a global value for the whole video, the standard deviation over space (of the values of the pixels of the difference frames) is computed. The maximum value over time is then calculated to get the final TI score for the clip. In this study, we also consider calculating the minimum and mean values.
(3) Colorfulness (3 **dim**): This is an important visual feature having a significant impact on the perceptual quality of a scene. Recently, in [1], the authors showed the performance of the state-of-the-art colorfulness metrics with the help of subjective experiment data. From this study, the metric proposed by Hasler *et. al.* [6] was selected, since it overcomes the others. Analogous temporal pooling methods are utilized after obtaining the colorfulness score per frame.
(4) Textural features (12 **dim**): Contrast is one of the most important textural indicators, strongly related to the physiological procedure of perceiving image quality. In addition to this, many textural features [11, 19], such as entropy, homogeneity and correlation (of a pixel to its neighbor over the whole image) have been used to characterize the image. These textural features can be extracted directly by exploiting Gray Level Co-occurrence Matrix (GLCM) [5]. More specifically, in this study, the contrast, entropy, homogeneity and correlation were computed using the GLCM, and the corresponding extracted features are denoted as GLCM-Contrast, GLCM-Entropy, GLCM-Homogeneity, and GLCM-Correlation respectively. After computing the four texture features frame-wise, similarly, minimum, maximum and mean values are computed across all the frames.

3.1.2 Feature Selection based on Content-Ambiguity . As pointed out in [17], some of the contents tend to be more complex for observers to judge confidently. For instance, a content with numerous camera motions in a dark scene tends to be more ambiguous than a content with few motion changes in a bright scene. Therefore, Li *et al.* [17] proposed a novel method (Li's recovery model) to recover the ground-truth of observers' opinions from noisy raw data, by jointly estimating the perceived quality of the stimulus, the bias & consistency of observers, and the ambiguity of contents.

As summarized in Section 2, a subjective study has been conducted in a similar context [21], where observers' opinions were collected for 20 contents from a top streaming platform via Absolute Category Rating (ACR). By employing Li's recovery model with the subjective data collected in [21], the ambiguity scores of the contents could be then estimated, *i.e.,* how difficult the content is for the observer to rate or how uncertain the observer is when rating the content.

Intuitively, using the ambiguity scores as labels, the features that reveal the ambiguity-level of the contents could be obtained by using a certain feature selection. However, the dataset released in [21] contains only 20 contents, which is almost equivalent to the feature dimensions (*i.e.,* 21). To avoid the issue of "the curse of the dimensions" [22], we simply select the top features based on their ranking, where the rankings are calculated based on the correlation between each feature's value and the ambiguity score. The

features with a higher correlation are more relevant in capturing the ambiguity.

3.1.3 Contents Clustering & Selection. According to our observations, contents that have lower ambiguity scores tend to gather together while the ones with higher ambiguity scores group into another cluster. Therefore, we cluster the 20 contents from [22] into 2 clusters using k-means clustering in the selected feature space (more clusters could be obtained with more data points).

In modern adaptive bitrate streaming systems, source contents are commonly segmented into a set of shots. Thus, the 229 HD and UHD candidate sources were first split into 10s-shots using ffmpeg with a scene-cut detection filter. The selected features of the shots were then extracted. Subsequently, they were assigned to one of the ambiguity clusters obtained in the previous step, and thus could be labeled by the corresponding ambiguity cluster.

In addition to the ambiguity labels, the SI and TI features are also computed (using 'mean' for temporal pooling, as SI-mean and TI-mean were found to be the most ambiguity-relevant dimensions, see Section 4.1 for more details). When selecting the contents, we ensured that:

- 20 HD SRCs and 10 UHD SRCs are selected
- 80% of the selected contents have been assigned to the cluster of larger ambiguity score (as they are more difficult for observers to rate), and the remaining 20% were selected from the low-ambiguity cluster,
- selected shots should be extracted from different contents (*i.e.*, two 10s shots from the same original 1 minute-content cannot be selected)
- selected shots should also cover a wide range of SI-TI complexity.

The thumbnails of the final selected HD and UHD SRCs are shown in Figure 1 and Figure 2.

3.2 PVS Generation

For each of the 20 HD SRCs selected in previous section, 13 quality points were generated for three encoding resolutions (1080p, 720p, 540p). For the 10 UHD contents also, 13 quality points were generated for three different encoding resolutions (2160p, 1080p, 720p). Thus, for both HD and UHD, each content is denoted as C_k in the remaining of this paper is associated with 40 representations R_n (*i.e.,* 39 PVS and the original uncompressed source).

The encoding recipes cover a wide range of quality. Only a subset of the encoding recipes are used in the subjective tests. To select this subset of PVS, an adapted JND test [29] has been performed by 5 experts to select 4 PVS for each of the selected content and each of the three encoding resolutions using the strategy described below. The reason to consider experts for this JND-based subjective test was twofold. First, experts are commonly considered as the "golden eyes": they know where to look to detect artifacts and have the ability to repeat a test in case of uncertainty to ensure the reliability of collected annotations. Second, experts are experienced in test design: they are familiar with the rating scales and can anticipate how naive observers interpret it.

In the JND expert test, a calibrated UHD "Grundig Finearts 55 FLX 9492 SL" with a 55-inch screen size was used. The viewing distance for UHD contents was set as 1.5H, where H is the height of the screened video, as recommended in ITU-R BT.1769 [10]. The viewing distance was set to 3H for HD contents. PVS with various encoding resolutions were upscaled by the video player to match the resolution of the source content (*i.e.*, viewing resolution).

Finally, 4 PVS (original source, 1st JND point from original source, lowest quality PVS, a higher quality PVS at 1 JND from the lowest quality PVS) were selected for each resolution. This process ensures to select enough significant pairs for the following subjective test and aims at focusing on sensitive regions where the Rate-Quality (R-Q) curves of different encoding resolutions cross each other. An example of original quality points for a single source and selected PVS is shown in Figure 3.

3.3 Experimental protocols

Degradation Category Rating (DCR) [25] always presents stimuli in pairs. More specifically, the PVS is always shown to the observers after the SRC, *i.e.,* the reference. DCR is able to avoid relevant errors that the ACR method may miss since subjective data collected via DCR has the minimum impacts from the biased opinions of the observers about the content, *e.g.,* whether the subject likes or dislikes the production quality, or different comprehensions of the quality scale. Furthermore, as the PVS is always compared to its reference in a continuous order, it is more suitable when the impairment is small. Therefore, it affords higher sensitivity, and is of greater advantages in comparing high-quality contents like HD, UHD contents.

Based on the discussion above, we employed DCR (5-point scale) for our subjective study. More concretely, the test was divided into 3 sessions including 1) UHD-DCR, 2) HD1-DCR (containing PVS from half of the selected HD SRCs), and 3) HD2-DCR (containing the remaining half PVS). Before the experimental session, participants had to sign a consent form and instructions were given. To avoid visual fatigue, the viewers were asked to take a five minutes break after half of the test samples. The total duration of each test session was approximately 55 minutes. The viewing environment was the same as the one for the JND expert test with the same UHD display.

In total, 72 remunerated viewers participated in this subjective experiment to ensure that each of the 3 test sessions would be annotated by 24 participants. All of the viewers are non-expert in subjective experiment, image processing or 4K related fields. All participants have either normal or corrected-to-normal visual acuity. Correct visual acuity was assured prior to this experiment using a Monoyer chart. Ishihara color plates were used to test color vision. All of the viewers passed the pre-experiment vision check.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Relevant Content Feature Selection

Table 1 shows the SRCC between the extracted features mentioned in previous section with the ambiguity scores for all devices for high quality content. Bolden features are the proposed features to be used during the k-mean clustering video selection.

4.2 Outlier Detection

Three methods to identify outliers in raw DCR subjective scores are considered in this section:

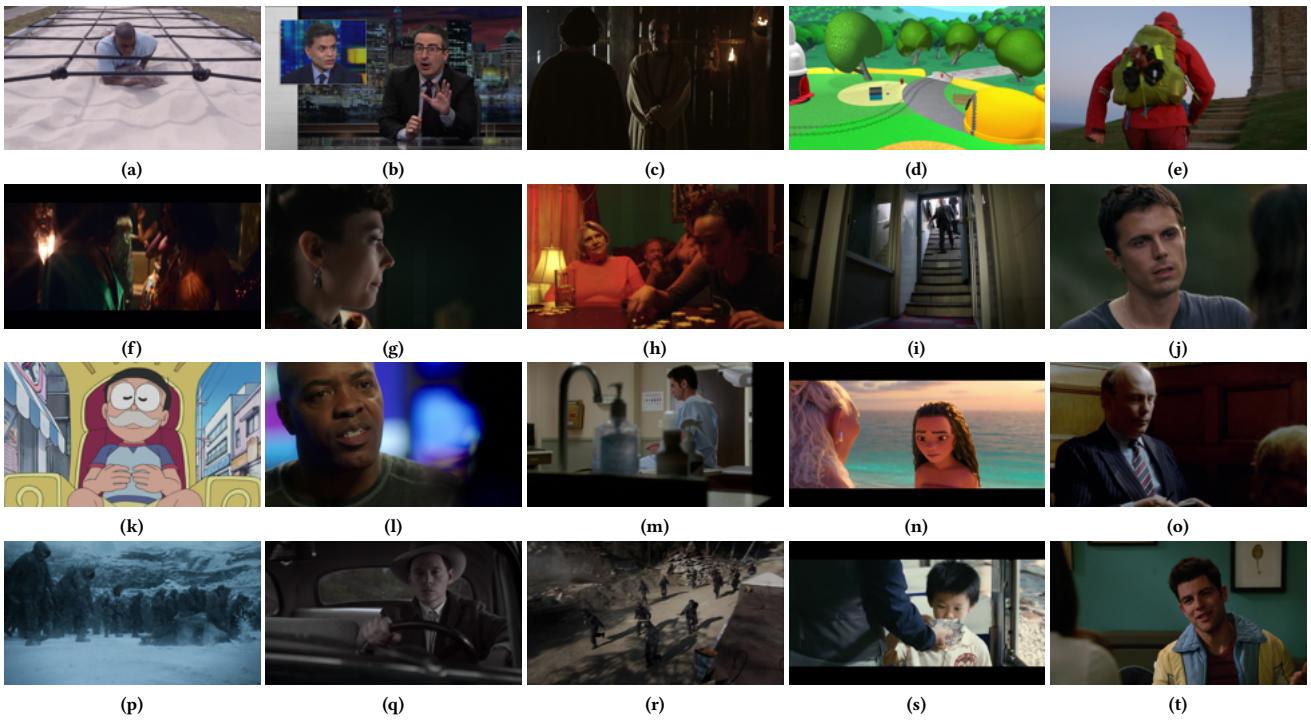

Figure 1: Thumbnails of the 20 selected HD contents.

Figure 2: Thumbnails of the 10 selected UHD contents.

- VQEG HDTV Annex I [28]
- ITU-R BT.1788 [7]
- ITU-R BT.500-12 [9]

Scores from subjects detected as outliers by those three methodologies are considered as outliers and are not considered in the following analysis. After the screening procedure, two outliers are found in UHD DCR and HD2 DCR test sessions while three outliers are found in HD1 DCR test session.

Results from HD1 and HD2 test sessions are combined for the following analysis and referred as the HD test scenario (sub-set).

4.3 Distinguishability of DCR vs. ACR

Ideally, if one subjective methodology is more distinguishable than another one, it should provide us with more significantly different pairs with a similar setting. Although the selected HD contents and PVS selection methodology from [21] are different from our study, some hints could be provided by the comparison of the 'significantly different pairs number' between DCR and ACR. To decide whether a pair of stimuli are significantly different or not, the Fisher exact test was employed. As there is a total of three groups of subjective data in [21], including the data collected using tablet, phone and screen (monitor), we only consider the data collected with screen. Table 2 and Table 3 show the total number of significant pairs obtained in both subjective tests. It could be observed that, the ratio

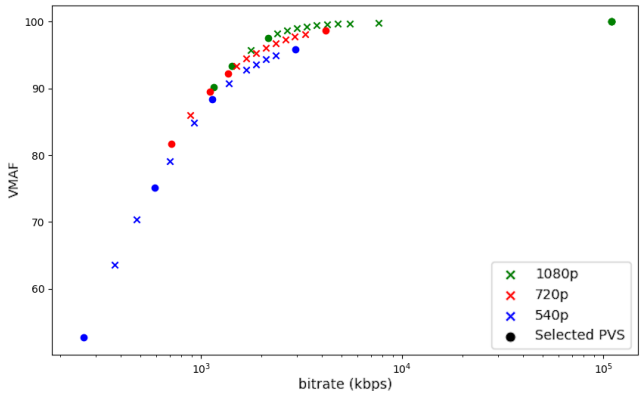

Figure 3: Selected PVS using JND searches for 3 encoding resolutions of a single SRC.

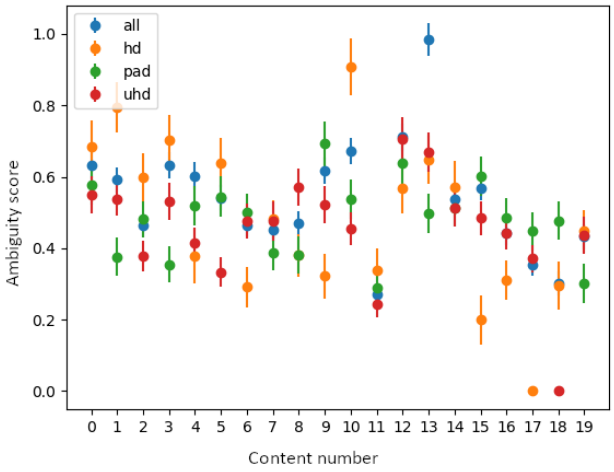

Figure 4: The ambiguity scores of the SRCs from the dataset [21] using Li's recovery model for each device.

of significant pairs obtained in the two tests for HD contents are 81.4% and 46.7% respectively. Therefore, by using DCR methodology, a significantly larger number of significant pairs could be obtained compared to ACR. There is a clear sign showing that the DCR methodology is of higher distinguishability compared to ACR.

4.4 Bench-marking Objective Video Metrics

4.4.1 Tested quality metrics. The Peak Signal to Noise Ratio (PSNR) [30], Structural Similarity Index (SSIM) [31] Multi-Scale Structural Similarity (MS-SSIM) [32], and Video Multi-Method Assessment Fusion (VMAF) [16] are four commonly used VQA metrics in the streaming industry. Therefore, similar to the experimental setup in [21], we bench-mark these four metrics.

4.4.2 Evaluation measures. To compare the performances of these VQA metrics, the following widely employed criteria are used:

- **Pearson Linear Correlation Coefficient (PCC):** measures the linear correlation between the subjective and objective scores [23].

Table 1: Ranking results of the candidate feature set.

Feature name	SRCC	Ranking
SI mean	-0.4496	5
SI min	-0.4361	7
SI max	-0.4361	7
TI mean	-0.3278	13
TI min	0	0
TI max	-0,1338	18
Colorfulness mean	-0.2661	16
Colorfulness min	-0.1293	19
Colorfulness max	-0.3504	11
GLCM-Contrast mean	-0.3759	10
GLCM-Contrast min	-0.4857	3
GLCM-Contrast max	-0.3203	14
GLCM-Homogeneity mean	0.2737	15
GLCM-Homogeneity min	0.2556	17
GLCM-Homogeneity max	0.4827	4
GLCM-Entropy mean	0.4241	9
GLCM-Entropy min	0.3489	12
GLCM-Entropy max	0.5098	2
GLCM-Correlation mean	-0.4481	6
Correlation min	-0.5353	1
GLCM-Correlation max	-0.0225	20

Table 2: Number of significant pairs for each methodology/protocol HD test sessions.

	DCR	ACR
UHD	563	-
HD	1075	514

Table 3: Percentage of significant pairs for each methodology/protocol HD test sessions.

	DCR	ACR
UHD	85.3%	-
HD	81.4%	46.7%

- **Kendall's correlation coefficient (KCC):** measures of rank correlation.
- **Spearman Rank Order Correlation Coefficient (SCC):** measures of rank correlation.
- **Root Mean Square Error (RMSE):** measures how spread out the prediction errors are.

The results are shown in Table 4. Although VMAF outperforms the other compared video quality metrics, it still performs poorly on both the HD and UHD set, *i.e.,* with PCC values of 0.7878 and 0.6466 correspondingly. In general, the tested metrics performs worse on the UHD set than the on the HD set.

In order to better evaluate the performance of different metrics within different quality range, especially the ones in the high-end range, in this study, the methodology proposed by Krasula *et al.* [12] is used. Their proposed model assumes that when comparing two stimulus, the capability of an objective metric depends on its

Table 4: Performances of quality metrics evaluated by commonly used performance evaluation measures.

Metrics	PCC	SCC	KCC	RMSE
HD				
$VMAF_{HD}$	**0.7878**	**0.7936**	**0.6131**	**0.5689**
PSNR	0.6021	0.6168	0.4442	0.7376
SSIM	0.7218	0.7304	0.5634	0.6393
MS-SSIM	0.6887	0.7020	0.5264	0.6697
UHD				
$VMAF_{UHD}$	**0.6466**	**0.6210**	**0.4593**	**0.6646**
PSNR	0.4555	0.4249	0.3021	0.7756
SSIM	0.5055	0.4879	0.4085	0.7537
MS-SSIM	0.4759	0.4395	0.3166	0.7733

Table 5: Performances of quality metrics evaluated by the 'Krasula' framework [12].

Metrics	AUC_{BW}	AUC_{DS}	CC_0	AUC_{BW}	AUC_{DS}	CC_0
HD						
	Low Quality range			High Quality range		
$VMAF_{HD}$	**0.93**	**0.74**	**0.91**	**0.92**	**0.69**	**0.93**
PSNR	0.75	0.54	0.68	0.86	0.59	0.82
SSIM	0.88	0.70	0.87	0.91	0.67	0.93
MS-SSIM	0.87	0.69	0.86	0.89	0.63	0.90
UHD						
	Low Quality range			High Quality range		
$VMAF_{UHD}$	0.86	0.57	0.79	0.63	**0.63**	**0.71**
PSNR	0.75	0.52	0.72	**0.66**	0.53	0.63
SSIM	**0.98**	**0.70**	**0.95**	0.57	0.62	0.71
MS-SSIM	0.87	0.56	0.82	0.58	0.52	0.63

ability to make reliable decisions about 1) whether the stimuli are qualitatively different and 2) if they are, which of them is of a higher quality. In other words, the 'Krasula' model is based on determining the classification capabilities of the objective models considering 'Better or Worse' and 'Different or Similar' scenarios.

Similar to the procedure described in [20], the stimuli were first assigned to the *low-quality* group if their MOS is smaller than a threshold τ, otherwise to the *high-quality* one. In this study, τ is set as 3 instead of 2.5, as this division provides a more balanced number of low and high-quality pairs. Afterwards, for each *low/high-quality* group, pairs were formed considering all possible combination within the group. Then, the pairs were further divided into two sets, including 1) the set composed with significantly different pairs of stimuli, and 2) the one containing the other pairs without significant differences. To decide whether a pair of PVS/stimuli is significant different, significance test is conducted by taking the individual subjective scores (raw opinion scores from all the observers from the DCR subjective test as input. To determine whether the preference for one stimulus over another is statistically significant, in this study we employed the Bonferroni method [4].

After categorizing pairs within each quality range into one group that contain pairs with significantly different quality scores, and another one that contain similar pairs , the second step is to pre-process the predicted scores by calculating the difference between the predicted scores of each pair of stimuli to obtain the predicted preference (by the objective quality metrics) of one stimulus over another.

With the pre-processed objective predicted and subjective scores, the **'Difference vs. Similar' Analysis** can be then conducted to check how well can the objective metric distinguish between significantly different and similar pairs, especially in the *low/high-quality* range. In this analysis, it is assumed that the difference of the objective scores predicted by a well-performing model, should be larger for significant pairs than for the non-significant ones. With this assumption, the objective metric can be indirectly considered as a binary classifier with categories 'Difference' versus 'Similar'.

The capability of the objective metric of categorizing similar and significantly different pairs can be then determined by employing the receiver operating characteristic (ROC) analysis on these two sets (ROC quantifies how well are the two sets are separated). Then, the performance of the metric can be verified with the area under the ROC curve (AUC). In the following part of this manuscript, it is denoted as AUC_{DS}.

Another important analysis, which can be done with the pre-computed subjective and predicted scores, is the **'Better Vs. Worse Analysis'**. The goal of this analysis is to see whether the objective metric is capable of picking out stimuli that are of higher/lower quality. Similar to the previous analysis, the set of pairs significant difference could be further divided into two sets, including one of higher quality and the other one with lower quality. Similarly, the ROC could be employed on these two sets. The performance of the under-test objective models can be then evaluated by checking the AUC value of the 'Better vs. Worse' ROC (denoted as AUC_{BW}.

Apart from AUC_{BW}, correct classification in 0 (CC_0) defined in [12] is also used as another quantifier to evaluate the performance with respect to whether the stimuli of better quality are assigned with higher objective scores by the objective model. Readers are recommended to refer to [12] for more details.

The results are presented in Table 5. For HD contents, similar to the previous analysis, $VMAF_{HD}$ achieves the best performances in both low and high quality range. Yet, its corresponding AUC_{DS} values are only 0.748 and 0.698, which means that it distinguishes poorly whether stimuli within a pair are similar or significantly different. For UHD set, SSIM exhibits superior performance than the other metrics in the low quality range while PSNR and $VMAF_{UHD}$ are superior to SSIM in the high quality range. It is obvious that the performances of the considered quality metrics are unsatisfactory in the high quality range of UHD set, as the maximum AUC_{BW}, AUC_{DS} and CC_0 values are only 0.662, 0.631 and 0.716 respectively.

5 CONCLUSION

To better evaluate the quality of high-quality HD and UHD contents, we present a subjective study in this paper. The Degradation Category Rating methodology guided by expert pre-selection of PVS based on JND was utilized in the study to collect more distinguishable subjective data for high-quality contents. A decent number of significantly different pairs were obtained via the subjective test. After dividing the entire quality range into high and low quality sub-ranges based on the mean opinion scores, a novel performance evaluation methodology is employed to bring out the

discriminability of the considered objective quality metrics. According to the experimental results, we found that most of commonly used video quality metrics like VMAF performs poorly for UHD contents, especially in the high-quality range.

REFERENCES

[1] Critina Amati, Niloy J. Mitra, and Tim Weyrich. 2014. A Study of Image Colourfulness. In *Workshop on Computational Aesthetics*. 23–31.

[2] Sung-Ho Bae, Jaeil Kim, Munchurl Kim, Sukhee Cho, and Jin Soo Choi. 2013. Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services. *IEEE Transactions on Broadcasting* 59, 2 (2013), 209–222.

[3] Kongfeng Berger, Yao Koudota, Marcus Barkowsky, and Patrick Le Callet. 2015. Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[4] J Martin Bland and Douglas G Altman. 1995. Multiple significance tests: the Bonferroni method. *Bmj* 310, 6973 (1995), 170.

[5] Robert M. Haralick, Its'hak Dinstein, and K. Shanmugam. 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics* SMC-3, 6 (1973), 610–621.

[6] David Hasler and Sabine E. Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, Vol. 5007.

[7] ITU. 2007. Methodology for the subjective assessment of video quality in multimedia applications. *Recommendation ITU-R BT.1788* (2007).

[8] ITU. 2008. Subjective video quality assessment methods for multimedia applications.

[9] ITU. 2009. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT.500-12* (2009).

[10] ITU-R BT.1769. 2006. Parameter values for an expanded hierarchy of LSDI image formats for production and international programme exchange. *Int'l Telecommunication Union* (2006).

[11] ME Jernigan and F D'astous. 1984. Entropy-based texture analysis in the spatial frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1984), 237–243.

[12] Lukáš Krasula, Karel Fliegel, Patrick Le Callet, and Miloš Klíma. 2016. On the accuracy of objective image and video quality models: New methodology for performance evaluation. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[13] Junghyuk Lee, Toinon Vigier, Patrick Le Callet, and Jong-Seok Lee. 2018. A Perception-Based Framework for Wide Color Gamut Content Selection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. 709–713.

[14] Jing Li, Suiyi Ling, Junle Wang, Zhi Li, and Patrick Le Callet. 2020. GPM: A Generic Probabilistic Model to Recover Annotator's Behavior and Ground Truth Labeling. *arXiv preprint arXiv:2003.00475* (2020).

[15] Jing Li, Rafal Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet. 2018. Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation. In *Advances in Neural Information Processing Systems*. 3475–3485.

[16] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. 2018. VMAF: The journey continues. *Netflix Technology Blog* (2018).

[17] Zhi Li and Christos G Bampis. 2017. Recover subjective quality scores from noisy measurements. In *2017 Data Compression Conference (DCC)*. IEEE, 52–61.

[18] Suiyi Ling, Yoann Baveye, Patrick Le Callet, Jim Skinner, and Ioannis Katsavounidis. 2020. Towards Perceptually-Optimized Compression Of User Generated Content (UGC): Prediction Of UGC Rate-Distortion Category. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[19] Suiyi Ling, Patrick Le Callet, and Zitong Yu. 2018. The role of structure and textural information in image utility and quality assessment tasks. *Electronic Imaging* 2018, 14 (2018), 1–13.

[20] Suiyi Ling, Jesús Gutiérrez, Ke Gu, and Patrick Le Callet. 2019. Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 204–216.

[21] Deepthi Nandakumar, Yongjun Wu, Hai Wei, and Avisar Ten-Ami. 2019. On the accuracy of video quality measurement techniques. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[22] Ivan V Oseledets and Eugene E Tyrtyshnikov. 2009. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM Journal on Scientific Computing* 31, 5 (2009), 3744–3759.

[23] Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58 (1895), 240–242.

[24] Margaret H. Pinson, Marcus Barkowsky, and Patrick Le Callet. 2013. Selecting scenes for 2D and 3D subjective video quality tests. *EURASIP Journal on Image and Video Processing* 2013, 1 (Aug. 2013), 50–61.

[25] ITUTP Recommendation. [n.d.]. ITU-Tp. 913. ([n. d.]).

[26] Sandvine. 2019. The global internet phenomena report. https://www.sandvine.com/press-releases/sandvine-releases-2019-global-internet-phenomena-report. [Online; accessed 04-June-2020].

[27] Rafael Sotelo, Jose Joskowicz, Matteo Anedda, Maurizio Murroni, and Daniele D Giusto. 2017. Subjective video quality assessments for 4K UHDTV. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 1–6.

[28] VQEG. 2007. VQEG HDTV Phase I Test Plan.

[29] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. 2017. VideoSet: A large-scale compressed video quality dataset based on JND measurement. *Journal of Visual Communication and Image Representation* 46 (2017), 292–302.

[30] Zhou Wang and Alan C Bovik. 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal processing magazine* 26, 1 (2009), 98–117.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

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