Benchmarking of quality metrics on ultra-high definition video sequences

Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi
Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

***Abstract*—The performance of objective quality metrics for high-definition (HD) video sequences is well studied, but little is known about their performance for ultra-high definition (UHD) video sequences. This paper analyzes the performance of several common objective quality metrics (PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM) on three different 4K UHD video sequences using subjective scores as ground truth. The findings confirm the content-dependent nature of most metrics (with VIF being the only exception), which has been reported previously for standard and high resolution video sequences. PSNR showed the lowest correlation with ground truth quality scores when the analysis was performed for all contents at once and thus is not recommended as a general metric for video quality, while VIF showed the highest Pearson (0.83) and Spearman (0.87) correlation coefficients and may be used as a general purpose metric. On the other hand, all studied metrics were accurate in distinguishing different quality levels for the same content. The results of several fittings between metric values and subjective ground truth scores demonstrated that logistic fitting provides the highest correlation. The results also indicated a shift in metrics values between synthetic and natural contents.**

***Index Terms*—Video quality assessment, objective metrics, ultra-high definition**

I. Introduction

Recent advances in hardware (both for acquisition and rendering) and compression algorithms (HEVC) have increased research interest in ultra-high definition (UHD) video content, such as 4K and 8K UHD. While UHD is positioned and marketed as an increasingly immersive experience, little is understood about its impacts on human visual perception and how to measure it. However, since UHD video content is meant for viewing in high-end home cinemas or in movie theaters, where high visual quality is one of the major factors constituting the quality of experience, an accurate measurement of its perceptual visual quality is important.

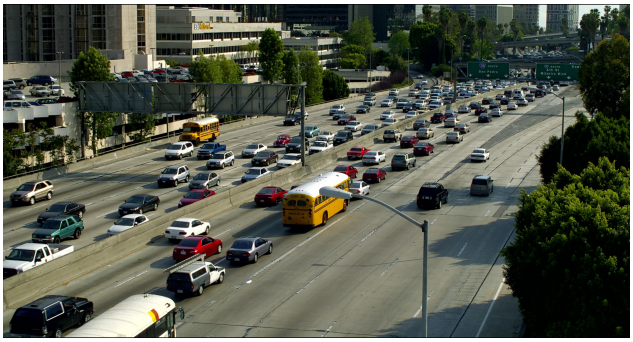
While the subjective evaluations remain the most accurate means of quantifying video quality, as measurements are done using human observers, they are time consuming and tedious, yielding such evaluations impractical in many applications. In practice, objective measures are often a more preferable more efficient alternative to subjective evaluations. However, objective metrics, especially the most commonly used, i.e., PSNR, are often criticized for their inaccurate prediction of perceptual video quality. The main reason for the inaccuracy is the lack of established exact relationship between values of a metric and perceived quality. This relationship should consider non-linearities and saturation effects of the human visual system.

Previous studies [1], [2], which used standard or high-definition content, have shown that PSNR is strongly content dependent and can only be used as measure of visual quality for each content separately. Sheikh *et al.* [3] evaluated several objective metrics on standard definition images and demonstrated the superior performance of VIF compared to other metrics, including PSNR, which showed the worst performance. Pedersen and Hardeberg [4] evaluated different objective metrics on many available image databases and showed that the performance of the metrics highly depends on the type of content and the types of distortions applied to degrade visual quality. Most of the previous work used standard and high definition images and video sequences in their evaluations, and none of the studies evaluated metrics performance on UHD content.

In our previous study [5], we performed a subjective quality evaluation to benchmark the performance of the upcoming H.265/HEVC video compression standard on 4K UHD content. Three original 4K UHD video sequences were considered: two with natural content and one with synthetic content. The contents were compressed with H.264/AVC and H.265/HEVC at five different bit rates, resulting in a total of thirty compressed video sequences. In this paper, we analyze the performance of several common objective quality metrics (PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM) on 4K UHD content, using the video sequences and corresponding ground truth subjective scores obtained in [5]. For each metric, objective scores were fitted to subjective scores using linear, cubic, and logistic fitting. As compliant with the standard procedure for evaluating the performance of objective metrics [6], [7], the following properties of the estimation of mean opinion scores (MOS) were considered in this study: accuracy, monotonicity, and consistency. Several performance indexes, such as Pearson and Spearman correlation coefficients, root-mean-square-error, and outlier ratio, were computed to compare the metrics estimation of MOS. Statistical tests were performed to determine if the difference between two metrics is statistically significant.

The rest of the paper is organized as follows. The dataset and corresponding subjective scores used as ground truth are described in Section II. The different metrics benchmarked in this study are defined in Section III. In Section IV, the methodology used to evaluate the performance of the metrics is described. Results are presented and analyzed in Section V. Finally, concluding remarks are given in Section VI.

(a) *PeopleOnStreet*



(b) *Traffic*

(c) *Sintel2*

(d) *Sintel39*

Figure 1: Sample frames of the individual contents considered in the subjective test.

II. Dataset and Subjective Scores

The dataset was composed of four ultra-high definition video contents, one for the training (referred to as *Sintel39*) and three for the test (referred to as *PeopleOnStreet*, *Traffic*, and *Sintel2*), with different visual characteristics, resolutions, and frame rates. All contents were five seconds long. The first frame of each content is shown in Figure 1. All test sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bits per sample. The video sequences were compressed with H.264/AVC and H.265/HEVC using the *Random Access* configuration. For each content and codec, five different bit rates were selected.

The evaluation was performed using a 56-inch professional high-performance 4K/QFHD LCD reference monitor Sony Trimaster SRM-L560. Thirty-six naive viewers evaluated the quality of each test sequence. The subjects were seated in three different positions (*Left*, *Centre*, and *Right*) with respect to the center of the monitor, at a distance approximately equal to 3.5 times the height of the screen. The laboratory setup had controlled lighting system to produce reliable and repeatable results. All subjects taking part in the evaluations underwent a screening to examine their visual acuity and color vision.

The Double Stimulus Impairment Scale (DSIS) methodology [8] was chosen as this methodology was selected by VCEQ and MPEG to evaluate the responses to the Joint Call for Proposals on Video Compression Technology [9]. Since the test sequences were only five seconds long and subjects are not used to watch UHDTV, Variant II was selected. A continuous scale ranging from 0 to 100, associated with five distinct impairment categories (*Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible but not annoying*, and *Imperceptible*) was used.

Before the start of the tests, oral instructions were provided to subjects explaining their task. Additionally, a training session was organized to allow subjects to familiarize with the assessment procedure. The video sequences used as training samples had quality levels representative of the labels reported on the rating scales and the experimenter explained the meaning of each label reported on the scale and related them to the presented sample sequences.

The overall experiment was split into two sessions. Between the sessions, each subject took a 15 minutes break before starting the next session. Each session included test materials corresponding to all contents, all the codecs under analysis, and only a subset of the bit rates, which were uniformly distributed across all the sessions. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each group of subjects, while the same content was never shown consecutively.

The subjective results were processed by first detecting and removing subjects whose scores appeared to deviate strongly from others in each test session. Then, the mean opinion score was computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% confidence interval, assuming a Student's t-distribution of the scores. More details about dataset, subjective evaluations, and computation of ground truth MOS can be found in [5].

III. Objective quality metrics

In this study, the performance of the following objective metrics was assessed:

1) PSNR: Peak Signal-to-Noise Ratio,
2) VSNR: Visual Signal-to-Noise Ratio [10],
3) SSIM: Structural Similarity Index [11],
4) MS-SSIM: Multi-Scale Structural Similarity Index [12],
5) VIF: Visual Information Fidelity[1] [13],
6) VQM: Video Quality Metric[2] [14].

All above objective metrics, except for VQM, were computed on the luma component of each frame and the resulting values were averaged across the frames to produce a global index for the entire video sequence.

Most of the objective metrics, except for VSNR, and VQM, were computed using our Video Quality Measurement Tool [15]. VSNR was obtained from its developer website [16]. VQM was obtained from the Institute for Telecommunication Sciences (ITS) website [17].

IV. Performance indexes

The results of the subjective tests can be used as ground truth to evaluate how well the objective metrics estimate perceived quality. The result of execution of a particular objective metric is a video quality rating (VQR), which is expected to be the estimation of the MOS corresponding to a pair of video data. To be compliant with the standard procedure for evaluating the performance of objective metrics [6], [7], the following properties of the VQR estimation of MOS were considered in this study: accuracy, monotonicity, and consistency.

First, a regression was fitted to each [VQR, MOS] data set using linear fitting (1), cubic fitting (2), and logistic fitting (3), with the constraint that the function is monotonic on the interval of observed quality values:

$$MOS_p(VQR) = a \cdot VQR + b \tag{1}$$

$$MOS_p(VQR) = a \cdot VQR^3 + b \cdot VQR^2 + c \cdot VQR + d \tag{2}$$

$$MOS_p(VQR) = a + \frac{b}{1 + \exp\left[-c\left(VQR - d\right)\right]} \tag{3}$$

where a, b, c, and d are the parameters of the fitting functions.

Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) were computed between MOS_p and MOS to estimate the accuracy of the VQR. To estimate monotonicity and consistency, the Spearman rank order correlation coefficient (SROCC) and the outlier ratio (OR), were computed between MOS_p and MOS, respectively. To determine whether the difference between different metrics is statistically significant, statistical tests were performed on these estimators, as described in the following subsections.

[1]Pixel domain version.

[2]NTIA General Model, no calibration.

A. Pearson correlation coefficient

The Pearson linear correlation coefficient (PCC) was computed between MOS_p and MOS to estimate the accuracy of the VQR

$$PCC = \frac{\sum_{i=1}^{M}\left(X_i - \overline{X_i}\right)\left(Y_i - \overline{Y_i}\right)}{\sqrt{\sum_{i=1}^{M}\left(X_i - \overline{X_i}\right)^2}\sqrt{\sum_{i=1}^{M}\left(Y_i - \overline{Y_i}\right)^2}} \tag{4}$$

where X_i and Y_i denote the ground truth subjective score (MOS) and predicted subjective score (MOS_p), respectively, and M is the total number of points.

Based on the assumption that MOS and MOS_p follow a bivariate normal distribution, the Fisher transformation of the Pearson correlation coefficient, $F(PCC)$, approximately follows a normal distribution with mean

$$z = F(PCC) = \frac{1}{2}\ln\frac{1 + PCC}{1 - PCC} \tag{5}$$

and standard deviation

$$\sigma_z = \sqrt{\frac{1}{M - 3}} \tag{6}$$

To determine whether the difference between two PCC values corresponding to two different metrics is statistically significant, a two-sample statistical test was performed. The null hypothesis under test was that there is no significant difference between correlation coefficients, against the alternative hypothesis that the difference is significant, although not specifying better or worse

$$\begin{aligned} H_0&: \quad PCC_1 = PCC_2 \\ H_1&: \quad PCC_1 \neq PCC_2 \end{aligned}$$

The observed value z_{obs} was computed from the observations for each comparison

$$z_{obs} = \frac{z_1 - z_2 - \mu_{z_1 - z_2}}{\sigma_{z_1 - z_2}} \tag{7}$$

where

$$\mu_{z_1 - z_2} = 0 \tag{8}$$

due to the null hypothesis and

$$\sigma_{z_1 - z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2} \tag{9}$$

If the observed value z_{obs} was inside the critical region determined by the 95% two-tailed z-value, then the null hypothesis was rejected at a 5% significance level.

If the sample size M was lower than 30 samples, then the z-value was replaced by a t-value corresponding to a two-tailed Student's t-distribution with $M - 1$ degrees of freedom.

B. Spearman rank order correlation coefficient

The Spearman rank order correlation coefficient (SROCC) was computed between MOS_p and MOS to estimate the monotonicity of the VQR

$$SROCC = \frac{\sum_{i=1}^{M} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{M} (x_i - \overline{x_i})^2} \sqrt{\sum_{i=1}^{M} (y_i - \overline{y_i})^2}} \tag{10}$$

where x_i and x_i denote the ranked variables of the ground truth subjective score (MOS) and predicted subjective score (MOS_p), respectively, and M is the total number of points.

To determine whether the difference between two SROCC values corresponding to two different metrics is statistically significant, a two-sample statistical test was performed similar to Section IV-A.

C. Root mean square error

The root-mean-square error (RMSE) was also computed between MOS_p and MOS to estimate the accuracy of the VQR

$$RMSE = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (MOS_i - MOS_{pi})^2} \tag{11}$$

where M is the total number of points.

Based on the assumption that MOS and MOS_p follow a normal distribution, the root mean square error follows approximately a chi-squared distribution with $M - d$ degrees of freedom, where d is the degrees of freedom of the fitting function.

To determine whether the difference between two PCC values corresponding to two different metrics is statistically significant, a two-sample statistical test was performed. The null hypothesis under test was that there is no difference between RMSE values, against the alternative hypothesis that the difference is significant, although not specifying better or worse

$$\begin{aligned} H_0&: \quad RMSE_1 = RMSE_2 \\ H_1&: \quad RMSE_1 \neq RMSE_2 \end{aligned}$$

The statistic defined in Equation (12) follows a F-distribution with M_1 and M_2 degrees of freedom

$$F_{obs} = \frac{RMSE_1^2}{RMSE_2^2} \tag{12}$$

The observed value F_{obs} was computed from the observations for each comparison. If the observed value F_{obs} was inside the critical region determined by the 95% two-tailed F-value with $M_1 - d$ and $M_2 - d$ degrees of freedom, then the null hypothesis was rejected at a 5% significance level.

D. Outlier ratio

The outlier ratio (OR) was computed between MOS_p and MOS to estimate the consistency of the VQR

$$OR = \frac{\text{total number of outliers}}{M} \tag{13}$$

where M is the total number of points and an outlier is defined as a point for which the error exceeds the 95% confidence interval of the mean MOS value

$$|MOS_i - MOS_{pi}| > CI_i \tag{14}$$

The outlier ratio follows a binomial distribution with mean

$$p = OR \tag{15}$$

and standard deviation

$$\sigma_p = \sqrt{\frac{p(1-p)}{M}} \tag{16}$$

To determine whether the difference between two OR values corresponding to two different metrics is statistically significant, a two-sample statistical test was performed. The null hypothesis under test was that there is no significant difference between outlier ratios, against the alternative hypothesis that the difference is significant, although not specifying better or worse

$$\begin{aligned} H_0&: \quad OR_1 = OR_2 \\ H_1&: \quad OR_1 \neq OR_2 \end{aligned}$$

If the sample size is large ($M \geq 30$), the distribution of differences of proportions from two binomially distributed populations can be approximated by a normal distribution.

The observed value z_{obs} was computed from the observations for each comparison

$$z_{obs} = \frac{p_1 - p_2 - \mu_{p_1 - p_2}}{\sigma_{p_1 - p_2}} \tag{17}$$

where

$$\mu_{p_1 - p_2} = 0 \tag{18}$$

and

$$\begin{aligned} \sigma_{p_1 - p_2} &= \sqrt{p(1-p)\frac{2}{M}} \\ p &= \frac{p_1 + p_2}{2} \end{aligned} \tag{19}$$

because the null hypothesis in this case considers that there is no difference between the population parameters p_1 and p_2.

If the observed value z_{obs} was inside the critical region determined by the 95% two-tailed z-value, then the null hypothesis was rejected at a 5% significance level.

If the sample size M was lower than 30 samples, then the z-value was replaced by a t-value corresponding to a two-tailed Student's *t*-distribution with $M - 1$ degrees of freedom.

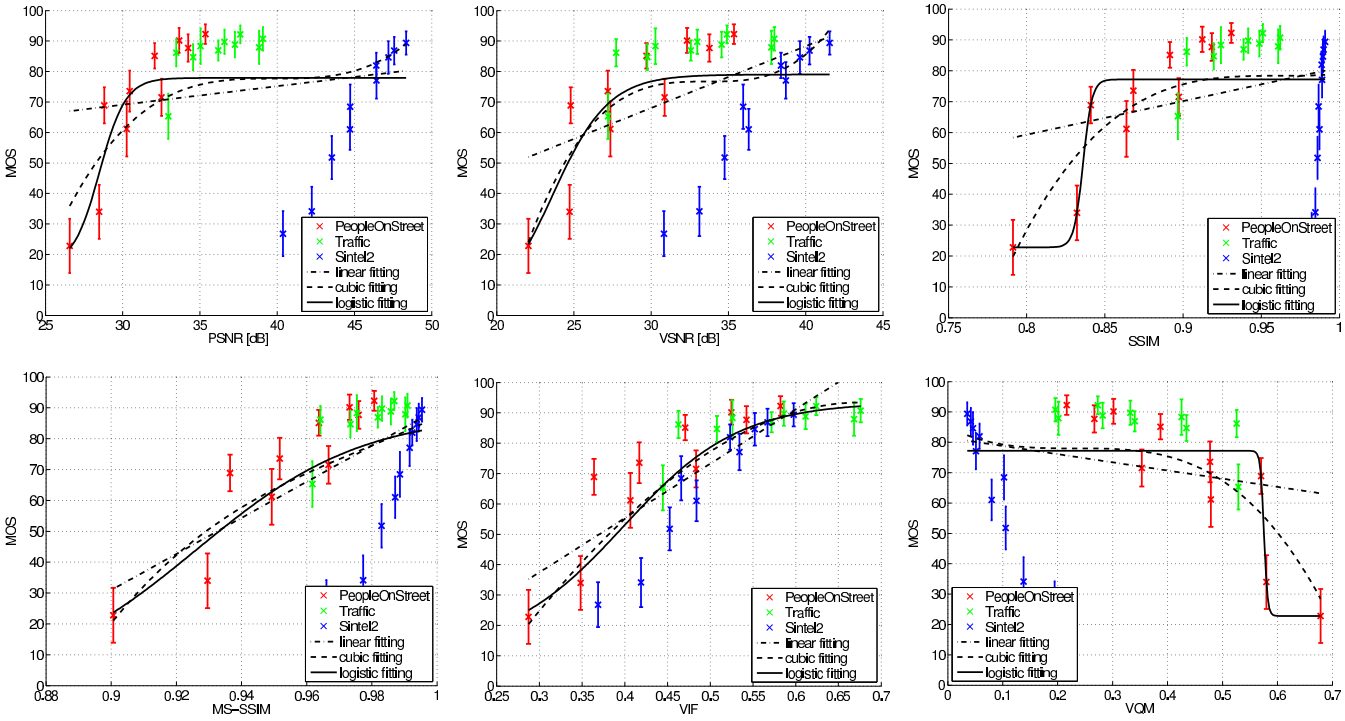


Figure 2: Subjective versus objective results.

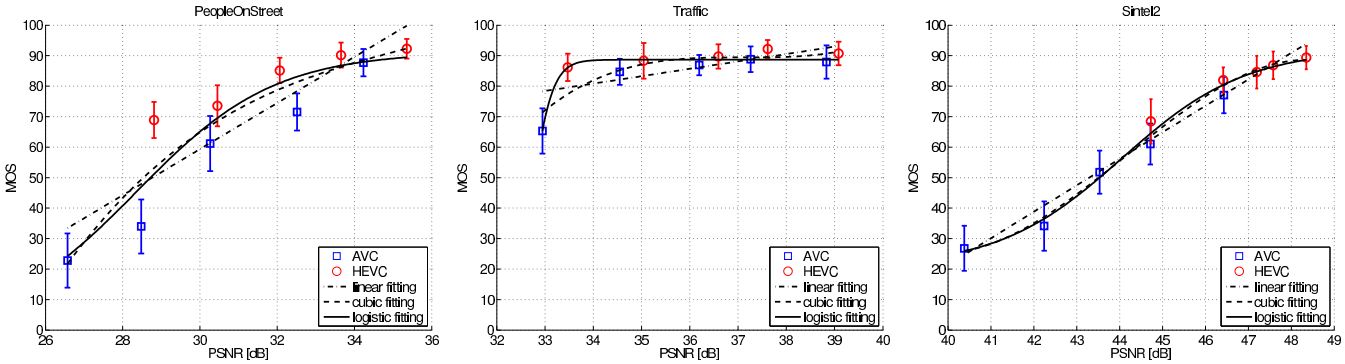


Figure 3: MOS versus PSNR.

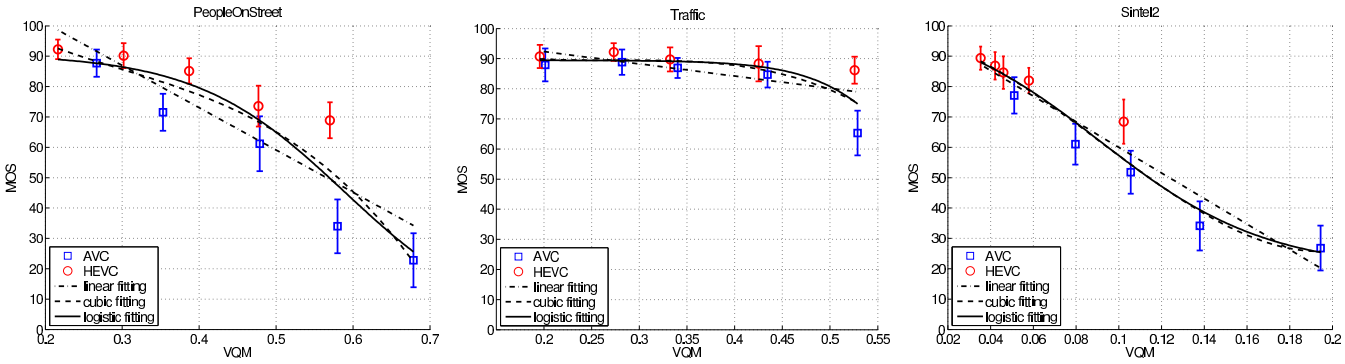


Figure 4: MOS versus VQM.

V. Results

Scatter plots of subjective versus objective results are presented in Figure 2 for the different metrics considered in this study. For all metrics, except for VIF, two well-defined clusters can be observed. One cluster is formed with all data points corresponding to the natural contents (*PeopleOnStreet* and *Traffic*), whereas another cluster is formed with all data points corresponding to the synthetic content (*Sintel2*). The amount of image noise is significantly different between natural and computer-generated contents, which can explain this clear separation. It seems that these objective metrics evaluated synthetic content as having higher overall quality compared to the natural scenes. This finding implies that objective perception depends (for several common metrics) on whether the content is naturally shot or computer-generated, which is not the case with the subjective perception. Although this finding is interesting and was not previously reported for standard or high definition video, the size of the evaluated dataset is too small (only one synthetic video and two natural scenes) to draw any definitive conclusion. This clustering effect is also due to the strong content dependency of these metrics, as reported in previous studies. However, VIF seems to be less content dependent as the data points are more packed into a single cluster.

The linear, cubic, and logistic fittings, as defined in Section IV, were applied in two different ways:

a) on all contents at once,
b) on each content separately.

In the latter case, the performance indexes were computed separately on each content and then averaged across contents. The fitting functions resulting from the different fittings applied on all contents at once are shown in Figure 2. Based on these graphs, it is expected that content dependent metrics will show lower performance in terms of the accuracy, consistency, and monotonicity indexes when compared to VIF. It is also expected that mapping objective scores to subjective scores using a linear fitting, which does not consider non-linearities and saturation effect of the human visual system, will exhibit lower correlation with ground truth subjective scores when compared to using a logistic fitting. However, for MS-SSIM and VIF, the relation between objective scores and MOS considering all data points seems to be more linear. Therefore, the PCC values should be similar between the different fittings for these two metrics. As it can be observed, the fitted cubic and logistic functions are quite similar on each metric. Therefore, the performance indexes obtained for these two fittings should be also quite similar.

The fitting functions resulting from the different fittings applied on each content separately are shown in Figure 3 and Figure 4 for PSNR and VQM, respectively. Because of the limited space, only these results are given, but they illustrate general trends. As it can be observed, mapping objective scores to subjective scores for each content independently significantly increased the correlation with ground truth subjective scores. In this case, the content dependency had no influence as the data points of only one content were used in the fitting process. Therefore, it is expected that the performance indexes will be good for all fittings and metrics when the fitting is applied on each content separately.

The accuracy, consistency, and monotonicity indexes, as defined in Section IV, are reported in Table I (a), Table II (a), and Table III (a) for the linear, cubic, and logistic fittings, respectively. The results reported in this tables confirm the analysis performed based on Figure 2 and Figure 3. For all metrics, except VIF, a clear improvement can be observed in terms of PCC, SROCC, RMSE, and OR when the fitting was applied on each content separately rather than on all contents at once. In this case, the PCC and SROCC were always higher than 0.82 and 0.89, respectively. In the case of VIF, the PCC and SROCC values were over 0.82 and 0.86, respectively, when the fitting was applied on all contents at once. These results confirm previous results obtained for standard and high definition video showing that PSNR, VSNR, SSIM, MS-SSIM, and VQM are highly content dependent, whereas VIF is less content dependent.

When the fitting was applied on each content separately, the obtained performance indexes were roughly similar between cubic and logistic fittings, and slightly better when compared to linear fitting. On the other hand, when the fitting was applied on all contents at once, the performance indexes obtained with cubic and logistic fittings were significantly better when compared to linear fitting. These findings are also consistent with evaluations of standard and high definition video. As predicted based on the graphical analysis, the PCC value of MS-SSIM and VIF were quite similar across the different fittings. Theoretically, if the fitting functions are strictly monotonic on the interval of observed quality values, then the SROCC value of one particular metric should be the same across fittings. However, because of the numerical precision of floating-point numbers, the logistic function can become monotonic instead of strictly monotonic, especially on the horizontal asymptotes. This phenomenon can be observed for SSIM and VQM when the fitting was applied on all contents at once, for example.

All metrics seem to have good and similar performance in terms of PCC, SROCC, RMSE, and OR when the fitting was applied on each content separately. However, when the fitting was applied on all contents at once, VIF seems to outperform other metrics. To determine if the difference between VIF and the other metrics is significant, statistical tests were performed according to Section IV. The results of the statistical tests are reported in Table I (b), Table II (b), and Table III (b) for the linear, cubic, and logistic fittings, respectively. Each entry in the table corresponds to the results of the statistical tests performed on the following performance indexes (from left to right): PCC, SROCC, RMSE, and OR. The statistical tests were performed to determine whether the difference between two performance index values corresponding to two different metrics was statistically significant. In these results, '$=$' means that there was no significant difference between the two metrics, whereas '$\neq$' means that the difference was significant.

Table I: Linear fitting.

(a) Accuracy, consistency, and monotonicity indexes.

	All contents				Average			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
PSNR	0.1862	0.1818	20.3207	0.7333	0.8521	0.9111	6.6611	0.3667
VSNR	0.4929	0.4131	17.9960	0.7333	0.8253	0.8949	7.5384	0.5000
SSIM	0.2858	0.1818	19.8200	0.7667	0.8706	0.9071	6.4677	0.4000
MS-SSIM	0.6158	0.4429	16.2963	0.7000	0.8787	0.9152	6.3531	0.2667
VIF	0.8252	0.8661	11.6833	0.6667	0.8473	0.9071	7.3221	0.4333
VQM	0.2444	0.1804	20.0552	0.7333	0.8306	0.9071	7.7061	0.4333

(b) Statistical analysis.

	PSNR	VSNR	SSIM	MS-SSIM	VIF	VQM
PSNR		= = = =	= = = =	= = = =	≠ ≠ ≠ =	= = = =
VSNR	= = = =		= = = =	= = = =	≠ ≠ ≠ =	= = = =
SSIM	= = = =	= = = =		= = = =	≠ ≠ ≠ =	= = = =
MS-SSIM	= = = =	= = = =	= = = =		= ≠ = =	= = = =
VIF	≠ ≠ ≠ =	≠ ≠ ≠ =	≠ ≠ ≠ =	= ≠ = =		≠ ≠ ≠ =
VQM	= = = =	= = = =	= = = =	= = = =	≠ ≠ ≠ =	

Table II: Cubic fitting.

(a) Accuracy, consistency, and monotonicity indexes.

	All contents				Average			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
PSNR	0.5328	0.1818	17.5020	0.7333	0.9273	0.9111	5.0571	0.1667
VSNR	0.5965	0.4131	16.5995	0.7000	0.9023	0.8949	5.9134	0.2667
SSIM	0.5768	0.1818	16.8960	0.8000	0.9248	0.9071	5.3742	0.2333
MS-SSIM	0.6284	0.4429	16.0885	0.7333	0.9291	0.9152	4.5779	0.2000
VIF	0.8689	0.8661	10.2365	0.4000	0.9212	0.9071	5.6887	0.2333
VQM	0.5365	0.1804	17.4542	0.7333	0.8797	0.9071	6.5988	0.3000

(b) Statistical analysis.

	PSNR	VSNR	SSIM	MS-SSIM	VIF	VQM
PSNR		= = = =	= = = =	= = = =	≠ ≠ ≠ ≠	= = = =
VSNR	= = = =		= = = =	= = = =	≠ ≠ ≠ ≠	= = = =
SSIM	= = = =	= = = =		= = = =	≠ ≠ ≠ ≠	= = = =
MS-SSIM	= = = =	= = = =	= = = =		≠ ≠ ≠ ≠	= = = =
VIF	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠		≠ ≠ ≠ ≠
VQM	= = = =	= = = =	= = = =	= = = =	≠ ≠ ≠ ≠	

Table III: Logistic fitting.

(a) Accuracy, consistency, and monotonicity indexes.

	All contents				Average			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
PSNR	0.5869	0.1818	16.7451	0.8333	0.9616	0.9111	4.4415	0.1333
VSNR	0.5792	0.4131	16.8598	0.7667	0.9533	0.9014	5.0867	0.2000
SSIM	0.6038	0.2761	16.4862	0.8000	0.9620	0.8998	4.6772	0.1667
MS-SSIM	0.6264	0.4429	16.1232	0.7333	0.9672	0.9079	3.8316	0.1667
VIF	0.8708	0.8661	10.1694	0.4000	0.9596	0.9071	4.9557	0.1667
VQM	0.6038	0.4750	16.4864	0.8000	0.8792	0.9071	6.7072	0.3000

(b) Statistical analysis.

	PSNR	VSNR	SSIM	MS-SSIM	VIF	VQM
PSNR		= = = =	= = = =	= = = =	≠ ≠ ≠ ≠	= = = =
VSNR	= = = =		= = = =	= = = =	≠ ≠ ≠ ≠	= = = =
SSIM	= = = =	= = = =		= = = =	≠ ≠ ≠ ≠	= = = =
MS-SSIM	= = = =	= = = =	= = = =		≠ ≠ ≠ ≠	= = = =
VIF	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠	≠ ≠ ≠ ≠		≠ ≠ ≠ ≠
VQM	= = = =	= = = =	= = = =	= = = =	≠ ≠ ≠ ≠	

When the fitting was applied on each content separately, the performance indexes were computed on only ten data points. To avoid violating some of the assumptions made in Section IV, the statistical tests were not performed in this case. Therefore, the statistical tests were performed only on the performance indexes computed when the fitting was applied on all contents at once. When linear fitting was applied, results of the statistical tests showed that the SROCC values of VIF and MS-SSIM were significantly different whereas there was no difference on the PCC, RMSE, and OR values. The PCC, SROCC, and RMSE values of VIF were significantly different to those of PSNR, VSN, SSIM, and VQM whereas there was no difference on the OR values. In all other cases, there was no significant difference between the PCC, SROCC, RMSE, or OR values. When cubic or logistic fitting was applied, results of the statistical tests showed that there was a significant difference between the PCC, SROCC, RMSE, and OR values of VIF and the other metrics. There was no significant difference between the PCC, SROCC, RMSE, or OR values of the other metrics.

VI. Conclusion

In this paper, the performance of several objective quality metrics was evaluated on three different 4K UHD video sequences. To evaluate the metrics performance, mean opinion scores collected during a formal subjective evaluation were used as ground truth. Results showed that metrics are content dependent except for VIF, which is consistent with previous findings for standard and high definition image and video content. Applying a logistic fitting increases the performance when compared to linear and cubic fitting. An interesting finding is that the majority of metrics showed a shift in objective values between synthetic and natural contents, with the exception being VIF. However, the number of ultra-high definition contents used in the dataset is not large enough to draw general conclusions about the content dependency of these metrics.

Acknowledgment

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), EC funded Network of Excellence VideoSense, and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

References

[1] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2012, pp. 37–38.

[2] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.

[3] H. Sheikh, M. Sabir, and A. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, November 2006.

[4] M. Pedersen and J. Y. Hardeberg, "Full-reference image quality metrics: Classification and evaluation," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 1, pp. 1–80, 2012.

[5] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in *Proceedings of SPIE*, ser. Applications of Digital Image Processing XXXV, vol. 8499, 2012.

[6] ITU-T Tutorial, "Objective perceptual assessment of video quality: Full reference television," International Telecommunication Union, 2004.

[7] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.

[8] ITU-R BT.500, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, January 2012.

[9] JCT-VC, "Joint Call for Proposals on Video Compression Technology," ITU-T SG16 Q.6 (VCEG) and ISO/IEC JTC1/SC29/WG11 (MPEG), Kyoto, JP, Tech. Rep. VCEG-AM91, MPEG-N11113, January 2010.

[10] D. Chandler and S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, September 2007.

[11] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[12] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1398–1402.

[13] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, February 2006.

[14] ITU-T J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," International Telecommunication Union, March 2004.

[15] VQMT. [Online]. Available: http://mmspg.epfl.ch/vqmt/

[16] VSNR. [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[17] VQM. [Online]. Available: http://vqm.its.bldrdoc.gov/