IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 13, 2021, accepted November 7, 2021, date of publication November 10, 2021,
date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127395

On the Link Between Subjective Score Prediction and Disagreement of Video Quality Metrics

LOHIC FOTIO TIOTSOP[1], (Member, IEEE), FLORENCE AGBOMA[2], GLENN VAN WALLENDAEL[3], (Member, IEEE), AHMED ALDAHDOOH[4], SEBASTIAN BOSSE[5], (Member, IEEE), LUCJAN JANOWSKI[6], MARCUS BARKOWSKY[7], (Member, IEEE), AND ENRICO MASALA[1], (Senior Member, IEEE)

[1]Control and Computer Engineering Department, Politecnico di Torino, 10129 Torino, Italy
[2]Global OTT Platforms, Sky U.K., Isleworth TW7 5QD, U.K.
[3]Department of Electronics and Information Systems, Ghent University -imec - IDLab, 35000 Ghent, Belgium
[4]IETR—UMR 6164, CNRS, INSA Rennes, University of Rennes, 35000 Rennes, France
[5]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, 10587 Berlin, Germany
[6]Department of Telecommunication, AGH University of Science and Technology, 30-059 Krakow, Poland
[7]Technische Hochschule Deggendorf, Deggendorf Institute of Technology, 94469 Deggendorf, Germany

Corresponding authors: Lohic Fotio Tiotsop (lohic.fotiotiosop@polito.it) and Enrico Masala (enrico.masala@polito.it)

This work was supported in part by Sky Group and Affiliates, and in part by the PoliTO Interdepartmental Centre for Service Robotics (PIC4SeR) (http://pic4ser.polito.it).

ABSTRACT Several video quality metrics (VQMs) have been proposed in many publications to predict how humans perceive video quality. It is common to observe significant disagreements amongst the quality predictions of these VQMs for the same video sequence. Following an extensive literature search, we found no publicised work that has investigated if such disagreements convey useful information on the accuracy of VQMs. Herein, a measure for quantifying the disagreement between VQMs is proposed. A small-scale subjective study is carried out to assess the effectiveness of our proposal. In particular, the proposed disagreement measure is shown to be extremely effective in determining whether the quality of any given processed video sequence (PVS) can be accurately predicted by the VQMs. This type of information is particularly useful for identifying video sequences that are likely to degrade the end-user's quality of experience (QoE). Our proposal is also useful in selecting the most effective PVSs to be employed in a subjective test. We show that the proposed disagreement measure can be effectively predicted from bitstream features. This establishes a link between the capability to accurately assess the quality of a PVS and the way it is encoded. In addition, an analysis is conducted to compare the performances of some well-known and widely used open-source metrics and two proprietary metrics. The two proprietary metrics are used by a large media company for enhancing its delivery pipeline. The outcome of this comparison highlights the suitability of the open-source VQM, Video Multi-method Assessment Fusion (VMAF), as a good benchmark quality measure for both the industrial and academic environments.

INDEX TERMS Objective measures, proprietary metrics, subjective test, video quality, metrics disagreement.

I. INTRODUCTION

A major concern for content providers and content aggregators is to guarantee high quality of experience (QoE) to their customers. The last decades have therefore witnessed numerous publications that have proposed novel algorithms to generate video quality metrics (VQMs) that can predict a mean opinion score (MOS [1], [2]. The MOS is the average of the opinion scores of end users when they are asked to rate or score their perception of the video quality during a subjective experiment. Quite often, significant differences occur between the MOS values predicted by these different VQMs, for the same processed video sequence (PVS). The study reported in this paper was carried out because, after an extensive literature search, no published works were found that investigated whether any useful information is obtainable about the accuracy of objective metrics from the differences and disagreements between the MOS predictions of the VQMs.

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

As a starting point in this study, a measure is proposed for quantifying the disagreements between VQMs. As a convention, its values range from 0 to 1 and can be computed for any PVS. The closer its value is to 1, the more the VQMs disagree on the perceptual quality (or MOS) of the PVS. Our study shows that the proposed measure is particularly useful in identifying i) PVSs for which the commonly used open-source VQMs and some proprietary VQMs are likely to deliver quality predictions that vary greatly from what the end user perceives, ii) PVSs for which the VQMs are likely to produce quality predictions that are close to the end-user scores. The proposed measure has the potential of being very useful in academia and industry, since it can determine if predictions made by VQMs are accurate or not.

In academia, this measure will facilitate the creation of effective tooling to identify appropriate subsets of PVSs to be used in subjective tests. This measure is useful for two additional reasons. Firstly, it saves time and resources by excluding from subjective experiments, PVSs whose end-user scores are accurately predictable using VQMs alone. Secondly, it can be used in identifying problematic PVSs for which VQMs are poor at predicting the end-user scores. Results from experiments using such PVSs are typically of great value to researchers.

In the media industry, it is of primary importance to be able to quickly and automatically identify the PVSs on which the quality predictions provided by the VQMs could be misleading. Misleading quality predictions often result in unexpected degradation of customers' QoE through inadequate resource provisioning. The results presented in this paper are the outcomes of a collaborative work with a global media company. Following kick-off consultations with that company, the scope of the collaboration was divided into three parts: i) automatic identification of PVSs for which VQMs are likely to produce inaccurate MOS estimation ii) determination of PVSs attributes, such as compressed bitstream features, which could affect the ability of a VQM to accurately predict the perceptual quality iii) benchmarking the performance of two proprietary VQMs (PVQMs) used internally by the company against well-known and widely used open-source metrics. Subjective experiments were conducted using an appropriate subset of PVSs carefully selected from a large dataset specifically created for this work. Results from these experiments show the effectiveness of the proposed measure in addressing the first and the second subject areas above. The analyses also produced conclusive results for performance comparisons of the two PVQMs against the widely used VQMs.

The key contributions of this paper are as follows:

1) This is the first publication proposing a measure that quantifies the disagreements between VQMs when predicting the perceptual quality or MOS of a given PVS. The work in this paper shows that the proposed measure is useful in automatically determining whether the quality scores predicted by VQMs are reliable or not. The proposed measure can thus be used to provide a preliminary answer to the following research question – "Which PVSs should be used in a subjective experiment to get the most out of it?"
2) This paper shows that for a given PVS, the proposed VQM disagreement measure can be estimated from the bitstream features of that PVS. This suggests there is possibly a relationship between the way a PVS has been encoded and how accurately a VQM can predict the perceptual quality of that PVS. In fact, bitstream features strongly depend on encoding settings and the proposed disagreement measure determines the difficulty of accurately assessing the quality of a PVS using a VQM.
3) This paper shows a comparison between two proprietary metrics and some well-known and widely used open-source VQMs. The proprietary metrics were highly optimised to operate in the real-world environment and are used in the content delivery pipeline of the global media company.

To perform the experiments, a dataset comprising 368 industry grade PVSs was created. Industry-grade (mezzanine format) content is minimally compressed during data acquisition [3]. This dataset differs from other widely used video quality datasets, which are typically built by using pristine-quality content and acquired without any compression. In media industries, content is usually of the mezzanine format, which is of high quality but not pristine. A decision was made to work with industry grade content to closely replicate the conditions encountered in actual media industry processing chains.

This study considered the following VQMs namely: Peak Signal to Noise Ratio (PSNR) [4], Structural Similarity Index Measure (SSIM) [5], Multi-Scale Structural Similarity Index Measure (MSSSIM) [6], Visual Information Fidelity (VIF) [7], Extended Weighted Peak Signal-to-Noise Ratio (XPSNR) [8], Video Multi-method Assessment Fusion (VMAF) [9], PVQM1 (the first proprietary metric), and PVQM2 (the second proprietary metric). Due to corporate legal considerations, the full names of the two proprietary metrics have been omitted above.

There are newer open-source VQMs than the ones listed above, some of which are presented in the ITU recommendation P.1203 [10], and others are based on Deep Learning approaches (a branch of Machine Learning). The academic and industry communities have not yet adopted these metrics on a large scale since many of them have not yet been tested in real-world environments. As such, the focus of this study was not on these more recent VQMs. Unlike the newer open-source metrics, the metrics considered for our study are those typically used by academic researchers for designing and evaluating state-of-the-art video processing applications [11]–[15]. Therefore, a measure, as the one proposed in this work, that provides information on the accuracy of these metrics, is of large interest for the scientific and industrial community.

To obtain the values for the proposed disagreement measure, we mapped all the VQMs onto the same scale. For each PVS, we counted the number of unique VQM pairs from the collection of possible VQM pairs, where one VQM provided a quality prediction that was perceptually different from the other VQM of the pair. We argue that this number, expressed as a fraction, is an effective indicator of the accuracy of the VQMs. In other words, if many VQMs disagree on the perceptual quality of a given PVS, then each VQM is also likely to wrongly estimate the MOS of that PVS. We are aware of the existence of some standardised techniques of comparing VQMs [16]. However, the work presented in this paper was aimed at investigating the implications of VQMs disagreements rather than directly comparing the metrics.

A support vector regression model was also trained and cross validated. Its accuracy shows that the proposed disagreement measure can be predicted from bitstream features such as the bit rate, the quantisation parameter and the motion vector components. This model has the following two purposes: i) identification of the bitstream features that contribute towards the VQM disagreements and thus the difficulty of objectively estimating the MOS of a PVS ii) the development of an efficient method for identifying, in a large set of PVSs, those for which it is strongly recommended to perform a subjective evaluation test.

To assess the effectiveness of the proposed measure, a small-scale subjective experiment was carried out on a subset of PVSs characterised by both low and high VQMs disagreements. The results showed the effectiveness of the proposed measure in deducing the accuracy of VQMs. A comparison analysis was then performed on all the VQMs relying on both the subjectively evaluated PVSs and the objectively evaluated ones. The results revealed that VMAF performed better when compared to all the other metrics involved in the study. The two PVQMs also showed better performance when compared to other open-source VQMs.

The paper is organised as follows. Section II presents a short review of previous works on the agreements and disagreements within a set of VQMs. Section III provides a description of the dataset used in this study. Section IV details the proposed VQMs disagreement measure. Section V describes the subjective experiment setup. Results are discussed in Section VI. The terms VQMs and metrics are going to be used interchangeably in this paper. Finally, conclusions are drawn in Section VII.

II. RELATED WORK

The idea of leveraging many objective metrics together to deliver more accurate assessments of perceptual quality has been investigated in the literature [17]. It has been shown that a machine learning (ML) model that takes, as input, a set of different VQMs computed on a given PVS, can yield improved quality predictions as opposed to using only single VQM. In [18], the authors designed a support vector regression model that jointly utilised several VQMs to provide a more accurate MOS estimations. The work presented in [19] argued that PVSs whose sources are characterised by a low spatial activity index are challenging to work with from the point of view of objective quality assessment. In that work, a neural network-based model was proposed to address such challenges. The model relied on the scores from many full-reference metrics in addition to the spatial and the temporal activity index to mitigate the inaccuracies of VQMs when estimating the quality of these PVSs. By feeding a ML based model with many different VQMs, the authors aimed at exploiting the diversities and similarities between the VQM scores in order to reach a better MOS estimation.

The approach of studying the differences between the predictions of many VQMs has not been exploited solely for accurate MOS estimations. In fact, in [20] the authors showed that the agreements between different VQMs, as measured by the Spearman and the Kendall rank order correlation coefficients, were related to the standard deviation of subjective ratings for a given PVS. They designed a neural network-based model that takes as input five VQMs and estimates the diversity among users' ratings. Still focusing on the quality scores as predicted by different VQMs, in [21], the authors proposed an approach based on Gaussian mixture models to find the range of quality values to which the MOS of a given PVS is expected to belong with a given probability.

In all the papers mentioned so far, the VQMs were studied together with ML models to enhance some aspects of the quality assessment processes. Despite the useful results reported in all these papers, their use of ML models means that they relied on black box models whose internal workings might not be trivial or easy to understand. Instead of using ML models, some other authors have exploited the information associated with the diversity or similarity between VQM scores in a more intuitive and easier to interpret way. In [22] and [23] the authors investigated the disagreements between PSNR, SSIM and the VIF at the frame and sequence level. In both works the authors analysed the behaviour of the three metrics on a given pair of PVSs. They evaluated, for different source content, the ability of these metrics to coherently rank the perceptual quality of a pair of PVSs.

The work in this paper differs from those in [22] and [23] in that the VQM disagreement measure focused on pairs of VQM metrics instead of PVSs, thus yielding an indicator that determines how difficult it is to assess the quality of a given PVS using VQMs. A small-scale subjective experiment was used in validating this concept, and the results showed that such a simple indicator could provide relevant information regarding the ability to accurately predict the perceptual quality of a PVS without resorting to a subjective experiment.

Furthermore, we observed that the proposed VQM disagreement measure is significantly correlated to the PVS bitstream features, and that, it can be predicted using several of such features. This allowed us to conclude that the way a PVS is encoded may enhance or negatively affect the accuracy of VQMs.

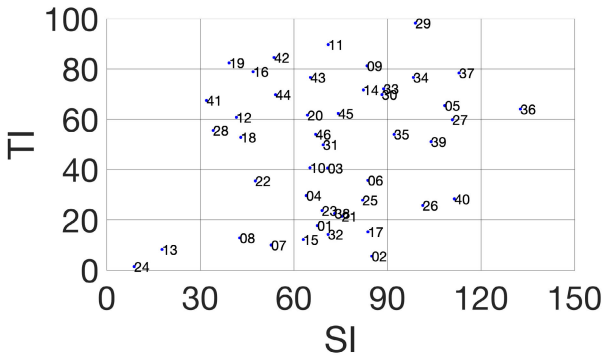


FIGURE 1. Assessing the heterogeneity of the 46 sources used to generate the PVSs contained in the dataset in terms of the spatial and temporal activity index. The labels indicate the different sources.

TABLE 1. Summary of the 8 hypothetical reference circuits (HRC) used on each of the 46 sources to generate the 368 (46 ∗ 8) PVSs in the dataset.

HRC	Resolution	Bit rate (kbps)
HRC1	512 x 288	365
HRC2	768 x 432	730
HRC3	768 x 432	1100
HRC4	960 x 540	2000
HRC5	1280 x 720	3000
HRC6	1280 x 720	4500
HRC7	1920 x 1080	6000
HRC8	1920 x 1080	7200

Another fundamental difference between the work in this paper and many others in the literature is the inclusion of proprietary VQMs. Researchers typically use open-source tools to benchmark their proposals. As such, VQM comparison studies have mostly focused on freely available metrics [24]. However, in some cases, open-source software are not properly optimised for effectively operating in real-world scenarios. To the best of our knowledge, there is just a small number of published works that have conducted VQM comparison studies involving proprietary VQMs [25]. Therefore, this work contributes in shedding light on the existence of a potential gap between the accuracy of well-known and widely used open-source VQMs and proprietary ones.

III. DATASET PREPARATION

A total of 46 Full HD (FHD) industry grade source videos were selected according to guidelines in [26]. These comprised a range of entertainment videos including sports, movies and animations. Depending on which region (Europe or US), the video frame rates per second (fps) were either 23.976, 25.000 or 29.970. Figure 1 shows the selected sources covered a wide range in terms of Spatial Information (SI) and Temporal Information (TI) according to [27].

The source videos were encoded using H.264/AVC constant bit rates. The Apple's HLS authoring specification [28] was used as guidelines in producing the eight hypothetical reference circuits (HRCs) summarised in Table 1. Some of the key encoding configurations included one-pass encoding preset, the instantaneous decoder refresh (IDR) interval was set to two seconds, with an option of inserting an I-frame if there was a scene change within a given IDR interval. The size of the video buffer verifier was set to 5 seconds and the deinterlacing mode was set to motion adaptive interpolation. A summary of the bit rates and resolutions are given in Table 1.

From each of the 46 source videos, eight PVSs were created resulting in a total of 368 PVSs. The PVSs in the dataset were also divided into two main categories, namely movies and sports. For sports content in Europe, the frame rates were interpolated from 25.00 fps to 50.00 fps. For sports content in the US, the frame rates were interpolated from 29.97 fps to 59.94 fps. This was done to reduce judder during playback, caused by camera panning movements. The frame rates for the movie content were untouched, so they were the same as the source videos.

The duration of each video was 10 seconds. But, allowing for an extra two seconds of content before and after the video, results in a total duration of 14 seconds. The purpose of the extra amount of time was to allow the video encoder to stabilise to the requested bit rate, thus removing quality fluctuations that may be present due to the rate control algorithm. Once each source was encoded, the FFMPEG application was used to trim off the extra four seconds of content.

The video quality of the 368 PVSs were evaluated using the eight considered VQMs - PSNR [4], SSIM [5], MSSSIM [6], VIF [7], XPSNR [8], VMAF [9] and the two proprietary VQMs PVQM1 and PVQM2.

The scores of each of these VQMs were recorded in a dataset, resulting in a total of 46 $sources \cdot 8\ HRCs \cdot 8\ VQMs = 2944$ objective quality scores to be analysed.

All eight VQMs considered in this study were full reference metrics, i.e., they evaluate the quality of a distorted signal by comparing it to the source. PSNR measures the quality of the distorted content by deriving its mean square error (MSE) with respect to the source pixels. SSIM evaluates the similarity between the source and the distorted signal by considering three main aspects, namely the luminance, the contrast and the preservation of the structures. MSSSIM implements the same steps as SSIM but at multiple scales. VIF uses natural scene statistics models to define the image information perceived by the human vision system (HVS). It then quantifies the amount of information shared between the source and the distorted signal. XPSNR is an enhancement of PSNR, which uses a *distance* between the source signal and the distorted signal considering some characteristics of the human vision system which are not considered when using the MSE alone. VMAF fuses together multiple elementary full reference metrics using machine learning. The rationale behind VMAF is that each elementary metric may have its own strengths and weaknesses with respect to the characteristics of the source video, the type of artefacts, and the degree of distortion. VMAF seeks to

preserve the strengths of the individual metrics and to deliver a more accurate final score.

PVQM1 is a machine learning based VQM. It was trained using a diverse range of interlaced and progressive video content including sports, TV shows and movies. Currently, it is used by the global media company to set the desired target MOS for content-aware encoding and for video-on-demand solutions. PVQM2 is based on a model of human vision system. The aim is to produce scores which are proximal to how human viewers would judge the perceptual quality. The design scope of PVQM2 includes both interlaced and 1080p TV viewing conditions.

Note that PSNR, SSIM, MSSSIM and VIF were originally developed for assessing the quality of still images. However, due to their analytical properties and low complexity, they are also the most used metrics for monitoring quality when designing video processing applications [11]. PSNR is even considered a kind of baseline in the context of video quality assessment. The Video Quality Experts Group (VQEG) [29] for instance, often uses PSNR as a benchmark for validation experiments, as was done during the performance evaluation of full reference VQMs in the HDTV experiment [30]. Many papers have compared PSNR, SSIM, MSSSIM and VIF to other Video Quality Measures (VQMs) [8], [31], [32]. Therefore, the consideration of these open-source metrics is not peculiar to the work reported here. Our study contributes to shedding light on the existence of a potential statistically significant gap between the accuracy of these widely used open-source VQMs and proprietary ones.

IV. PROPOSED VIDEO QUALITY METRICS DISAGREEMENT MEASURE

One of the major issues addressed in this work was how to objectively identify the PVSs for which VQMs are likely to produce inaccurate MOS estimations. To this end, we propose a measure based on the disagreements between the scores provided by a set of VQMs. Such a measure enables the establishment of whether a VQM would accurately estimate the perceptual quality of a given PVS as shown in Section VI.

Let denote by D_{pvs} the value of the proposed measure of VQMs disagreement for a given PVS. To formally define D_{pvs}, we introduce the following parameters:

- n, the number of VQMs used to evaluate the perceptual quality of the PVS;
- $VQM_1, VQM_2, \ldots, VQM_n$, the n VQMs used to evaluate the quality of the PVS;
- The respective predicted scores of the VQMs vqm_1^{pvs}, $vqm_2^{pvs} \ldots, vqm_n^{pvs}$

In order to compute D_{pvs}, one of the VQMs is chosen as the reference metric. Assume that VQM_1 is the reference metric, let the following functions

- f_i $(i = 1, 2, \ldots, n)$ be for mapping each VQM_i from its original scale to the VQM_1 scale.
- δ_1 denote the VQM_1 sensitivity, which is the minimum variation in quality perceptible by most human viewers if the quality were to be predicted using VQM_1.

For instance, it has been empirically observed that two pictures having VMAF scores that differ by less than seven points are likely to be judged as equal in terms of perceptual quality [33]. Therefore, for VMAF, the δ would be seven. The consideration of the VQM sensitivity is not a peculiarity of this work; similar approaches have already been proposed in the literature [34].

Relying on the previously introduced notation, D_{pvs} is defined as follows:

$$D_{PVS} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}_{\left(\left|f_i\left(vqm_i^{pvs}\right) - f_j\left(vqm_j^{pvs}\right)\right| > \delta_1\right)}}{\binom{n}{2}} \quad (1)$$

where $\mathbb{1}$ is the indicator function, whose value is 1 if the subscript proposition is true and 0 otherwise. The denominator in Eq (1) is the total number of unique pairs of metrics that can be formed using the n VQMs. The numerator counts the number of these pairs for which the two metrics that constitute the pair disagree on the perceptual quality of the PVS. Two metrics are said to disagree when the absolute value of the difference between the predicted scores (using the reference metric scale) is greater than δ_1.

In this work, VMAF was chosen as the reference VQM and δ_1 was set to 7. Furthermore, the mapping functions have been computed by performing a least square fitting of each of the VQMs to VMAF using third-order polynomial functions [16]. The diagram in Figure 2 summarises the implementation steps for the computation of the proposed disagreement measure.

For any PVS, $D_{pvs} \in [0, 1]$. The closer the value of D_{pvs} is to one, the larger the disagreement between the VQMs regarding the perceptual quality of the PVS. We argue that the larger the value of D_{pvs} for a given PVS, the more likely it is that VQMs will be inaccurate when assessing the perceptual quality of that PVS. To verify such a statement, we conducted a subjective experiment whose details are provided in the next section.

V. SMALL SCALE SUBJECTIVE EXPERIMENT

A subjective experiment was conducted to investigate the reliability of the proposed measure. Due to time constraints, the experiment was conducted on a small scale.

A. SELECTION OF THE PROCESSED VIDEO SEQUENCES TO TEST

Since we aimed at investigating the implications of VQM disagreements, viewers were shown PVSs on which the VQMs strongly agreed and those for which the VQMs strongly disagreed.

The VQMs disagreement value D_{pvs}, as described in Section III was computed for each of the 368 PVSs in the dataset. Afterwards, the PVSs were sorted in ascending order of D_{pvs}. From this, the following were found: i) at the lowest scale, 31 PVSs had $D_{pvs} < 0.2$ ii) at the highest scale, 36 PVSs had $D_{pvs} > 0.6$. These PVSs at the lowest and highest scales were selected for the subjective test dataset.

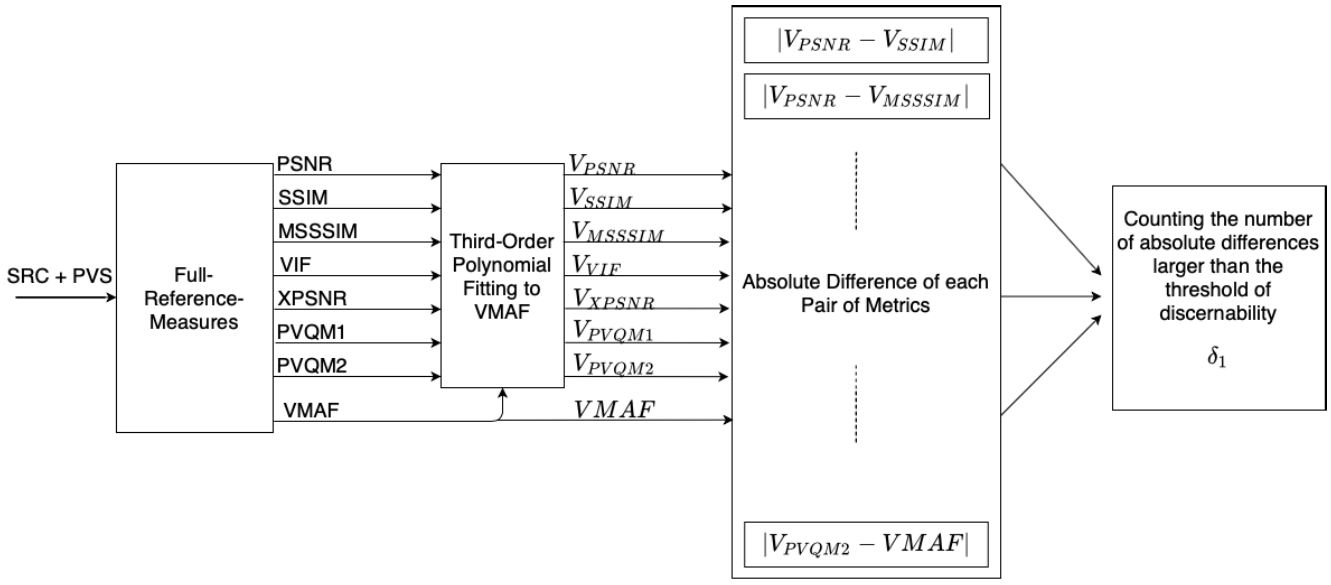


FIGURE 2. The diagram summarises the implementation steps of the proposed disagreement measure. VMAF is chosen as the reference metric, hence, the VQM sensitivity δ_1 is set to 7. V_{PSNR} is the quality score obtained after performing a least square fitting of the PSNR to the VMAF scale using a third-other polynomial function. The same definition holds for all the other VQMs. By considering eight different VQMs, in total, 28 absolute differences were computed that corresponded to the number of unique pairs of VQMs that can be formed by selecting two VQMs from the eight available.

In addition to these 67 PVSs (31 + 36), 16 additional PVSs were added onto the dataset to ensure viewers evaluated a dataset whose perceptual qualities covered the entire quality scale, as this is a good practice in designing subjective experiments.

B. EXPERIMENT SETUP

A total of 16 subjects (viewers) working in the media industry participated in this subjective experiment across two laboratories in Italy and Germany. The subjects were non-experts. The Double Stimulus Impairment Scale (DSIS) method was used. In this method, the subjects are shown both the source video and the PVS. The DSIS method closely follows how most of the full reference metrics operate; that is by computing the perceptual differences between the original reference video and the degraded test video. By adopting the DSIS, we aimed at aligning the subjective evaluation as closely as possible to how full reference metrics operate. This was to mitigate against any extraneous sources of inaccuracies not directly related to the VQMs.

The source video was shown first, followed by the encoded one (PVS) as illustrated in Figure 3. After watching the source video, the PVS was shown two seconds later. The subjects were then given six seconds to rate their perception and the annoyance of artefacts within the PVS against the source video using a 5-grade impairment scale. The scale consisted of the following five options: ''Very annoying'', ''Annoying'', ''Slightly annoying'', ''Perceptible but not annoying'', ''Imperceptible''. To aid in the computation of the MOS values, the five options were assigned unique numeric scores (ratings) from 1 to 5 respectively. For each subject, the viewing distance to the monitor was fixed in accordance with the relevant ITU recommendations [27].

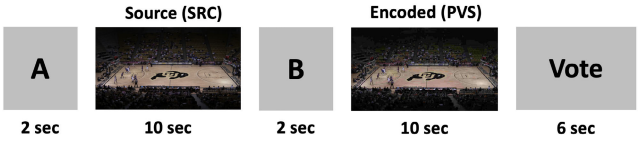


FIGURE 3. Procedure adopted during the subjective test. First, the subject watches the source video, then after two seconds the PVS, and finally provides a rating (or score) of the quality within the next six seconds.

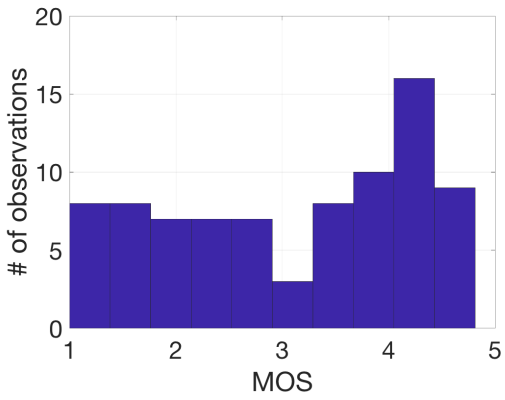


FIGURE 4. The histogram of the MOS values shows a distribution that is not far from a uniform one. This is fundamental since a different distribution of subjective scores could significantly bias the analysis' conclusions.

VI. RESULTS

In this section, we begin by assessing the reliability of the subjective ratings (MOSs) that were obtained during the subjective experiment. We then compare PVQM1 and PVQM2 to some well-known and widely used open-source VQMs. Thereafter, we assess the effectiveness of the proposed VQM disagreement measure. Finally, we show that

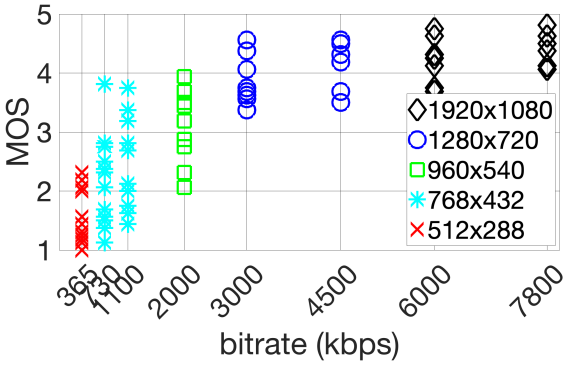


FIGURE 5. The MOS values for all the PVSs included in the test. Higher MOS values were obtained in correspondence to higher bit rates (kbps) and resolutions.

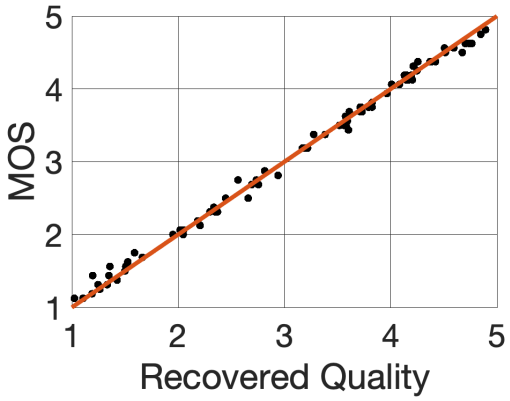


FIGURE 6. The results show that, on average, the subjects consistently evaluated the quality of the sequences used during the subjective test since the so called "Recovered Quality" of each processed video sequence does not differ significantly from the MOS.

the PVS bitstream features can be used to effectively predict the proposed VQM disagreement measure.

A. SUBJECTIVE RATINGS: DISTRIBUTION AND QUALITY

In assessing the reliability of the subjective ratings, we note that a fundamental requirement for a well-designed subjective experiment is that the subjective scores (MOSs) are uniformly distributed over the chosen quality scale or, at least, fully cover such a scale. Figure 4 shows a histogram of the MOS values obtained in our subjective experiment. The histogram shows the MOS scores span across the quality scale, and the numbers in the different bins are reasonably well balanced.

Figure 5 presents the MOS values as a function of the bit rate and the resolution. It is evident that subjects were consistent in discerning between low and high video qualities. For example, the video quality at 512 × 288@365kbps and 768 × 432@730kbps were rated lower than those encoded at 1280 × 720@3000kbps. For 1280 × 720 and 1920 × 1080 resolutions, an increment in bit rate from 3000 kbps to 4500 kbps and from 6000 kbps to 7800 kbps respectively did not result in noticeable difference in perceived quality.

TABLE 2. Comparing all VQMs in terms of accuracy.

Metrics	PLCC	SROCC	RMSE
PSNR	0.43	0.61	1.05
SSIM	0.49	0.57	1.02
MSSSIM	0.65	0.72	0.88
VIF	0.69	0.68	0.85
XPSNR	0.80	0.81	0.70
PVQM1	0.79	0.76	0.72
PVQM2	0.84	0.84	0.63
VMAF (v.0.6.1)	**0.91**	**0.91**	**0.50**

To further investigate the reliability of the MOS values, we applied Netflix's SUREAL software that implements the model proposed in [35] for subjective quality recovering. We chose such a model because there has been some evidence of its superiority over traditional approaches such as BT.500 [36] and Z-score normalisation [37]. See [35] for more details. The model recovers the so called "true subjective quality" for each PVS while automatically estimating and removing subjects' biases and inconsistencies. Figure 6 shows comparisons between the MOS obtained from the subjective test and the recovered quality (the "true subjective quality") values by the SUREAL software. As seen in Figure 6, there was a very good agreement between the two sets of values. This suggests that there were no PVSs whose evaluation had been particularly problematic to the subjects, making the dataset suitable for research despite its limited size.

B. PROPRIETARY VS OPEN SOURCE VQMS: STATISTICAL ANALYSIS

Table 2 shows the values of key statistical indicators normally used in assessing the accuracy of VQMs. The indicators are Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC) and the Root Mean Square Error (RMSE). Before computing the RMSE and the PLCC, a least square fitting of the VQMs scores to the MOS values was carried out using a logistic function as recommended in [34]. Except for VMAF and XPSNR, the open-source VQMs yielded lower correlation coefficients as compared to the two-proprietary metrics, PVQM1 and PVQM2. The PSNR and SSIM, which are still widely used within the research community had the following correlation coefficient values to the MOS values. For PSNR, the PLCC, SROCC and RMSE values were 0.43, 0.61 and 1.05 respectively. For SSIM, the same statistical indicator values were 0.49, 0.57 and 1.02 respectively. Since the correlation values are significantly less than 1, and the RMSE values are significantly greater than 0, this suggested that there were no strong similarities between the quality predictions of these two VQMs and the MOSs. On the other hand, VMAF showed higher performance than both PVQM1 and PVQM2, See Table 2.

Statistical tests were carried out to check whether the differences, in terms of accuracy, between the VQMs were statistically significant. Table 3 shows the results of the

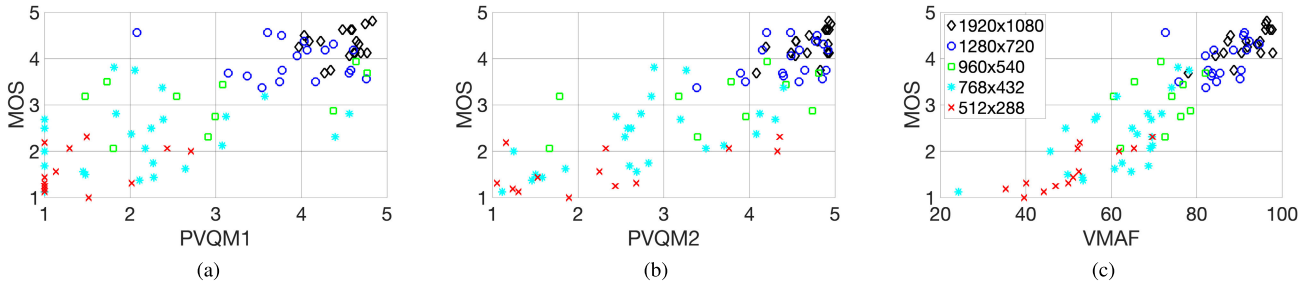


FIGURE 7. Visual comparison of the accuracy of proprietary metrics to VMAF in terms of MOS prediction. Each point corresponds to a processed video sequence and the colour represents the resolution.

TABLE 3. Statistical tests on PLCC values computed between the metrics and the MOS. Each table cell shows 1 when the PLCC of the metric in the row is higher, with statistical significance, than that of the metric in the column. VMAF predictions correlate to the MOS significantly better than other metrics.

	PSNR	SSIM	MSSSIM	VIF	XPSNR	PVQM1	PVQM2	VMAF
PSNR	-	0	0	0	0	0	0	0
SSIM	0	-	0	0	0	0	0	0
MSSSIM	1	0	-	0	0	0	0	0
VIF	1	1	0	-	0	0	0	0
XPSNR	1	1	1	0	-	0	0	0
PVQM1	1	1	1	0	0	-	0	0
PVQM2	1	1	1	1	0	0	-	0
VMAF	1	1	1	1	1	1	1	-

Z-tests conducted on each pair of VQMs. In Table 3, the value in a cell is "1" when the PLCC of the VQM in the row was statistically higher than that of the VQM in the column. VMAF predictions correlated with the MOSs significantly better than all the other VQMs. PVQM1 was seen to be significantly better than the PSNR, SSIM and MSSSIM, while PVQM2 showed superior performance when compared to PSNR, SSIM, MSSSIM and VIF.

Figure 7 shows a visual comparison between the MOS and the proprietary VQMs. VMAF (an open-source VQM) was included here for the sake of comparison. Figure 7a shows the scatter plot for PVQM1. A larger spread of points was observed when compared with PVQM2 and VMAF; see Figure 7b and 7c, respectively. This is also in line with the lower performance of PVQM1 observed in Table 2. The lower performance of PVQM1 was mostly perceptible on PVSs with lower resolutions.

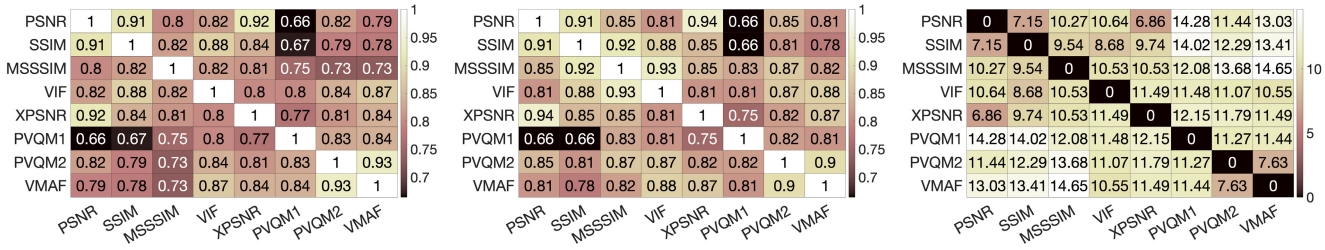
We observe that, in general, VQMs originally designed for image quality assessment (IQA) such as PSNR, SSIM, MSSSIM and VIF have reported lower performances than those of PVQM1, PVQM2 and VMAF which were developed for video. This could be explained by the fact that metrics for IQA do not consider the characteristics of the temporal dimension of the video such as the motion masking effects. However, it was important to verify this expectation as suggested and carried out by VQEG while comparing VQMs [30].

Regarding the fact that VMAF performed better than the proprietary VQMs, it is acknowledged that VMAF has a different history and circumstance to the other open-source VQMs considered in this study. Open-source VQMs, in general, mainly originate from academia where access to resources is often constrained in terms of funding and the availability of large libraries of test PVSs. However, VMAF was the result of extensive R&D efforts aimed at optimising the delivery pipeline of a major media company - Netflix. We therefore hypothesise that the design and development of VMAF may have benefited from large number of resources available to many proprietary and commercial VQMs. So, although VMAF is open-source, it is optimised enough to measure the quality as much as or better than certain commercial and proprietary tools. In fact, the results presented so far show that VMAF, as open-source metric, is a reliable benchmark from both the research and industry points of view.

The results in Table 2, Table 3 and Figure 7 compare the performances of all the VQMs in terms of MOS prediction. These results were obtained by considering only the subset of PVSs used during the subjective test. We also compared the open-source VQMs and the two proprietary VQMs using the objective scores from all the 368 PVSs. We performed a least square fitting of all the VQMs to the VMAF scale using a third-order polynomial function. This also enabled us to compare the VQMs also in terms of RMSE. The mutual RMSE (MRMSE) between two different VQMs denoted by VQM1 and VQM2 was defined as follows:

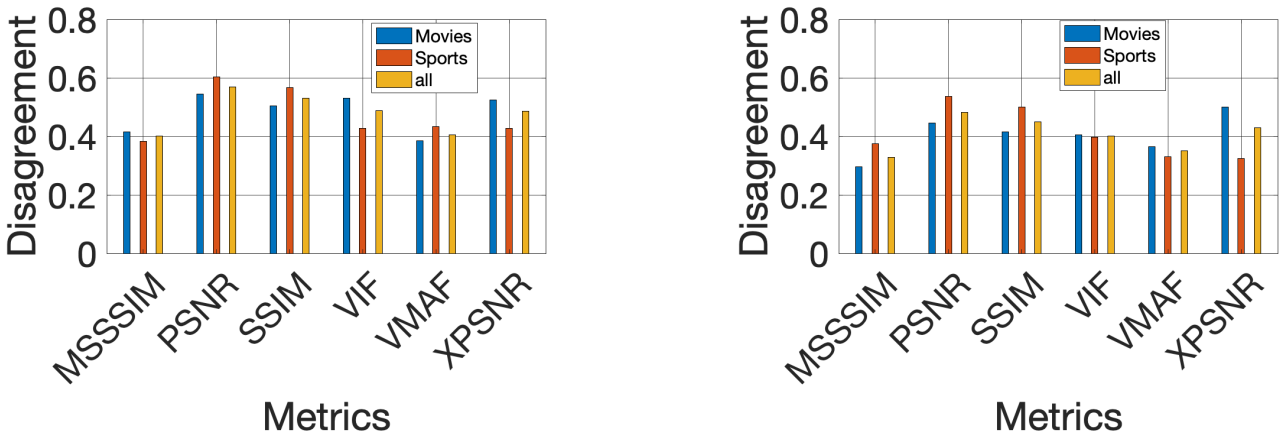
$$MRMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(V^i_{VQM1} - V^i_{VQM2}\right)^2} \qquad (2)$$

where M is the number of PVSs in the dataset (368). V^i_{VQM1} and V^i_{VQM2} are the scores obtained for the i-th PVS in the dataset after mapping VQM1 and VQM2 to the VMAF scale. The word "mutual" is used here to highlight the fact that none



(a) PLCC (b) SROCC (c) MUTUAL RMSE

FIGURE 8. Evaluating the correlation and mutual RMSE between all the metrics used in the study. In general, the proprietary metrics (PVQM1 and PVQM2) showed higher correlation to state of the art open-source metrics, as expected.



(a) PVQM1 vs Open-source VQMs (b) PVQM2 vs Open-source VQMs

FIGURE 9. Evaluating the fraction of PVSs on which the PVQMs disagree with each open-source VQM. The analysis indicates that PSNR and SSIM are more likely to measure a quality that would be perceptually different than that indicated by the PVQMs especially on sports content.

of the two metrics is the ground truth, but rather they were being compared against each other.

Figure 8 show the results obtained for the PLCC, SROCC and MRMSE between each pair of VQMs. While the SROCC was computed maintaining each VQM in its original scale, the PLCC was computed after performing the fitting of all the VQMs to the VMAF scale. In the correlation matrices in Figures 8, a correlation of at least 0.65 was obtained in all cases. Such correlations were statistically different from zero even when the statistical tests of significance were conducted at a 99% confidence level. This means that none of the VQMs in this study was totally inconsistent with respect to the other VQMs.

PVQM2 showed a strong alignment with VMAF. Their PLCC, SROCC and MRMSE values were 0.93, 0.90 and 7.63 respectively. The MRMSE value was close to the threshold of 7.00 and would suggest that the two metrics, on average, measure the same perceptual quality. Compared to the other proprietary metric, PVQM1 had slightly lower PLCC, SROCC values and slightly higher MRMSE values with VMAF. This agrees with the results of the subjective tests where PVQM1 showed lower accuracy than VMAF and PVQM2.

We note the high PLCC and SROCC correlation values of 0.92 and 0.94 respectively between PSNR and XPSNR. The equivalent correlation values between VIF and VMAF were 0.87 and 0.88. See Figure 8a and Figure 8b. Such high values could be explained by the fact that PSNR and VIF are key elements in the design of XPSNR and VMAF respectively. Therefore, this inherent correlation between PSNR and XPSNR on the one hand, and VIF and VMAF on the other hand, suggests that correlation values alone may not be enough to correctly evaluate the reliability and accuracy of the metrics. Despite the high correlation between the XPSNR and the PSNR, these two metrics yielded significantly different performances in terms of MOS prediction as shown in Table 2.

C. PROPRIETARY VS OPEN SOURCE VQMS: DISAGREEMENT ANALYSIS

We continued the analysis by evaluating the fraction of PVSs on which a proprietary metric disagreed with an open-source metric. As mentioned in Section IV, we consider two VQMs to disagree when their quality predictions, reported on the VMAF scale, differ by more than seven points. This approach considers only the range of quality variation in which the human eye is sensitive. The analysis was conducted separately for movies and sports PVSs to assess whether the content type could affect the disagreement measure.

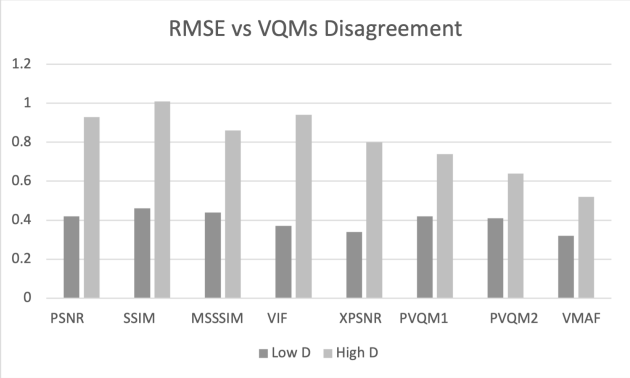


FIGURE 10. The VQMs' accuracy, in terms of RMSE, for low and high disagreement conditions. Lower RMSE is better. For all the metrics, in case of high disagreement, the predicted quality is expected to be affected by larger error.

TABLE 4. Statistical analysis of the variance of the MOS prediction error. In case of high VQMs disagreement, each metric is expected to be more inconsistent with statistical significance.

Metrics	Low D	High D	F test: p_values	Decision
PSNR	0.32	**1.23**	0.000	**yes**
SSIM	0.30	**1.14**	0.000	**yes**
MSSSIM	0.25	**0.85**	0.000	**yes**
VIFp	0.25	**0.94**	0.000	**yes**
XPSNR	0.14	**0.66**	0.000	**yes**
PVQM1	0.20	**0.58**	0.001	**yes**
PVQM2	0.20	**0.43**	0.014	**yes**
VMAF	0.12	**0.32**	0.002	**yes**

The results are shown in Figure 9. For sports content, both PSNR and SSIM disagree with the two PVQMs more than the other open-source VQMs. In fact, for 60% of the Sports PVSs, PVQM1 measured a quality that was perceptually different from those predicted by PSNR or SSIM (see Figure 9a). This percentage is reduced to 55% for PVQM2 (see Figure 9b). Note that XPSNR and PVQM2 agree significantly on Sports PVSs more than the movies PVSs.

The analysis suggests that both PSNR and SSIM were more likely to yield quality estimations that differ from the proprietary VQMs, and that viewers would be able to perceive the difference in video qualities.

D. EFFECTIVENESS OF THE PROPOSED VQMS DISAGREEMENT MEASURE

This section outlines in detail, the effectiveness of the proposed VQM disagreement measure as an indicator of VQM accuracy.

Firstly, we looked at the RMSE as an indicator of VQMs accuracy. Figure 10 shows the RMSE values for two groups of PVSs. The first group of PVSs is where the disagreement between VQMs is low (Low D); the second group of PVSs is where the disagreement between VQMs is high (High D). On average, in cases of high disagreement (High D), each VQM yielded a prediction affected by a larger deviation from the MOS. The analysis in Figure 10 indicates that a higher RMSE is expected. On the other hand, when the metrics agree, (i.e., Low D) the average of the observed RMSE values was around 0.4. This is quite interesting since this value is close to the average mutual RMSE that would be observed between MOS values obtained for the same PVSs evaluated in two different subjective experiments [38]. Therefore, this result seems to indicate that, if the proposed disagreement measure for a given PVS yields a small value, then the VQMs will provide good approximations of the perceived quality that is obtained in a subjective test for that PVS.

Statistical tests (F-test) were performed to show that VQNs are more inconsistent when predicting the MOS in case of large disagreement. In measuring the VQMs inconsistency, the variances of the residuals were taken. Residuals are the differences between the quality score predicted by the VQMs and their corresponding MOSs. Table 4 reports on the variance of each VQM's residuals for PVSs with low and high VQMs disagreements, as well as the p-value of the F-test. The F-test was performed to verify whether the variance of the residuals for each VQM was significantly larger for cases with high VQM disagreements.

Table 4 shows that for all the VQMs, the p-value of the F-test was smaller than 0.01. This means that at a 99% confidence level, the prediction error of each VQM has a variance that is larger when VQMs disagree.

This lack of accuracy observed in cases where VQMs disagreed was not caused by subject inconsistency. It was caused by intrinsic limitations in the VQMs themselves. It can be seen, for instance, that the proposed VQM disagreement measure is poorly correlated to the subject opinions' standard deviation (SOS) as shown in Figure 11a. This meant subjects did not experience any less or any more difficulty in rating the perceptual quality for cases of high VQM disagreements. We also used Netflix's SUREAL software to compute the inconsistency that affected the ratings of each individual subject who participated in the test. It can be seen in Figure 11b that each subject's inconsistency did not seem to be consistently larger in cases of high VQM disagreements.

Therefore, the indication is that the proposed VQM disagreement measure allows for the identification of PVSs whose quality will be difficult to accurately predict using a VQM. In any case, such PVSs do not pose specific challenges to human viewers because their perceptual quality can be effectively determined using subjective tests. The proposed disagreement measure can therefore be considered as a tool to identify only the PVSs for which subjective evaluation is strongly recommended, thereby reducing the number of PVSs to be used in a subjective test.

We examined the dependencies of the disagreement measure on the number and the types of VQMs (i.e. open-source or proprietary). We considered, as a reference value, the disagreement measure obtained by using in the Eq (1) all the eight VQMs considered in this work. Then, we computed the disagreement measure using only n VQMs (e.g., $n = 5, 6, 7$) chosen from the eight available VQMs, each time considering all possible combinations of the n

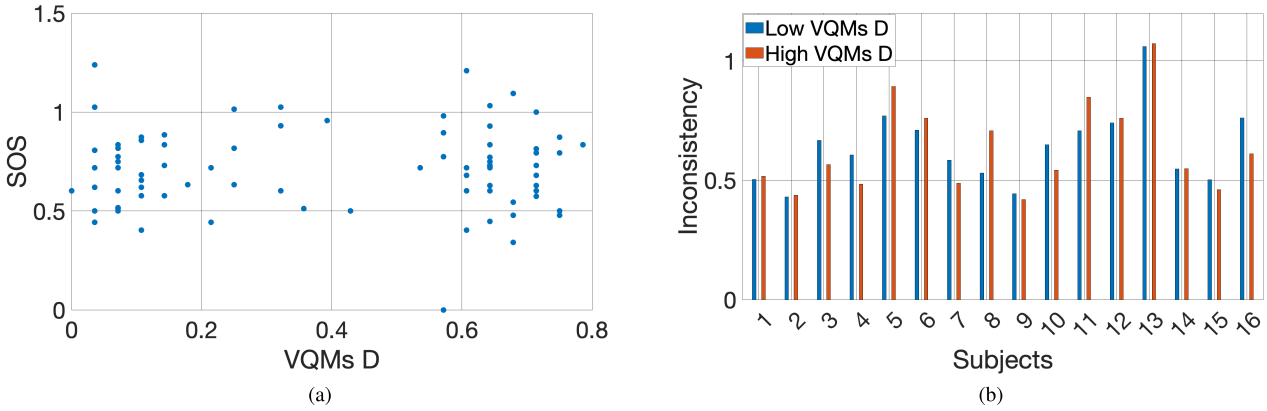


FIGURE 11. SOS and individual subjects' inconsistency as function of the proposed VQMs disagreement. Subjects seem to experience the same difficulty in assessing the quality of a PVS independently on the disagreement of the VQMs scores.

TABLE 5. Comparing the drop in performance of the VQMs when used on PVSs whose quality is difficult to evaluate, i.e., PVSs reporting a high VQM disagreement (High D). The drop (Δ) for each statistical index was computed by taking the difference between the value obtained when the metrics are expected to be highly accurate, i.e., when there is low VQM disagreement (Low D), and the one obtained in case of high VQM disagreement.

Metric	PLCC (Low D)	PLCC (High D)	Δ PLCC	Δ RMSE	Δ Variance of MOS prediction error
PSNR	0.87	0.43	-0.44	+0.51	+0.91
SSIM	0.85	0.24	-0.61	+0.55	+0.84
MSSSIM	0.86	0.51	-0.35	+0.42	+0.6
VIFp	0.90	0.42	-0.48	+0.57	+0.69
XPSNR	0.91	0.64	-0.27	+0.46	+0.52
PVQM1	0.87	0.69	**-0.18**	**+0.32**	**+0.38**
PVQM2	0.88	0.78	**-0.10**	**+0.23**	**+0.23**
VMAF	0.93	0.86	**-0.07**	**+0.20**	**+0.20**

VQMs out of eight. For example, for $n = 5$, there were 56 distinct combinations. For each combination, the RMSE between the obtained values and the reference values was computed. So, for $n = 5$, 56 values of RMSE were obtained. Note that by considering all possible combinations of VQMs for each value of n, this experiment also accounted for the impact of the VQM type used to compute the disagreement measure.

Figure 12 shows the minimum, the average, and the maximum values of RMSE for each value of n. When all combinations of five VQMs were considered, the average of the RMSE values was 0.12. For combinations where n was greater than five VQMs, an average RMSE of less than 0.08 was observed. This is less than 10% of the range [0, 1], which represents the range of variation of the disagreement measure. This average RMSE value can therefore be considered very reasonable. For the minimum and maximum RMSE values, we noted that the difference between them did not exceed 0.07 for any combination of n VQMs. This difference of 0.07 represents 7% of the variation range of the disagreement measure. So, using any combination of VQMs to estimate the reference disagreement value would not vary the average estimation error by more than 7% of the variation range of the disagreement measure.

The results obtained for the RMSE showed that the proposed disagreement measure is not very sensitive to the number and type of VQMs used to compute it.

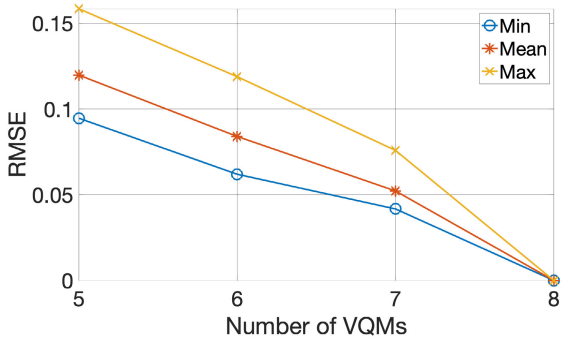


FIGURE 12. Analysis of the impact of the number of VQMs on the proposed disagreement measure. The values of the disagreement computed by using all the eight VQMs considered in this paper is taken as the reference disagreement value or ground truth. The Figure shows the RMSE between the reference disagreement value and the disagreement computed using n (n = 5, 6 and 7) VQMs. For each value of n, all possible combinations of n metrics out of eight are used to compute the disagreement. The minimum, the mean and the maximum value obtained for each n is then reported.

To further study the impact of the VQM type on the proposed disagreement measure, we computed the measure using only open-source VQMs and then checked whether the measure remained a good indicator of the accuracy of the VQMs. The results are shown in Figure 13. As observed, the results were very consistent with those shown in Figure 10 where the disagreement was obtained considering all eight metrics. In other words, when there was high disagreement

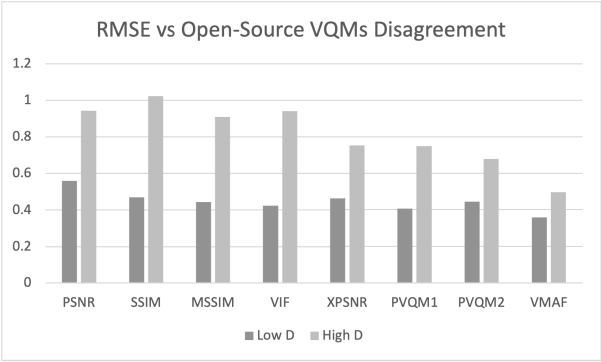


FIGURE 13. The video quality metrics' accuracy, in terms of RMSE, for low and high disagreement of open-source VQMs. Lower RMSE is better. For all the metrics, in case of high disagreement, the predicted quality is expected to be affected by larger error.

from the open-source metrics considered in this study, a lower accuracy was observed from the metrics when used in predicting the MOS. This result is quite interesting because even if the two PVQMs were not considered, the proposed disagreement measure still provided significant indications on the accuracy of all VQMs. This suggests that our proposal could be used to deduce the accuracy of any metric in the literature that had the same design scope as those considered in this study.

In Table 5, we compared the performance drop of the different VQMs when used on PVSs whose quality assessment was challenging rather than on those that were easy to evaluate. The results in Figure 10 and Table 4, show that the challenging PVSs were those corresponding to a higher value of disagreement measure, and vice versa. Therefore, for each statistical indicator in Table 5, the drop Δ was calculated by taking the difference between the values obtained respectively on the PVSs with high disagreement and those with low disagreement. It is very interesting to note that excluding VMAF, all open-source VQMs had a higher accuracy drop than the proprietary ones when moving from low to high disagreement PVSs. Specifically, PVQM1, which is a proprietary metric, had the greatest drop in accuracy, it showed a +0.32 RMSE increase and a -0.18 MOS correlation decrease. On the other hand, the lowest performance drop observed among open-source metrics (excluding VMAF) was +0.42 and -0.27 for RMSE and PLCC respectively. Similar considerations can be made for the variance of the MOS prediction error. These results showed that VMAF and the PVQMs were more robust when a PVS was more likely to confuse or mislead VQMs. These VQMs may therefore be expected to deliver better estimations of quality on challenging PVSs. Finally, we note that, for all VQMs, lower PLCC values were observed in correspondence with PVSs with high VQM disagreement.

E. TOWARDS MODELLING AND PREDICTING VQM DISAGREEMENT

The bitstream features of each of the 368 PVSs were extracted. The key features of the bitstream information were the bit rate, the average quantization parameter (QP), standard deviation of QP over the PVS's frames, the average motion vector (MV) components, standard deviation of MV components, percentage of Intra and Inter coded blocks, the percentage of each block size and the percentage of skipped blocks. These features were extracted at the single block level and later pooled into a single value using both the average and the Minkowski norm for each PVS. A total of 104 features were extracted for each PVS.

A backward sequential feature selection algorithm [39] was then used to find the bitstream features that were important in predicting the VQM disagreement. The features that were seen to have major importance were the average QP, the average MV in each direction X and Y, the percentage of Intra blocks in a slice and the percentage of 2Nx2N Intra coded blocks. We also experimentally found that the best pooling strategy is the Minkowski norm when the exponent is set to $p = 1.3$.

After determining the best set of features, they were regressed to the disagreement measure using different machine learning (ML) algorithms. We considered a few models such as linear regression model (LM), regression tree (RT), neural network (NN) with a single hidden layer having four neurons, support vector regression model with a Gaussian kernel (SVR Gaus) and support vector regression model with a radial basis function kernel (SVR rbf). The 368 PVSs were divided into 10 folds, and all the models were trained on 9 folds and tested on the one left out.

TABLE 6. PLCC values obtained when comparing different machine learning models for regressing the bitstream features to the proposed measure of VQMs disagreement. Support vector regression with the radial basis function as kernel yielded the best performance.

Folds	LM	RT	NN	SVR (Gaus)	SVR (rbf)
Fold 1	0.65	0.81	0.78	0.85	**0.93**
Fold 2	0.53	0.70	0.60	0.65	**0.80**
Fold 3	0.47	0.59	0.59	0.57	**0.74**
Fold 4	0.42	0.46	0.55	0.77	**0.91**
Fold 5	0.50	0.73	0.64	0.78	**0.88**
Fold 6	0.40	0.54	0.52	0.65	**0.83**
Fold 7	0.48	0.41	0.48	0.61	**0.78**
Fold 8	0.73	0.75	0.72	0.84	**0.90**
Fold 9	0.65	0.73	0.75	0.82	**0.95**
Fold 10	0.64	0.68	0.74	0.75	**0.75**
Overall	0.56	0.66	0.65	0.74	**0.86**

TABLE 7. SROCC values obtained when comparing different machine learning models for regressing the bitstream features to the proposed measure of VQMs disagreement. Support vector regression with the radial basis function as kernel yielded the best performance.

Folds	LM	RT	NN	SVR (Gaus)	SVR (rbf)
Fold 1	0.59	0.68	0.60	0.78	**0.88**
Fold 2	0.54	0.66	0.60	0.67	**0.84**
Fold 3	0.48	0.63	0.62	0.64	**0.79**
Fold 4	0.38	0.45	0.48	0.72	**0.87**
Fold 5	0.53	0.74	0.59	0.71	**0.86**
Fold 6	0.52	0.56	0.54	0.65	**0.84**
Fold 7	0.56	0.47	0.51	0.68	**0.85**
Fold 8	0.76	0.75	0.72	0.86	**0.92**
Fold 9	0.68	0.74	0.79	0.84	**0.95**
Fold 10	0.67	0.67	0.66	0.73	**0.73**
Overall	0.58	0.65	0.62	0.74	**0.87**

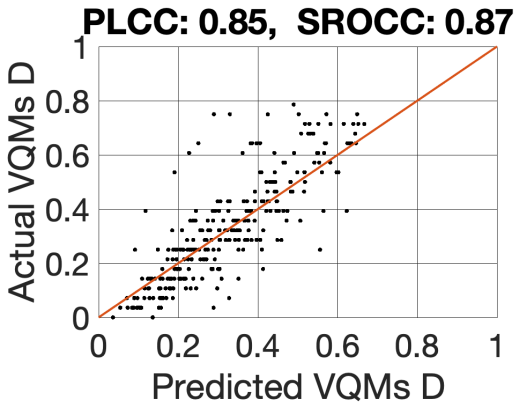


FIGURE 14. Accuracy of the final SVR model on all the data. Despite some outliers, in general the model has been able to satisfactory model the metrics disagreement, yielding high linear (0.85) and rank correlation values (0.87).

The results are shown in Table 6 and Table 7. The overall performance was determined by computing the inverse transform of the average Fisher's Z transformation of single correlation scores as recommended in [40].

For all testing conditions, the linear model yielded a PLCC and a SROCC significantly different from 0, and showed lower performance than other algorithms. Thus, the relationship between the selected features and the VQMs disagreement is probably not trivial. The SVR-based models, and particularly the SVR model (with an rbf kernel), provided the highest performance, reaching a global linear and rank correlation of 0.85 and 0.86 respectively.

The final SVR model (with an rbf kernel) was trained using all the data available in the dataset. The scatter plot in Figure 14 illustrates the performance of the final SVR model on the whole dataset. In general, its predictions correlated quite well with the actual value of the VQMs disagreement.

The proposed VQMs disagreement measure in Eq (1) was related to the VQMs accuracy through the results in Figure 10 and Table 4. The final SVR model (with an rbf kernel) was also able to accurately predict the proposed VQMs disagreement measure using the PVS bitstream features as shown in Figure 14. This suggests that it is possible to determine the accuracy of any VQM on a given PVS, by just relying on its bitstream features, without the need to compute many full reference VQMs (particularly the proprietary ones). In other words, there is a link between the ways a PVS is encoded and the difficulty in accurately evaluating its quality with VQMs.

VII. CONCLUSION

In this work, a way to quantify VQM disagreement was proposed. A dataset comprising 368 PVSs was created for the analysis. A subset of those PVSs was selected for subjective evaluation based on the proposed VQM disagreement measure.

Unlike many studies in the literature that analysed only open-source video quality metrics, our study considered two proprietary metrics used in the content delivery chain by some media industries to optimise their content preparation and delivery pipeline. A comparison analysis between some well-known and widely used open-source metrics and the proprietary metrics was conducted. The results showed that VMAF yielded better performance than the proprietary metrics with statistical significance, while the latter showed higher accuracy than most of the open-source metrics.

It was shown that the proposed VQM disagreement measure can be used to determine a VQM's accuracy when estimating the MOS. Statistical analyses showed that when the VQMs agreed, the commonly predicted objective score was an accurate estimation of the MOS. The proposed disagreement measure can therefore be considered as a tool to identify only the PVSs for which subjective evaluation is strongly recommended, thereby reducing the number of PVSs to be used in a subjective test. Finally, it was observed that the proposed VQM disagreement measure can be effectively predicted from bitstream features. This shows that there is a link between the way a PVS is encoded and the difficulty in objectively assessing its perceptual quality.

The small-scale subjective experiment that was carried out in the context of this work showed promising results. Future work will consider the possibility of designing larger datasets for a deeper investigation into the potential implications of the disagreement of video quality measures.

ACKNOWLEDGMENT

This work was supported in part by Sky Group and Affiliates, and in part by the PoliTO Interdepartmental Centre for Service Robotics (PIC4SeR) (http://pic4ser.polito.it). The results presented here were a joint collaborative effort between the academic parties and Sky Group. The authors would like to acknowledge the following individuals for their expertise throughout this project: Emanuele D'Addea (Sky Italy) and Adrian Baniewicz (Sky Germany).

REFERENCES

[1] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[2] A. Raake, S. Borer, S. Satti, J. Gustafsson, R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, G. Heikkilä, S. Broom, C. Schmidmer, B. Feiten, U. Wüstenhagen, T. Wittmann, M. Obermann, and R. Bitto, "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204," *IEEE Access*, vol. 8, pp. 193020–193049, 2020.

[3] T. Richter, J. Keinert, S. Foessel, A. Descampe, G. Rouvroy, and J.-B. Lorent, "JPEG-XS—A high-quality mezzanine image codec for video over IP," *SMPTE Motion Imag. J.*, vol. 127, no. 9, pp. 39–49, 2018.

[4] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[7] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[8] C. R. Helmrich, M. Siekmann, S. Becker, S. Bosse, D. Marpe, and T. Wiegand, "Xpsnr: A low-complexity extension of the perceptually weighted peak signal-to-noise ratio for high-resolution video quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2727–2731.
[9] Netflix. (May 2018). *VMAF Video Multi-Method Assessment Fusion V.0.6.2*. [Online]. Available: https://github.com/Netflix/vmaf
[10] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Video Quality Estimation Module*, document ITU-T Rec. P.1203.1, Jan. 2019.
[11] H. Liu, M. Lu, Z. Ma, F. Wang, Z. Xie, X. Cao, and Y. Wang, "Neural video coding using multiscale motion compensation and spatiotemporal context model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3182–3196, Aug. 2021.
[12] A. S. Panayides, M. S. Pattichis, M. Pantziaris, A. G. Constantinides, and C. S. Pattichis, "The battle of the video codecs in the healthcare Domain–A comparative performance evaluation study leveraging VVC and AV1," *IEEE Access*, vol. 8, pp. 11469–11481, 2020.
[13] S. Deng, J. Han, and Y. Xu, "VMAF based rate-distortion optimization for video coding," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.
[14] A.-N. Moldovan and C. H. Muntean, "QoE-aware video resolution thresholds computation for adaptive multimedia," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2017, pp. 1–6.
[15] Y. Li and X. Mou, "Joint optimization for SSIM-based CTU-level bit allocation and rate distortion optimization," *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 500–511, Jun. 2021.
[16] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T Rec. P.1401, Jul. 2012.
[17] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.
[18] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C.-J. Kuo, "A fusion-based video quality assessment (fvqa) index," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Asia–Pacific*, Dec. 2014, pp. 1–5.
[19] L. F. Tiotsop, A. Servetti, and E. Masala, "Full reference video quality measures improvement using neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2737–2741.
[20] L. F. Tiotsop, T. Mizdos, M. Uhrina, M. Barkowsky, P. Pocta, and E. Masala, "Modeling and estimating the subjects' diversity of opinions in video quality assessment: A neural network based approach," *Multimedia Tools Appl.*, vol. 80, pp. 3469–3487, Sep. 2020.
[21] L. F. Tiotsop, E. Masala, A. Aldahdooh, G. V. Wallendael, and M. Barkowsky, "Computing Quality-of-Experience ranges for video quality estimation," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
[22] A. Aldahdooh, E. Masala, G. Van Wallendael, and M. Barkowsky, "Comparing temporal behavior of fast objective video quality measures on a large-scale database," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.
[23] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.
[24] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[25] C. Mello, M. Saraiva, D. Phoenix, and R. Nishihara, "A comparative study of objective video quality assessment metrics," *J. Universal Comput. Sci.*, vol. 23, pp. 505–527, 01 2017.
[26] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, pp. 1–12, Dec. 2013.
[27] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P.910, Apr. 2008.
[28] Apple. (2020). *HLS Authoring Specification for Apple Devices*. [Online]. Available: http://apple.co/39VrP6t
[29] (May 2021). *Video Quality Experts Group (VQEG)*. [Online]. Available: http://vqeg.org
[30] VQEG. (Jun. 2010). *Report on the Validation of Video Quality Models for High Definition Video Content (V. 2.0)*. [Online]. Available: http://bit.ly/2Z7GWDI
[31] Z. Sinno, A. K. Moorthy, J. De Cock, Z. Li, and A. C. Bovik, "Quality measurement of images on mobile streaming interfaces deployed at scale," *IEEE Trans. Image Process.*, vol. 29, pp. 2536–2551, 2020.
[32] T. Rahim, M. A. Usman, and S. Y. Shin, "Comparing H.265/HEVC and VP9: Impact of high frame rates on the perceptual quality of compressed videos," 2020, *arXiv:2006.02671*.
[33] *Personal Communication*, Netflix Developers, Los Gatos, CA, USA, Jan. 2021.
[34] *Method for Specifying Accuracy and Cross-Calibration of Video Quality Metrics (VQM)*, document ITU-T Rec. J.149, Mar. 2004.
[35] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 52–61.
[36] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-T Rec. BT.500, Jan. 2012.
[37] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[38] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. SPIE*, vol. 5150, pp. 573–582, Jun. 2003.
[39] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning From Data*. New York, NY, USA: Springer, 1996, pp. 199–206. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4612-2404-4_19#citeas
[40] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.

LOHIC FOTIO TIOTSOP (Member, IEEE) received the M.Sc. degree in mathematical engineering from the Politecnico di Torino, Italy, where he is currently pursuing the Ph.D. degree in control and computer engineering. His primary research interests include advanced statistical methods and machine learning algorithms applied to multimedia problems.

FLORENCE AGBOMA received the Ph.D. degree in computing and electronics systems from the University of Essex, U.K., in 2009. She is currently the Streaming Quality Manager at Sky, U.K. Her work includes improving the streaming quality of Sky's OTT services, inclusive of Live and VOD. Her interests include PayTV analytics, quality of experience management, and emerging broadcast TV technologies.

GLENN VAN WALLENDAEL (Member, IEEE) received the M.Sc. degree in computer science engineering from Ghent University, Belgium, in 2008, and the Ph.D. degree from IDLab, Ghent University, with the financial support of the Research Foundation—Flanders (FWO). Since 2019, he has been working as a Professor at Ghent University and imec on topics, such as the efficient representation and compression of visual information, including 360 degree video, light field, virtual reality, and the different operations on these modalities, such as (scalable) compression, transcoding, encryption, watermarking, personalized delivery, and quality estimation.

AHMED ALDAHDOOH received the master's degree in multimedia and data management from the Polytech Nantes, Nantes University, France, in 2014, and the Ph.D. degree in IT and its applications from LS2N, Polytech Nantes, in 2017. In 2020, he joined IETR, INSA Rennes, as a Research Engineer. His main research interests include content-aware video delivery, video quality, image and video processing, error concealment, deep learning, safety of deep learning, and adversarial examples detection.

SEBASTIAN BOSSE (Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and information technology from RWTH Aachen, Germany, and the Dr.-Ing. degree from Technical University Berlin, Germany, in 2018. He is currently with the Department of Vision and Imaging Technologies, Fraunhofer Heinrich Hertz Institute, Berlin, Germany, where he heads the Interactive and Cognitive Systems Group. His research interests include human–machine interaction, human perception models and its applications in multimedia systems, and machine learning.

LUCJAN JANOWSKI received the Ph.D. degree in telecommunications from the AGH University of Science and Technology, Krakow, Poland, in 2006. In 2007, he was a Postdoctoral Researcher with the Laboratory for Analysis and Architecture of Systems, Centre National de la Recherche Scientifique, Paris, France. From 2010 to 2011, he was a Postdoctoral Researcher with the University of Geneva, Geneva, Switzerland. From 2014 to 2015, he was a Postdoctoral Researcher with the Telecommunications Research Center Vienna, Vienna, Austria. He is currently an Assistant Professor with the Department of Telecommunications, AGH University of Science and Technology. His research interests include statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation.

MARCUS BARKOWSKY (Member, IEEE) received the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Germany, in 2009. He joined the University of Nantes, France, in 2010, then in 2018, he obtained the professorship on interactive systems and the Internet of Things at the Deggendorf Institute of Technology, University of Applied Sciences, Germany. His activities range from designing 3-D interaction and measuring visual discomfort using psychometric measurements to computationally modeling spatial and temporal effects of the human perception.

ENRICO MASALA (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Torino, Italy, in 2004. He is currently an Associate Professor with the Politecnico di Torino. His main research interests include multimedia quality optimization of communications over packet networks, with special attention to particular scenarios, such as remote control applications, 3-D video, and cloud for multimedia. He is also involved in the management of the Politecnico di Torino Interdepartmental Center for Service Robotics (PIC4SeR).