Computing Quality-of-Experience Ranges for Video Quality Estimation

2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)

Computing Quality-of-Experience Ranges for Video Quality Estimation

Lohic Fotio Tiotsop[1], Enrico Masala[1], Ahmed Aldahdooh[2], Glenn Van Wallendael[3], Marcus Barkowsky[4]

Comp. Eng. Dept.[1] *Politecnico di Torino* Torino, Italy

IT Department[2] *Univ. College of Applied Sciences* Gaza, Palestine

imec - IDLab[3] *Ghent University* Ghent, Belgium

Deggendorf Inst. of Technology (DIT)[4] *University of Applied Sciences* Deggendorf, Germany

first.lastname@polito.it, adahdooh@ucas.edu.ps, Glenn.VanWallendael@UGent.be, marcus.barkowsky@th-deg.de

***Abstract*—Typically, the measurement of the Quality of Experience for video sequences aims at a single value, in most cases the Mean Opinion Score (MOS). Predicting this value using various algorithms has been widely studied. However, deviation from the MOS is often handled as an unpredictable error. The approach in this contribution estimates intervals of video quality instead of the single valued MOS. Well-known video quality estimators are fused together to output a lower and upper border for the expected video quality, on the basis of a model derived from a well-known subjectively annotated dataset. Results on different datasets provide insight on the suitability of the well-known estimators for this particular approach.**

***Index Terms*—Quality of Experience, video quality, large scale evaluation, QoE ranges**

I. Introduction

In general, subjective scores are modeled as a Gaussian distribution and thus, subjective experiments lead to a mean opinion score (MOS) and a Gaussian-modeled confidence interval for this MOS. The assumption that this distribution is indeed Gaussian and that a single value is sufficient, is challenged by the diversity of source contents, the number and type of degradation, especially multi-dimensional degradations where dimensions can be, for instance, image distortions and temporal distortions. Objective measures [1] may try to predict such confidence intervals but, in addition, they introduce further uncertainties that are, at least, bound to the quality range. For instance, just noticeable difference (JND) measures are usually better for high quality (small differences) than for low quality.

This paper, instead, aims to be a first step towards a more generic approach in which we propose to estimate a QoE range instead of a single QoE value. For simplicity, we are not trying to model an individual distribution per sequence but we are restricting to a minimum and maximum value. To the best of our knowledge, this is the first attempt to work in this direction. The motivation and examples of the usefulness of such approach are presented in more details in Section II.

Our proposed approach should not to be confused with confidence interval estimation for video quality, which starts from the idea of modeling the intrinsic uncertainty of opinions of human subjects as a Gaussian distribution.

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

According to our proposed approach, we attempt to compute a score range, i.e., a minimum and maximum value, starting from well-known objective full-reference video quality measures (VQMs) that can be easily computed for each processed video sequence (PVS). The underlying idea is that the use of several VQMs, each based on different approaches, could somehow capture the multi-dimensional degradations that may affect PVSs.

To test this idea, we rely on the results of the VQEG HDTV Phase I experiment [2] (VQEG-HD) which is one of the most extensive subjectively-annotated publicly-available datasets, with a large variety of high resolution (1920x1080) content. We assume that such dataset reasonably covers the large majority of cases in which the video quality research community could be interested. As it will be shown by the results of this paper, this assumption seems quite reasonable when our results are tested on other independently created datasets. To make the method more practical, when determining the extremes of the ranges, we accept that in a given percentage of cases the actual MOS value does not lie within the range. However, this percentage can be tuned as desired, making the method flexible in this regard.

II. Motivation

In Section I we highlighted several reasons that cause an intrinsic difficulty in trying to determine a single QoE value for each test case, e.g., each PVS, regardless of the goodness of the algorithm used to estimate such single QoE value. In this work we argue that it may be worth investigating the QoE estimation problem from a different angle, i.e., attempt to predict a QoE range instead of a single QoE value.

From a practical point of view, estimating a range without even attempting to compute a single QoE estimate can be useful in many situations. For instance, for quality assurance purposes, it could be enough to know that a certain minimum QoE is met, whereas the actual QoE value is of less interest. Another example could be the design of a dataset for a subjective experiment in which it is desirable to perform a first screening to ensure that the contained samples are a good choice in terms of variety of quality.

We again stress that we are not seeking to determine a confidence interval, since we are interested in determining the bounds but without any assumption about where the actual

QoE value could be. To this aim, we believe that taking advantage of all the variability in quality estimation provided by the different approaches employed by well-known objective quality measures could yield interesting results, as it will be shown in the remainder of the paper.

III. Range Estimation

We consider the VQEG-HD experiment [2]. On that dataset we computed a set of full-reference VQM measures for each PVS, in particular PSNR, SSIM, MSSSIM, VIF [3], and VMAF 0.6.2 [4]. Since most of the measures do not easily handle interlaced video, we restrict our analysis to non-interlaced sequences, i.e., the vqeghd1, vqeghd3 and vqeghd5 subsets. The sequences considered in the VQEG-HD experiment have been chosen in order to cover a large set of content, conditions, and quality ranges, therefore we expect it to be a rather good representation of the conditions that can be encountered in the majority of real-world applications. Please note that for the VQEG-HD complete dataset, created by joining several experiments (see Chapter 7 of [2]), the MOS scores range in [0.82, 5.26].

Let $V = (vqm_1, vqm_2, \ldots, vqm_n)$ be a vector containing the objective scores determined for a PVS by n VQMs and $\alpha \in$ [0,1] that models our tolerance, i.e., the percentage of cases in which the MOS can be outside the range, we should find the values mos_{Min} and mos_{Max} such that

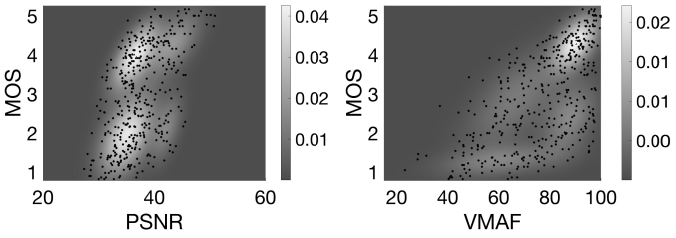
$$\begin{array}{rcl} \Pr\left(MOS \leq mos_{Min}|V\right) & = & \alpha/2, \\ \Pr\left(MOS \geq mos_{Max}|V\right) & = & \alpha/2, \end{array} \quad (1)$$

where Pr means probability. While considering all the VQMs together is certainly the most desirable approach to the problem, initially we consider, for simplicity and for easier graphical interpretation, each VQM individually.

First, we propose to model the points in the dataset for a given VQM as a joint distribution $f(VQM, MOS)$ using a 2D Gaussian mixture model (GMM):

$$f(VQM, MOS) = \sum_{i=1}^{k} \pi_i \cdot N\Big((VQM, MOS)|\mu_i, \Sigma_i\Big) \quad (2)$$

Where $N\left((VQM, MOS)|\mu_i, \Sigma_i\right)$ is the p.d.f. of a bivariate normal distribution with mean μ_i and covariance matrix Σ_i and k is the number of components. Note that we use GMM since, with a suitable number of components, they can approximate any distribution, but using GMM does not imply that the MOS distribution of a single PVS is assumed to be Gaussian. The parameters (π_i, μ_i, Σ_i) have been estimated from the data collected during the VQEG-HD experiment using the expectation maximization (EM) algorithm. Such algorithm provides some criteria to determine which is the best number of components to use. We employed the Bayesian information criterion (BIC) to determine an optimal number of components to use for each VQM, i.e., the point at which the BIC curve (as a function of k) becomes almost flat. In practice, this means that either $k = 3$ or $k = 4$ depending on the VQM can be used [5]. Fig. 1 shows that the proposed model accurately fits the density of the points in the dataset.



(a) $f(PSNR, MOS)$ (b) $f(VMAF, MOS)$

Fig. 1: The 2D GMM obtained for the PSNR and VMAF, with the original points in the dataset.

Due to lack of space we only present the results for the PSNR and VMAF. The other VQMs exhibit similar behaviour.

Once a suitable 2D GMM is fitted to the data it is possible to compute the conditional probability for a given VQM interval. We divided the useful interval of each VQM in the dataset into 100 equal parts. For the center vqm_j of each interval:

$$\begin{array}{l} G(vqm_j, mos) = \\ \Pr\left(MOS \leq mos | vqm_j - \delta \leq VQM \leq vqm_j + \delta\right) \end{array} \quad (3)$$

where $\delta = (\max(VQM) - \min(VQM))/100$. For each of those vqm_j points, the following equations are solved for mos to determine the MOS bounds for that particular vqm_j value:

$$\begin{array}{rcl} G(vqm_j, mos_{Min}^{vqm_j}) & = & \alpha/2, \\ G(vqm_j, mos_{Max}^{vqm_j}) & = & 1 - \alpha/2. \end{array} \quad (4)$$

The 100 $mos_{Min}^{vqm_j}$ values are interpolated to obtain a continuous curve $mos_{Min}(\mathrm{VQM})$. The same applies to the max curve. Fig. 2 shows the curves obtained by interpolating the 100 $mos_{Min}^{vqm_j}$ (and max) points, for different α values, as well as the original points in the dataset. It can be observed, for instance, in Fig. 2a that, for a PVS with PSNR equal to 47 dB, the MOS is expected to be in the range [3,5] with 90% probability.

The min and max values for each single VQM obtained in Eq. (4) can later be combined together to obtain a global min and max MOS value for a specific PVS. In this work, we propose to average the min and max values of each of the N VQMs to obtain the global value:

$$\begin{array}{l} mos_{Min}^{\mathrm{PVS}} = \frac{1}{N} \sum_n \left(mos_{Min}(\mathrm{VQM}_n^{\mathrm{PVS}})\right) \\ mos_{Max}^{\mathrm{PVS}} = \frac{1}{N} \sum_n \left(mos_{Max}(\mathrm{VQM}_n^{\mathrm{PVS}})\right) \end{array} \quad (5)$$

Therefore, at the end of the procedure, when a new PVS with unknown MOS is presented to our system, we first compute the five VQM values, then we use the curves $mos_{Min}(\mathrm{VQM})$ and $mos_{Max}(\mathrm{VQM})$ to obtain the min and max MOS for each VQM, then we aggregate those values using the average to form the final MOS range for that PVS.

IV. Results

To validate the effectiveness of our proposed system, we employed two datasets different from the used VQEG-HD dataset containing sequences not considered by the proposed system. Both datasets include high resolution content (1920x1080). The first is the Netflix Public Dataset [6], which includes

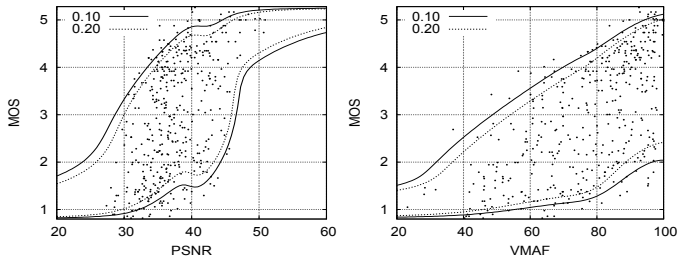


Fig. 2: $mos_{Min}(VQM)$ and $mos_{Max}(VQM)$ curves as a function of the VQM values, for two different α values, shown in the legend. Each point represents a PVS in the VQEG-HD dataset. $MOS \in [0.82, 5.26]$ due to realignment of the VQEG-HD subsets [2].

70 subjectively annotated PVSs covering the full MOS range. The second is the VQEG JEG-Hybrid Large Scale Database (JEG-DB) [7] which includes 19,840 1080p PVSs obtained by compressing a few source sequences in HEVC format using a large set of coding parameters, including bitrates ranging from 500 Kbps to 16 Mbps.

First, we focus on the Netflix subjectively-annotated dataset. Table I shows that our proposed system can compute MOS ranges accurately despite it being constructed from an unknown dataset. In particular, the fraction of MOS values outside the range is close to the expected one, determined by the α value. In all cases, the number of PVSs falling outside the range differ from the expected one for max 8 units.

Then, we consider the JEG-DB, a use case with a huge number of PVSs for which no MOS information is available. The top part of Fig. 3 shows the distribution of the length of the ranges we obtained, as a function of the center of the predicted ranges. Clearly, as α increases, the size of the range (the MOS "spread") decreases. Moreover, as expected, when the bitrate of the PVS is at one extreme (low or high), the range size is reduced, i.e., there is less doubt on the MOS position, respectively low or high. However, for intermediate values, the range size increases. Despite not having MOS values, it is however possible to spot interesting sequence behaviours. In the bottom part of Fig. 3 for instance, we highlighted the points corresponding to two source contents (all others are in grey): the blue shows a sequence (a cartoon) which exhibits a quite peculiar behavior in terms of MOS (less uncertainty for high quality), whereas manual inspection showed that the red points correspond to sequences with some digital noise in the original source. This simple analysis underlines the usefulness of being able to estimate even just MOS ranges to identify interesting behaviors in a large database of video sequences.

V. Conclusions

In this work we presented a different approach to MOS estimation that attempts to predict a MOS range rather than a single value. A methodology to create such values has been proposed and applied to the VQEG-HD dataset. The methodology is able to determine the MOS range for a PVS as a function of 5 well-known objective quality metrics.

TABLE I: No. of PVS whose MOS value is outside the range.

α	VQEG-HD		Netflix Public	
	Expected	Actual	Expected	Actual
0.01	4/415	0/415	1/70	0/70
0.05	21/415	21/415	4/70	4/70
0.10	42/415	44/415	7/70	13/70
0.15	63/415	70/415	11/70	19/70
0.20	84/415	85/415	14/70	23/70

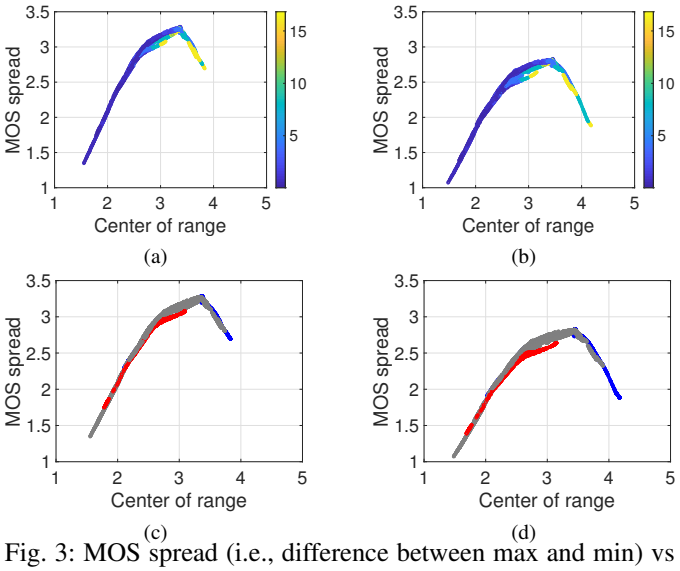


Fig. 3: MOS spread (i.e., difference between max and min) vs center of range on JEG-DB. α=0.10 (left), 0.20 (right). Colors represent actual PVS bitrate (Mbps) (top), different sources (bottom).

Results provide significant insight when cross-tested against other video quality datasets. Work is ongoing to improve the technique by using other modeling functions, cross-testing on other datasets, and to refine the method by exploring, for instance, different strategies to combine predictions by various VQMs.

VI. Acknowledgment

This work has been supported in part by PIC4SeR (http://pic4ser.polito.it). Some of the computational resources were provided by HPC@POLITO (http://www.hpc.polito.it).

References

[1] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1–19, 2013.

[2] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)," http://bit.ly/2Z7GWDI, Jun. 2010.

[3] P. Hanhart and R. Hahling, "Video quality measurement tool (VQMT)," http://mmspg.epfl.ch/vqmt, Sep. 2013.

[4] Netflix, "VMAF - video multi-method assessment fusion v.0.6.2," https://github.com/Netflix/vmaf, May 2018.

[5] A. Aldahdooh, E. Masala, G. Van Wallendael, P. Lambert, and M. Barkowsky, "Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures," *Signal Processing: Image Communication*, vol. 74, pp. 32 – 41, 2019.

[6] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data Compression Conference (DCC)*, April 2017, pp. 52–61.

[7] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnström, N. Staelens, and P. Le Callet, "Objective video quality assessment-towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98B, no. 1, pp. 2–11, 2015.