POLITECNICO DI TORINO
Repository ISTITUZIONALE

Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach

Original
Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach / FOTIO TIOTSOP, L., Mizdos, T., Uhrina, M., Barkowsky, M., Pocta, P., Masala, E.. - In: MULTIMEDIA TOOLS AND APPLICATIONS. - ISSN 1380-7501. - STAMPA. - 80:(2021), pp. 3469-3487. [10.1007/s11042-020-09704-w]

Availability:
This version is available at: 11583/2846319 since: 2020-09-22T10:26:00Z

Publisher:
Springer

Published
DOI:10.1007/s11042-020-09704-w

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

14 October 2026

Multimedia Tools and Applications (2021) 80:3469–3487
https://doi.org/10.1007/s11042-020-09704-w

Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach

Lohic Fotio Tiotsop[1] · Tomas Mizdos[2] · Miroslav Uhrina[2] · Marcus Barkowsky[3] · Peter Pocta[2] · Enrico Masala[1]

Received: 7 April 2020 / Revised: 16 July 2020 / Accepted: 25 August 2020 /
Published online: 22 September 2020
© The Author(s) 2020

Abstract
Subjective experiments are considered the most reliable way to assess the perceived visual quality. However, observers' opinions are characterized by large diversity: in fact, even the same observer is often not able to exactly repeat his first opinion when rating again a given stimulus. This makes the Mean Opinion Score (MOS) alone, in many cases, not sufficient to get accurate information about the perceived visual quality. To this aim, it is important to have a measure characterizing to what extent the observed or predicted MOS value is reliable and stable. For instance, the Standard deviation of the Opinions of the Subjects (SOS) could be considered as a measure of reliability when evaluating the quality subjectively. However, we are not aware of the existence of models or algorithms that allow to objectively predict how much diversity would be observed in subjects' opinions in terms of SOS. In this work we observe, on the basis of a statistical analysis made on several subjective experiments, that the disagreement between the quality as measured by means of different objective video quality metrics (VQMs) can provide information on the diversity of the observers' ratings on a given processed video sequence (PVS). In light of this observation we: i) propose and validate a model for the SOS observed in a subjective experiment; ii) design and train Neural Networks (NNs) that predict the average diversity that would be observed among the subjects' ratings for a PVS starting from a set of VQMs values computed on such a PVS; iii) give insights into how the same NN based approach can be used to identify potential anomalies in the data collected in subjective experiments.

Keywords Video quality · Subjective testing · Opinions diversity · Neural networks

1 Introduction

There is a growing interest for machine learning (ML) techniques in the research community due to their ability to extract information from data without necessarily making assumptions

✉ Enrico Masala
enrico.masala@polito.it

Extended author information available on the last page of the article.

about a model underlying the data [14]. Depending on the techniques, they can create new predictions on the basis of new input data or provide insights in the observed data. Also the media quality assessment research community leveraged such possibilities proposing several systems that are expected to predict the subjective quality, i.e., the Mean Opinion Score (MOS), of samples starting from a number of different features extracted from the media content by means of algorithms [3, 31]. Other approaches focus on the Quality of Experience (QoE) by studying its correlation with the Quality of Service when the media is transmitted through a network [2, 19].

The research in ML for media quality assessment has been mostly restricted to the quality prediction [4, 5, 8] whereas the problem of predicting the deviation from the MOS, despite being a hot topic within the media quality assessment research community, has benefited only slightly from the success of such an approach [16]. In fact, in a number of recent papers, relying on statistical methods, authors highlighted the inability of the MOS to fully capture all the aspects necessary to measure the perceived quality of a media. In [6], the deviation from the MOS is handled by determining ranges of quality of experience (QoE) rather than a single MOS value. The authors in [24] illustrated the fundamental advantages of using the distribution of opinion scores to assess the quality rather than the MOS, thus underlining the importance of explicitly taking into account the opinions' diversity when assessing the perceived visual quality.

Also the analysis of data coming from subjective experiments has taken limited advantage of ML methods to figure out potential anomalies and thus enhancing the quality of data [1]. Traditional techniques, in fact, focus on using standard statistical approaches (e.g. outlier detection, likelihood estimation, etc.) to deal with the problem of identifying unusual and strange behavior in the data [10, 11, 11–13, 15].

In this work, we focus on the Standard deviation of the Opinions of Subjects (SOS). The SOS is looked at as a measure of observers' opinions diversity. We argue that it is possible to model it as the sum of two components, i.e., i) a deterministic component called ground truth SOS (gtSOS) that can be estimated through the use of neural networks (NNs) by exploiting the disagreement between the objective quality computed by different video quality metrics (VQMs) that are provided as input features to the NN; ii) a random term modeling the two main sources of errors caused by subjective experiments, i.e., the quantization of the rating scale and the limited number of subjects involved in any experiment.

More precisely, we model the diversity in users' opinions by distinguishing between the SOS directly observed in a subjective experiment (with a finite and often very limited number of observers' rating on a discrete scale) and gtSOS, i.e., the standard deviation that would be observed if an infinite or very large number of subjects were asked to assess the quality of the same processed video sequence (PVS) on a continuous scale.

The gtSOS is thus intended to be a measure of how much the intrinsic complexity of a PVS contributes to generate diversity among the subjects' ratings. Complexity is indeed influenced by many factors such as, for instance, the amount of details and motion, as well as potentially different types of distortions in the PVS.

In addition, we designed and trained NNs aiming at predicting the gtSOS. The same NNs can also be exploited to find peculiar behavior not immediately evident in the data of a subjective experiment.

The contributions of this paper is threefold: i) we model the SOS observed for each PVS in a subjective experiment as the sum of the gtSOS and a stochastic normally distributed component modeling the error introduced by the experimental settings; ii) we show that the gtSOS is well correlated to the disagreement of VQM values by considering the Spearman and the Kendall rank correlation on many subjectively annotated datasets; iii) we showcase

the usefulness of the proposed NN based approach to identify potential anomalies in the data collected in subjective experiments.

The paper is organized as follows. The SOS importance in video quality assessment as well as the innovativeness of the work are discussed in Section 2. The proposed SOS model is discussed in Section 3, followed by the Section 4 where we illustrate how, by exploiting NNs and the proposed model, it is possible to highlight potential anomalies in the data collected during a subjective experiment. Section 5 is devoted to the design and training of NNs specific for gtSOS prediction. Conclusions are drawn in Section 6.

2 The SOS in video quality assessment

The SOS has traditionally been used for computing 95% confidence intervals (CIs) for the MOS as follows:

$$CI = MOS \pm \frac{SOS \cdot \tau_{n-1}^{97.5}}{\sqrt{n}} \tag{1}$$

where n is the number of opinion scores from which the MOS is computed and $\tau_{n-1}^{97.5}$ is the 97.5% quantile of a Student's t-distribution with $n-1$ degrees of freedom.

The CIs allow to distinguish between PVSs which are consistently evaluated (those with a small CIs) and the PVSs whose quality is subject to high uncertainty (those with large CIs). According to (1), computing the CI requires the collection of opinions from different observers in order to calculate the SOS and MOS. CIs can therefore only be calculated after carrying out a subjective experiment. This precludes the possibility of being able to use them in the case of, e.g. real-time streaming quality monitoring to automatically determine which PVSs need to be granted more resources in an attempt to reduce the high uncertainty that affects their visual quality.

This problem may be solved using an estimated CI. This would require an estimate not only of the MOS but also of the SOS. Unfortunately, while many advances have been made in estimating the MOS using the features extracted from the PVS, this has not been the case for the SOS. To the best of our knowledge, only in one paper [9] the authors studied the SOS in relation to the MOS, postulating that the SOS is linked to the MOS through a second order polynomial. This postulate is useful for estimating the SOS if and only if the MOS is available. Therefore, it does not solve the problem related to the CI estimation at all. Furthermore, this way of estimating the SOS yields a measure that strongly depends on the context in which the subjective experiment, whose data are used to compute the MOS, was conducted. So, the estimated SOS is therefore no longer a measure of the intrinsic ability of a PVS to confuse observers when evaluating its quality but rather a good metric for analyzing the reliability of the data gathered during a specific subjective experiment.

This work explores, for the first time, the possibility of estimating the observers' diversity of opinions on a given PVS using only features extracted from it, namely the VQMs. More precisely, we highlight the sources of errors that may affect the SOS calculated from the raw data of a subjective experiment. We therefore introduce the gtSOS that can be computed from the PVS' characteristics and therefore represents the level of consensus that would be observed among the opinions collected if the PVS would ideally be evaluated by a very high number of observers. The gtSOS, being an estimate of the SOS but not affected by the errors introduced by any subjective experiment, results in a more stable and reliable measure of the observers' opinions diversity.

Researchers in various scientific fields have introduced a considerable number of sophisticated metrics aiming at measuring the level of diversity or consensus between the opinions of subjects collected in studies based on the Likert scale [26, 30]. However, the media quality assessment research community still did not adopt such metrics. Instead, the SOS remains, until now, the only measure of the observers' diversity of opinions in this case. In many other studies in which opinions are gathered on a Likert scale there is the possibility of re-adjusting the experimental setup or the questionnaire before resubmitting it to the attention of the participants. Moreover, it is possible to iterate in this way until reaching a certain level of consensus among the subjects involved in the study. Unfortunately, this is not the case for media quality assessment, since it is influenced by so many factors even unknown to the persons who design the experiment [23]. This makes the implementation of a consensus-based process difficult in media quality assessment, thus precluding the deployment of the related sophisticated consensus measures. The gtSOS, if interpreted as a measure of consensus in the video quality assessment community, therefore acquires even more significance and importance since it represents a first step towards the development of objective consensus measures within the media quality assessment community.

3 The SOS model

In this section, we investigate and model the SOS observed during actual subjective experiments. To this aim, we propose a model, which introduces the gtSOS that represents a measure of the uncertainty intrinsically associated with the perceived visual quality of a PVS. By predicting such a value, we expect to measure how much reliable would be any estimation of the perceived quality of a PVS. The ability to predict such a value has important practical implications. For instance, to maximize the Quality of Experience (QoE) for final users, it would be better to make sure that the PVSs whose visual quality is difficult to predict consistently receive higher attention, thus ensuring that all users experience a uniform and high satisfaction level [18, 21].

Therefore, in the following we distinguish between the subjectively measured standard deviation observed during a subjective experiment and the gtSOS of the PVS, i.e. the standard deviation that would be observed with an infinite number of observers voting on a continuous (i.e. non-quantized) scale. We will regard the gtSOS value, being experiment-independent, as an intrinsic characteristic of the PVS.

The standard deviation observed directly in subjective tests with a limited number of subjects differs from the gtSOS since it is affected by two main sources of error:

1. The quantization of votes: We observe that, typically, the main focus of a subjective evaluation experiment is to measure the average perceived quality in terms of MOS rather than the spread of opinions in terms of standard deviation [9]. When the standard deviation is needed, it is computed from quantized votes. Consider, for example, that in a five point Absolute Category Rating (ACR) scale experiment, for a given PVS all observers may choose the same score, yielding to an integer MOS value and a computed standard deviation of zero. This actually occurred in experiments even with 24 observers. The VQEG-HD1 [27] can serve as a good example in this context. However, we assume that having a standard deviation equal to zero is induced by the use of a quantized ACR-scale, since it would be really improbable that all observers perfectly agree for a given PVS if a continuous scale is used.

2. The inaccuracy of subjective experiments with a limited number of observers: The statistics of the samples, such as the mean of the samples and their standard deviation are consistent estimators. As the sample size increases, they become more stable and converge to the exact value of the estimated parameters. Unfortunately, subjective experiments are typically conducted with a limited number of subjects. In this case, the standard deviation of the opinions can become, with a not negligible probability, an unstable estimator of the intrinsic ability of the PVS to confuse the viewer in terms of quality perception.

It is worth noting that the aforementioned sources of error are to be taken into account when analyzing the diversity of opinions in any study where the ratings of a finite number of individuals are collected using an ordinal Likert scale. Therefore the approach presented in this work is not limited to the video quality assessment field and can be adopted to analyze the level of consensus among subjects' opinions in other research fields.

Taking these two sources of error into account, we propose to model the measured standard deviation SOS_{exp}^{pvs} of the subjects' opinions for a given PVS observed during a subjective experiment (here indexed by exp) as the sum of two components, i.e. a deterministic component $gtSOS^{pvs} \in [0, \infty)$ intrinsic to the PVS itself and, a non-predictable, stochastic and normally distributed component $D_{exp} = (err_{exp}^{quant} + err_{exp}^{subj})$ dependent on the experimental settings, i.e. the effect of quantization (err_{exp}^{quant}) and the use of a finite number of subjects (err_{exp}^{subj}). This leads to the following SOS model:

$$SOS_{exp}^{pvs} = gtSOS^{pvs} + D_{exp}^{pvs}. \tag{2}$$

In order to model the systematic component of the standard deviation ($gtSOS^{pvs}$), we investigate the possibility of exploiting the disagreement of objective metrics computed on the PVS. In fact, since different VQMs are designed to take into account different aspects of the human visual system, we expect that there could be artifacts, which a certain VQM might be very sensitive to, while others less so, similarly as human observers are too.

To confirm such intuition, i.e. the existence of a significant link between the disagreement of objective VQMs and the ability of a PVS to induce diversity among observers' opinions, we conducted a statistical analysis aiming to verify whether greater diversity of opinions is observed in the presence of greater disagreement of VQM values.

The analysis is conducted on five subjectively annotated datasets, i.e., the ITS4S dataset [7, 22], the Netflix public dataset [15] and three datasets released by the Video Quality Expert Group (VQEG): the VQEG-HD1 [27], VQEG-HD3 [27], and VQEG-HD5 [27]. Sample images taken from SRCs in those datasets are shown in Fig. 1. On each dataset, we sort the sequences in ascending order of SOS. Then we measure, by means of the Spearman Rank Order Correlation Coefficient (SROCC) and the Kendall Rank Order Correlation Coefficient (KROCC), the agreement of three VQMs, i.e., the Peak Signal to Noise Ratio (PSNR), the Structural Similarity Image (SSIM) [29] and the Visual Information Fidelity (VIF) [25], on 50 sequences having recorded the lowest SOS as well as on 50 ones with the highest SOS. Unlike the ITS4S and the Netflix public dataset in which all the PVSs are affected only by coding distortion, the three VQEG datasets involved in our analysis consider both coding and transmission distortion. Therefore, for these datasets, our analysis was also made on the basis of the type of distortion in order to reach a more precise conclusion. The results are shown in Figs. 2 and 3 for the SROCC and KROCC respectively. It can be noted that in all the cases in which the PVSs are only affected by coding distortion, the VQM values show greater correlation on the set of sequences with the less diversification

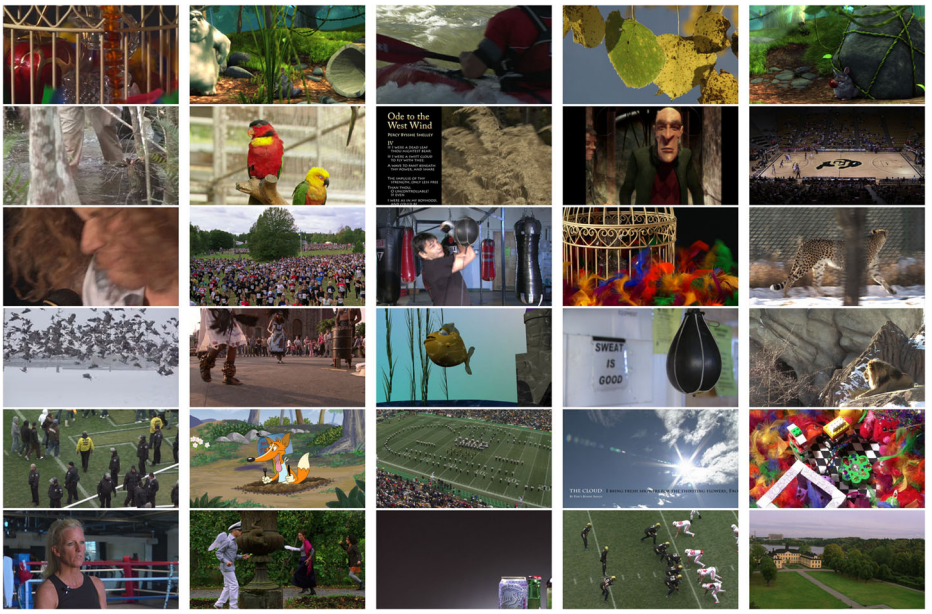


Fig. 1 Sample images, one dataset per column. From left to right: ITS4S, Netflix Public, VQEG HD1, HD3, HD5

of opinions (low SOS). This greater correlation of VQMs in presence of greater agreement between human observers is not clearly observed in the case of PVSs whose quality is corrupted by transmission artifacts. We believe that this behavior can be explained by the

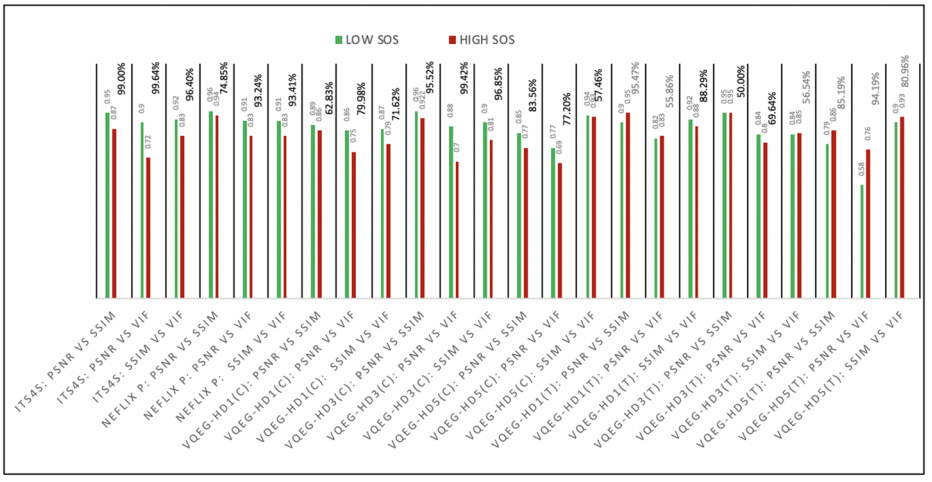


Fig. 2 Correlation coefficient (Spearman rank order) between pairs of VQMs (PSNR, SSIM, VIF), in a given subjective experiment (the ITS4S, Netflix public dataset, VQEG-HD1, HD3 and HD5), when the PVSs with low (green) or high (red) SOS are considered. The statistical significance of the difference is indicated in percentage. For PVSs affected by coding (C) distortions, low SOS always implies higher VQM correlation. For transmission (T) distortions this is not always the case (percentage in grey)

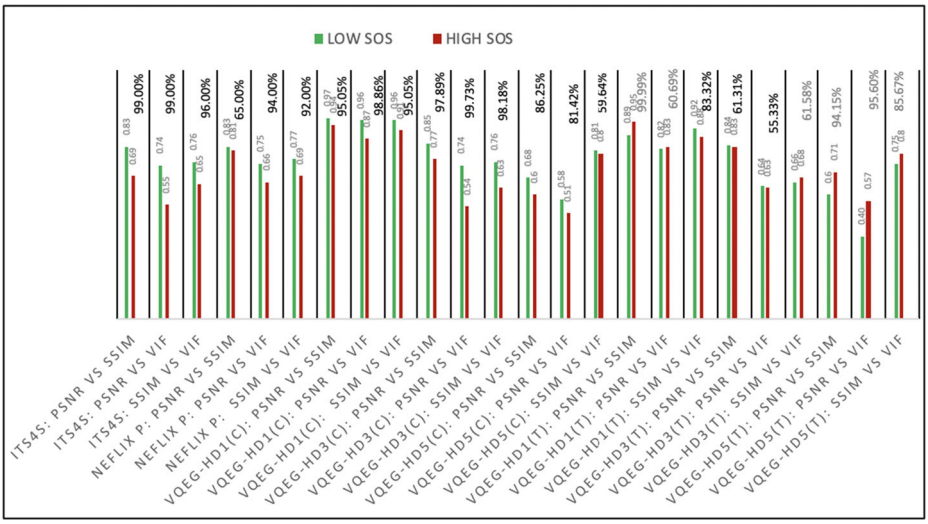


Fig. 3 Correlation coefficient (Kendall rank order) between pairs of VQMs (PSNR, SSIM, VIF), in a given subjective experiment (the ITS4S, Netflix public dataset, VQEG-HD1, HD3 and HD5), when the PVSs with low (green) or high (red) SOS are considered. The statistical significance of the difference is indicated in percentage. For PVSs affected by coding (C) distortions, low SOS always implies higher VQM correlation. For transmission (T) distortions this is not always the case (percentage in grey)

fact that the considered VQMs have empirically demonstrated poor accuracy in adequately capturing quality impairment due to transmission.

The purpose of the analysis is to verify whether the correlation of VQMs when the SOS is small is greater than the one observed when the SOS is high. To make sure that our conclusions are independent of the particular datasets that we are using, when a correlation is greater than another we verify whether the difference is significant or just an artifact of the data or, equivalently, only a matter of chance. For this reason, we performed statistical tests to determine how confident one should be in stating that a certain correlation value is greater than another. The percentages in Figs. 2 and 3 show these confidence levels for each pair of correlations under comparison.

For instance, for the ITS4S dataset, the correlation between the PSNR and the SSIM on PVSs with low SOS can be considered greater than the one observed in presence of high SOS with 99% of confidence. Hence the difference between the two values cannot reasonably be considered as a result of chance. Similar high values of confidence are observed among all other pairs of VQMs for the ITS4S, the VQEG-HD3 when limiting the analysis to PVSs with coding distortion and finally also for the Netflix public dataset. On the other hand, in the case of PVSs affected by coding distortion in the VQEG-HD1 and HD5 datasets, although the correlation coefficients between the VQMs observed in the presence of low SOS are larger than those observed in correspondence of high SOS, the percentages of confidence are less than 95%.

In summary, the analysis reveals that the degree of agreement between the PSNR, SSIM and the VIF, measured through the SROCC and the KROCC, is generally greater when calculated on PVSs affected by coding distortion for which observers have expressed opinions characterized by little diversity. On the other hand, for the PVSs affected by transmission distortion this preliminary analysis does not allow such a conclusion. However, this does not

preclude the existence of a more sophisticated agreement measure than the SROCC and the KROCC between the VQMs that may explain the diversity of the observers' opinions when rating PVSs whose quality is impaired by transmission artifacts. Such a measure could be found by fitting the VQMs to the SOS using a highly nonlinear function as done later in the work.

In light of the previous discussion, we formulate the following hypothesis: the $gtSOS^{pvs}$ can be estimated from the values of a number of objective measures computed on the PVS, on the basis of their disagreement. In this work, the PSNR, the SSIM, the VIF [25], the Multi-Scale Structural Similarity Image (MS-SSIM) [28], and the Video Multimethod Assessment Fusion (VMAF) [20] are considered. Hence

$$gtSOS^{pvs} = f(PSNR, SSIM, VIF, MSSSIM, VMAF) + \epsilon_{obj}^{pvs} \tag{3}$$

where ϵ_{obj}^{pvs} is an error term modeling the inability of completely predicting $gtSOS^{pvs}$ by only considering the values of a set of objective measures as features, and f a function mapping the information related to the objective metrics' disagreement to the gtSOS. The estimation of the function f will be discussed in the next section.

In summary, our analysis argues that the SOS_{exp}^{pvs} observed for any PVS during a subjective experiment is a realzation of a normally distributed random variable due to the D_{exp}^{pvs} component, and has a mean $gtSOS^{pvs}$ that we propose to estimate by exploiting the disagreement between the different VQMs modelling the characteristics of the sequence. Further insights into the validity of such a statement will be given in Section 4.

4 SOS model validation and anomalies detection in subjective experiments

In this section, we estimate the function f in (3) through NNs, then we investigate the validity of the model proposed in (2) through various numerical experiments and finally we illustrate the capability of the whole system to highlight potential anomalies in the data collected during a subjective experiment.

4.1 SOS model validation

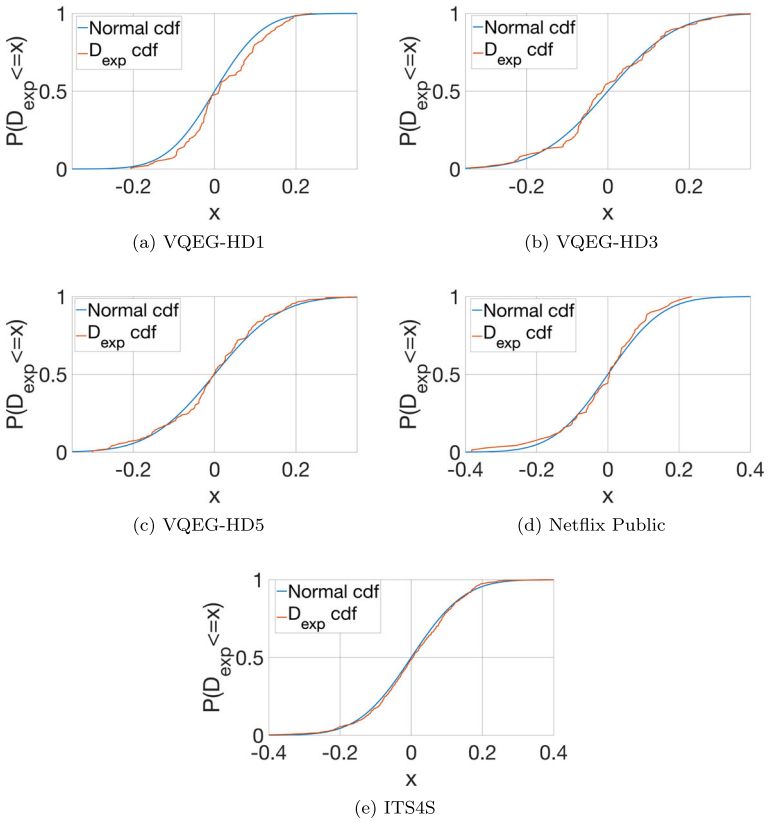
To validate the SOS model in (2) as well as the ability of VQMs to capture diversity among observers' ratings, an approximation of the function f is needed. This can be done by fitting the VQMs to the SOS observed during a subjective experiment, using any ML algorithm tailored for regression. An impressive number of ML algorithms has been proposed in literature, however NN based models and support vector regression (SVR) have empirically demonstrated greater accuracy in the field of media quality assessment. To estimate the function f we therefore naturally evaluated both NN as well as SVR based models. However, we have experimentally observed that NNs, for the task of interest, lead to a prediction of the gtSOS which correlates better with the SOS when cross validating the obtained models. We rely therefore on a NN to approximate the function f. The NN is trained using the five aforementioned VQMs, as an input, and the target is SOS_{exp}^{pvs}. However, on the basis of the model in (2) and the assumption in (3), the stochastic component D_{exp} is not predictable, and from the disagreement of the values of the objective metrics that the NN receives as input, interesting information can be gained only for the prediction of the deterministic component of the SOS. Therefore, we can assimilate the result of the NN prediction to the $gtSOS$.

Since subjective experiments are expensive and time consuming, it is very difficult or probably even impossible to find datasets that contain reliable subjective evaluations for a very high number of PVSs. This precludes the possibility of using on these datasets, deep NNs, i.e., NNs with more than one hidden layer, or even single hidden layer NNs with a large number of neurons on the hidden layer. In fact the high number of parameters and consequently the number of degrees of freedom of these NNs would lead to overfitting the dataset. In the context of this study, overfitting would yield an estimate of the gtSOS affected by the peculiarities of the specific subjective experiment which is reflected in the data used for training. Such estimate of the gtSOS would therefore no longer be an intrinsic characteristic of the PVS since it suffers from two sources of error due to subjective experiment settings, i.e., scale quantization and limited number of observers, as previously discussed. To overcome this problem, in Section 5 we will adopt a data augmentation approach. More precisely, we will generate more data artificially from the ones actually collected during a subjective experiment in order to be able to use a deep NN. Given that the focus of this Section is to validate the model in (2) for each subjective experiment involved in our study, we simply investigated several single hidden layer NNs with few neurons on the hidden layer, to determine the structure that would work best to estimate f without already generating other data that could biased the accuracy of the proposed model in representing the SOS values actually observed during a subjective experiment.

We experimentally found that f can be effectively approximated by a NN with 5 neurons on the input layer, i.e. one for each VQM, a single hidden layer with 4 neurons and finally an output layer with one neuron delivering the gtSOS estimation.

In order to validate the model in (2), we estimate the function f on five different annotated datasets, i.e. the VQEG-HD1, VQEG-HD3, VQEG-HD5, Netflix public and ITS4S dataset. Once the function f is known, it is possible to i) estimate the value of $gtSOS^{pvs}$ for each PVS, thus identifying contents whose quality is intrinsically difficult to assess consistently (i.e., high $gtSOS^{pvs}$); ii) deduce from (2) the value of the stochastic component D_{exp} for each PVS. From the set of D_{exp} values, we estimate the empirical cumulative distribution of D_{exp} that we then compare with the cumulative distribution of a Gaussian random variable with zero mean and standard deviation equal to the one derived from the set of D_{exp} values. The results are shown in Fig. 4. In all the cases, the empirical cumulative distribution of D_{exp} seems to be very well approximated by a Normal cumulative distribution. This is coherent with the proposed SOS model. Figures 5, 6, 7, 8 and 9 report the comparison between the predicted gtSOS and the SOS for all the aforementioned datasets. On the various training sets, i.e., when training the NN using all the data in the dataset, the obtained PLCC values range from 0.30, in the worst case, up to 0.82, whereas in cross validation the observed PLCC values range from 0.29 to 0.77. However, the SROCC values are somewhat lower. In fact, on the various training sets they range from 0.24 to 0.69, and in cross validation from 0.23 to 0.62. This difference with respect to the PLCC values is an artifact of the quantization of the scale on which the subjective tests are conducted. In fact, the computation of the SOS value on ordinal data increases the probability of getting ties, the presence of which typically leads to an underestimation of the SROCC.

We performed statistical tests aiming at verifying whether the PLCC and SROCC values in the aforementioned ranges can be considered statistically different than zero with 95% of confidence while taking into account the size of each dataset i.e., the number of PVSs evaluated in the dataset. In all cases, the test result revealed that the obtained PLCC and SROCC values can be considered greater than zero with statistical significance. Therefore, the hypothesis that it is possible to obtain information about the diversity observed in the



(a) VQEG-HD1 (b) VQEG-HD3

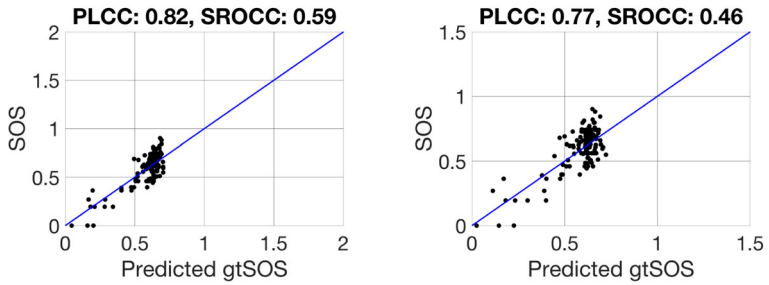
(c) VQEG-HD5 (d) Netflix Public

(e) ITS4S

Fig. 4 Empirical and estimated cumulative density function of D_{exp} for several datasets

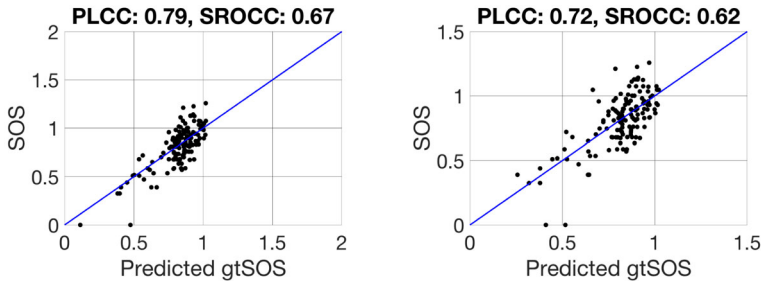
opinions expressed by different observers about the visual quality of a PVS using only some VQMs calculated on that PVS cannot be rejected.

We notice that lower PLCC and SROCC values have been observed in the case of the ITS4S dataset in comparison to those obtained on the other datasets. We attribute this behavior to the fact that, unlike the other subjective experiments considered in this work, the one of the ITS4S was designed for the development of no-reference metrics. Therefore, during



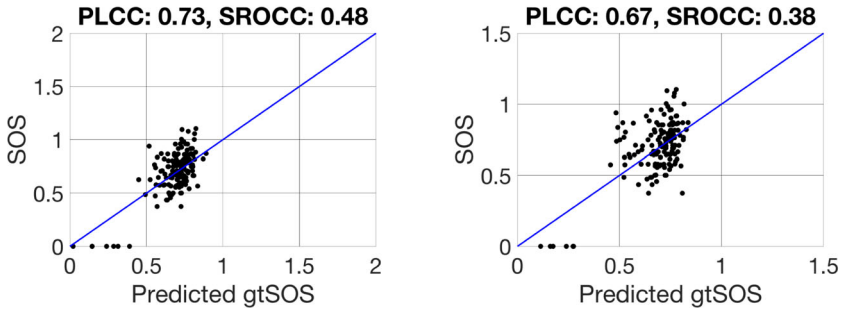
(a) Training set (b) 5-fold cross validation

Fig. 5 VQEG-HD1 dataset: gtSOS prediction from the VQM values for each PVS



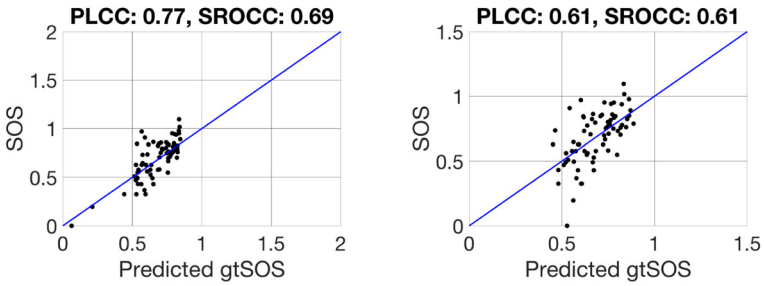
(a) Training set (b) 5-fold cross validation

Fig. 6 VQEG-HD3 dataset: gtSOS prediction from the VQM values for each PVS



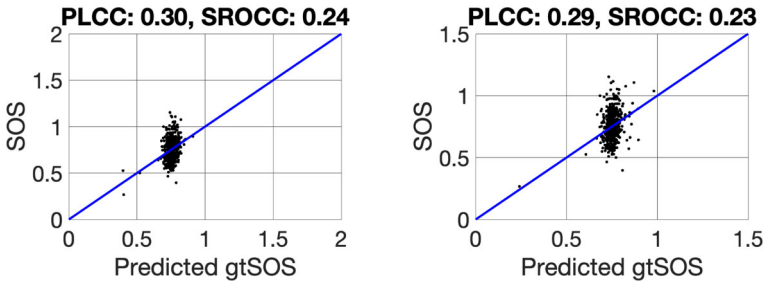
(a) Training set (b) 5 folds cross validation

Fig. 7 VQEG-HD5 dataset: gtSOS prediction from the VQM values for each PVS



(a) Training set (b) 5 folds cross validation

Fig. 8 Netflix Public dataset: gtSOS prediction from the VQM values for each PVS



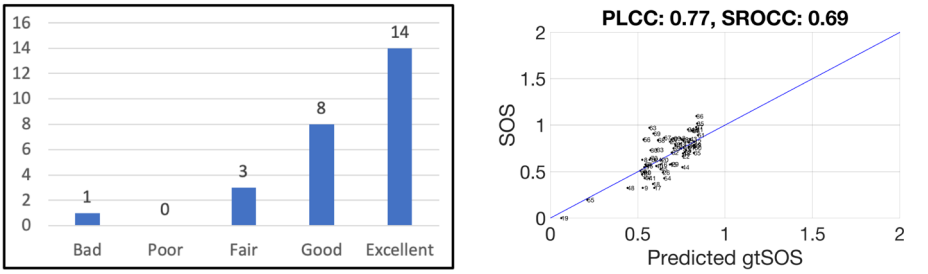
(a) Training set (b) 5 folds cross validation

Fig. 9 ITS4S dataset: gtSOS prediction from the VQM values for each PVS

the experiment, the source (SRC), i.e. the original content, was never shown to the observers. Hence, the full reference VQMs considered in this study did not allow to obtain as much information on the diversity between the opinions of the observer as in the other cases. Nevertheless, the obtained PLCC and SROCC can be considered significant with 95% of confidence.

4.2 Anomalies detection

In literature, some studies [13] addressed the issue of identifying potential anomalies in a subjective experiment due to the presence of peculiar contents or subject behavior. For instance, an observer may just assign random votes or the grading of a specific sequence may be remarkably inconsistent. The presence of such anomalies may negatively affect the accuracy of objective measures developed, relying on raw data collected during subjective experiments. The typical approach adopted for anomalies detection is to model the observer opinion on each sequence using the normal distribution [11, 13, 15] and then estimate the related parameters to identify unexpected situations. While using the normal distribution is very convenient from the theoretical point of view, in practice the use of such a distribution may not always be the best option. For instance, the normal distribution can not effectively model the opinions' distribution for PVSs with very high or very low perceived visual quality as illustrated in Fig. 10a, which shows the score distribution for a specific PVS in the Netflix dataset.



(a) Subjective scores distribution for PVS #63 (b) Predicted gtSOS, with labelled PVSs.

Fig. 10 Analysis done on the Netflix Public dataset. PVS #63 is far from the 45-degree line. Inspecting the score distribution for PVS #63 reveals that an opinion score equal to 1 seems to be anomalous (left)

In this work, we approach the problem differently. Our analysis is based on the proposed SOS model described by (2). The term D_{exp} in the model intends to represent the part of the inconsistency in the votes introduced by the experimental settings. As such, it also models the average inconsistency of the sample of people chosen for the experiment. Therefore, such an estimate allows to determine the sequences for which a high inconsistency of the votes has been observed and also those for which, due the quantization of the scale, the observed SOS is less than that, which could have been observed considering a greater number of subjects in the subjective experiment voting on a continuous scale.

Our procedure to find potential anomalies can be summarized as follows. Starting from the data of the subjective experiment under examination, we estimate the function f as discussed before, then from (2) and (3) we obtain, for each PVS, the following estimate:

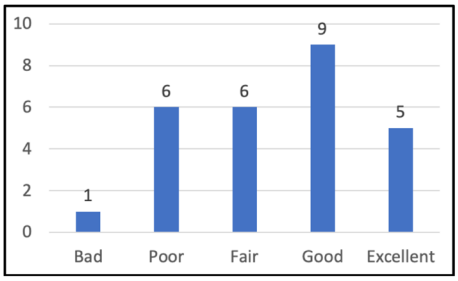
$$D_{exp} \approx SOS_{exp} - f(PSNR, SSIM, VIF, MSSSIM, VMAF) \tag{4}$$

We thus obtain a set of values having a normal distribution with zero mean as indicated by the model in (2). The PVSs, whose evaluation we believe may be affected by anomalies, are those for which the estimated D_{exp} value is an outlier of this distribution. In practice, denoting with D_{exp}^{pvs} the value of D_{exp} for a given PVS and by $std_{D_{exp}}$ the standard deviation of D_{exp}, we suggest to give a closer look to the ratings of each PVS for which:

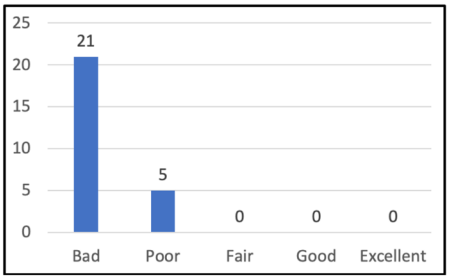
$$\left|D_{exp}^{pvs}\right| > 3 \cdot std_{D_{exp}} \tag{5}$$

and carefully consider an examination of such anomalies before using the data.

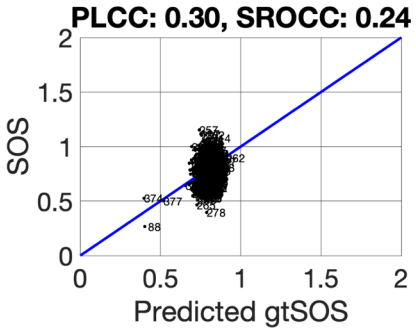
In order to investigate the effectiveness of the method in practice, we tested it on the Netflix public dataset and the ITS4S dataset. In Fig. 10, we report again the comparison between the predicted gtSOS and the SOS after determining the function f on the Netflix public dataset. We labeled the PVSs to facilitate the interpretation of the results. For any PVS, D_{exp}^{pvs} is estimated by subtracting the predicted $gtSOS^{pvs}$ from the SOS_{exp}^{pvs}. Consider, for instance, PVS #63 for which the condition in (5) holds. The ratings collected in the subjective experiments are shown in Fig. 10a. For such a PVS, even if the mode of the distribution of the subjects' opinion is equal to 5 ("Excellent") and 22 observers out of 26 rank the quality of the PVS at least 4, i.e. "Good", there is surprisingly an observer ranking it as 1, i.e. "Bad". It is therefore reasonable to be skeptical about the latter rating. This is indeed more curious when we notice that there are even sequences, such as PVS #19, where the previous anomalous observer is in a full agreement with all the observers. In the case of the ITS4S dataset shown in Fig. 11, we analyzed the scores collected for PVS #257 and #278 that exhibit a high value of $|D_{exp}|$. We notice that the individual subjects' ratings for PVS #257 (shown in Fig. 11a) are almost uniformly distributed between "Poor" and "Excellent" leading to an observed SOS value, which is significantly larger than the predicted gtSOS that would suggest that the intrinsic difficulty of evaluating the PVS should be lower. Therefore, the PVS content characteristics should be investigated in more details. On the contrary, for PVS #278 (shown in Fig. 11b), a low value of the SOS is observed since 21 observers rated its perceived visual quality as 1 ("Bad") and 5 observers rated it as 2 ("Poor"). However the analysis indicates that the observed SOS underestimates the gtSOS and thus the intrinsic capacity of such PVS to confuse the observer in terms of quality perception. This suggests that higher diversity among the opinions should be expected in case more ratings are gathered. This is therefore another interesting case for further investigation. For instance, such a PVS could be reevaluated asking many observers to vote on a continue scale in order to make sure that the low SOS value previously observed is not just due to the scale quantization effect and the use of a limited number of observers.



(a) Subjective scores distribution for PVS #257



(b) Subjective scores distribution for PVS #278



(c) Predicted gtSOS, with labelled PVSs

Fig. 11 Analysis done on the ITS4S dataset. PVS #257 and #278 are far from the 45-degree line. Inspecting the score distribution for PVS #257 reveals a close-to-uniform distribution, while the analysis suggested that the observed low SOS for PVS #278 may not reliably represent its intrinsic ability to confuse observers in terms of quality perception

5 Deep neural network based model for gtSOS prediction

In this section, we try to design and train a NN that can be used to predict the gtSOS in general, not limiting the analysis to a single subjective experiment, as done in the previous section in order to validate the model in (2). The aim is therefore to train a model that can provide hints about the uncertainty that characterizes the perceived visual quality of a PVS. In order to train the model we choose, as a training set, the data collected during the VQEG-HD1 and VQEG-HD5 experiments, restricting the analysis to PVSs affected only by the coding distortions, since we employ a set of VQMs which have traditionally shown higher accuracy in assessing the quality of PVSs corrupted by this type of distortion only.

In our work, we intend the gtSOS to be a characteristic of a PVS. Therefore, its estimate must be detached from the influence of subjective experiments used to compute it. To this aim, we introduce a stochastic component in order not only to inform the whole process about the fact that the subjective data available for the training process represents only one of many possible scenarios but also to derive a probabilistic model useful for data augmentation. This then allows us to effectively use a deep NN for gtSOS prediction. More precisely, each data point in our training dataset is considered to be a sample of the following 6-dimensional random vector: $(PSNR, SSIM, VIF, MSSSIM, VMAF, SOS)$. This is in line with the model in (2) that explicitly considers the SOS for each PVS s as a random

variable. This, coupled with the variability of VQM values for the same subjective quality, suggest that data points available in our training dataset can be considered as realizations of a 6-dimensional random vector. On the basis of this observation, we attempt to derive the multivariate distribution from which additional data can be generated to better train the NN. In particular, we aim at reducing the influence of the settings of the subjective experiments chosen for the training.

We propose to model such multivariate distribution using a 6-dimensional Gaussian Mixture Model (GMM), i.e.

$$g(VQM_{pvs}, SOS_{exp}^{pvs}) = \sum_{i=1}^{k} \pi_i \cdot N\Big((VQMs, SOS_{exp}^{pvs})|\mu_i, \Sigma_i\Big) \tag{6}$$

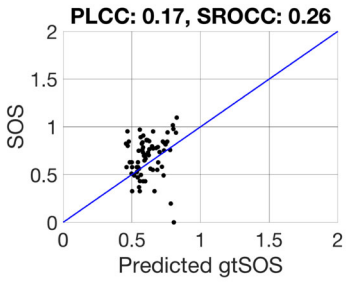
where $VQMs = (PSNR, SSIM, VIF, MSSSIM, VMAF)$, $N\left(VQMs, SOS_{exp}^{pvs}|\mu_i, \Sigma_i\right)$ is a probability density function of a multivariate normal distribution with mean μ_i and covariance matrix Σ_i and k is a number of components of the GMM. The parameters (π_i, μ_i, Σ_i and k) of the GMM are estimated using a maximum likelihood estimation. Denoting by M the number of PVSs in the training set, we solve the following optimization problem

$$(\pi_i, \mu_i, \Sigma_i, k) = \arg\max \left(\prod_{s=1}^{M} \left(\sum_{i=1}^{k} \pi_i \cdot N\Big((VQMs_{pvs}, SOS_{exp}^{pvs})|\mu_i, \Sigma_i\Big)\right)\right) \tag{7}$$

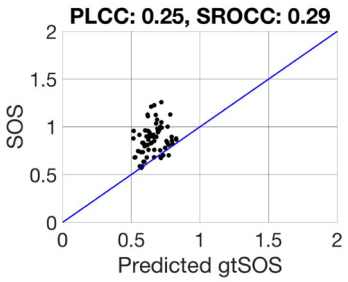
where $VQMs_{pvs}$ are values of the objective measures computed on a PVS. The problem in (7) is solved by using the Expectation-Maximization algorithm (EMA). More details about the EMA can be found in [17].

Once the parameters of the GMM are obtained, to augment the data for the training process, we simulate more data points by the GMM. This approach allowed us to exploit the prediction capability of deep NNs that would otherwise have led to overfitting if the deep NNs would be trained on the initially available limited size datasets. We performed extensive numerical experiments to determine the NN architecture that best fits our need. The best results were obtained using a NN with 5 neurons on the input layer, i.e. one for each objective measure, three hidden layers having 11, 17 and 5 neurons respectively and finally an output layer with 1 neuron that provides the desired estimation of the gtSOS value.

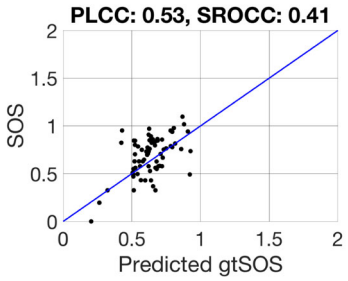
To evaluate the effectiveness of the deep NN based model, we tested it on the Netflix public dataset and the VQEG-HD3 dataset that have not been considered during the training process. The results are shown in Fig. 12 (bottom part). On the Netflix public dataset, the gtSOS predicted by the trained deep NN, when compared to the SOS, yielded a PLCC of 0.5 and a SROCC of 0.41. While on the VQEG-HD3 dataset the PLCC and the SRCC between the predicted gtSOS and the actual SOS reached respectively 0.48 and 0.44. Although these values were tested to be greater than zero with 95% of confidence, they are lower than those reported previously when training and cross validating small networks on the data collected during a single subjective experiment. However, we are confident that the accuracy of this model can be further improved if it would be possible to use data from a subjective experiment designed specifically to create a good predictor for the gtSOS value. This is not the case for typical subjective experiments that are designed to cover, as uniformly as possible, the quality scale in terms of MOS of the chosen PVSs, but often do not take into account what could be the SOS for each PVS. However, in order to effectively train machine learning algorithms for gtSOS prediction, a sufficiently uniform coverage of the



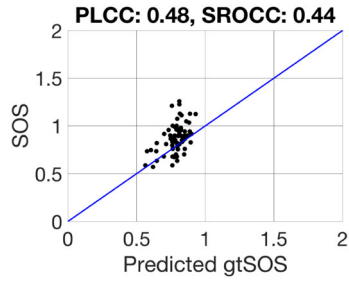
(a) Netflix public dataset, no data augmentation



(b) VQEG-HD3, no data augmentation



(c) Netflix public dataset, with data augmentation



(d) VQEG-HD3, with data augmentation

Fig. 12 Accuracy of the deep NN based model when predicting the gtSOS without (top) and with (bottom) the data augmentation. In all the cases, the NN is trained on the VQEG-HD1 and VQEG-HD5 (coding artifacts only)

SOS range is required to avoid models that need to extrapolate the results for certain conditions. Therefore, it is necessary to design a subjective experiment with this aim in mind since the beginning.

Finally, to evaluate the effectiveness of our data augmentation approach, i.e. the simulation of more training data points by the GMM, we trained a shallow NN, having the structure presented in Section 4, on a training set composed of the VQEG-HD1 and VQEG-HD5 experiments without the data augmentation, i.e., simulating more training data from the GMM. Testing this NN on the Netflix public dataset and VQEG-HD3 dataset yields the results shown in Fig. 12 (top part). The much lower PLCC values ($0.17 < 0.53$, $0.25 < 0.48$) as well lower SROCC values ($0.26 < 0.41$, $0.29 < 0.44$) compared to those reported in the bottom part of the corresponding picture show the strong need for data augmentation as well as its effectiveness. This further confirms our belief that gathering enough data during a subjective experiment specifically designed for gtSOS modeling would potentially improve the performance obtained in this study.

6 Conclusions

In this work, we showed how machine learning techniques and neural networks in particular can be a helpful tool in analyzing the details of subjective experiments. Neural networks, typically used in the literature to predict only the mean subjective quality, can also be a helpful tool in analyzing the data coming from subjective experiments in order to identify, for

instance, anomalies or behavior, which are not immediately found by using the traditional analysis approaches. Our analysis focus on analyzing and modeling the diversity observed among the subjects' opinions in subjective experiments. In particular, we model the standard deviation of the ratings of different observers on single PVSs, arguing that it is distributed according to a normal distribution whose mean, referred to as the ground truth SOS in this work, can be effectively estimated by exploiting the value of a set of VQMs computed on the PVS. Relying on this model, we showed that it is possible to identify PVSs that might present anomalies when the subjects' scores are considered together with their variance. The identified cases can then be manually analyzed to better investigate potential causes. Moreover, we also showed that it is possible to train neural networks that, taking VQMs values as an input, can predict how much diversity would be observed among subjects' votes if a PVS would be subjectively evaluated. While training and cross validating the neural network on the same subjective experiment, we showed that the prediction is significantly correlated with the standard deviation observed in the actual subjective experiment. Finally, by applying a data augmentation approach, we trained a deep neural network that is supposed to predict the ground truth standard deviation of any PVS affected by compression artifacts after receiving, as an input, only the VQMs values computed on that PVS. This deep neural network provided correlations equal to about 0.5. Although this correlation is demonstrated to be statistically significantly different from zero with 95% of confidence, it still remains somehow low. However, the approach looks promising. Therefore, future research activities will be devoted to conduct a subjective experiment intentionally designed to collect data in a way that will improve the accuracy of such a deep neural network model.

Acknowledgements Part of this work has been supported by PIC4SeR (https://pic4ser.polito.it).

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Aldahdooh A, Masala E, Van Wallendael G, Lambert P, Barkowsky M (2019) Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures. Signal Process Image Commun 74:32–41
2. Aroussi S, Mellouk A (2014) Survey on machine learning-based QoE-QoS correlation models. In: 2014 International conference on computing, management and telecommunications (commantel). IEEE, pp 200–204
3. Bhattacharya A, Palit S (2020) Measurement of image degradation: a no-reference approach. Multimed Tools Appl 79(9):5545–5572

4. Bosse S, Maniry D, Müller K, Wiegand T, Samek W (2018) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219. https://doi.org/10.1109/TIP.2017.2760518
5. Ding Y, Zhao Y, Zhao X (2017) Image quality assessment based on multi-feature extraction and synthesis with support vector regression. Signal Process Image Commun 54:81–92
6. Fotio Tiotsop L, Masala E, Aldahdooh A, Van Wallendael G, Barkowsky M (2019) Computing quality-of-experience ranges for video quality estimation. In: 2019 Eleventh international conference on quality of multimedia experience (qoMEX), pp 1–3. https://doi.org/10.1109/QoMEX.2019.8743303
7. Fotio Tiotsop L, Servetti A, Masala E (2020) Full reference video quality measure improvement using neural networks. In: Proc. Intl. Conf. acoustics, Speech, and Signal Processing (ICASSP)
8. Freitas PG, Akamine WY, Farias MC (2018) Using multiple spatio-temporal features to estimate video quality. Signal Process Image Commun 64:1–10
9. Hoßfeld T, Schatz R, Egger-Lampl S (2011) SOS: The MOS is not enough!. In: 2011 Third international workshop on quality of multimedia experience, pp 131–136. https://doi.org/10.1109/QoMEX.2011.6065690
10. ITU-T Rec. BT.500: Methodology for the subjective assessment of the quality of television pictures (2012)
11. ITU-T Rec. P.1401: methods, metrics and procedures for statistical evaluation qualification and comparison of objective quality prediction models (2012)
12. ITU-T Rec. P.910: Subjective video quality assessment methods for multimedia applications (1996)
13. Janowski L, Pinson M (2015) The accuracy of subjects in a quality experiment: a theoretical subject model. IEEE Trans Multimed 17(12):2210–2224
14. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260
15. Li Z, Bampis CG (2017) Recover subjective quality scores from noisy measurements. In: 2017 Data compression conference (DCC), pp 52–61. https://doi.org/10.1109/DCC.2017.26
16. Mocanu DC, Pokhrel J, Garella JP, Seppänen J, Liotou E, Narwaria M (2015) No-reference video quality measurement: added value of machine learning. J Electron Imaging 24(6):1–15. https://doi.org/10.1117/1.JEI.24.6.061208
17. Moon TK (1996) The expectation-maximization algorithm. IEEE Signal Process Mag 13(6):47–60
18. Mu M, Romaniak P, Mauthe A, Leszczuk M, Janowski L, Cerqueira E (2012) Framework for the integrated video quality assessment. Multimed Tools Appl 61(3):787–817
19. Mushtaq MS, Augustin B, Mellouk A (2012) Empirical study based on machine learning approach to assess the QoS/QoE correlation. In: 2012 17th european conference on networks and optical communications. IEEE, pp 1–7
20. Netflix: VMAF - video multi-method assessment fusion v.0.6.2. https://github.com/Netflix/vmaf (2018)
21. Paudyal P, Battisti F, Carli M (2016) Impact of video content and transmission impairments on quality of experience. Multimed Tools Appl 75(23):16461–16485
22. Pinson MH (2018) A video quality dataset with four-second unrepeated scenes. NTIA Technical Memo TM-18-532
23. Reiter U, Brunnström K, De Moor K, Larabi MC, Pereira M, Pinheiro A, You J, Zgank A (2014) Factors Influencing Quality of Experience. Springer International Publishing, Cham, pp 55–72
24. Seufert M (2019) Fundamental advantages of considering quality of experience distributions over mean opinion scores. In: 2019 Eleventh international conference on quality of multimedia experience (QoMEX), pp 1–6. https://doi.org/10.1109/QoMEX.2019.8743296
25. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
26. von der Gracht HA (2012) Consensus measurement in delphi studies: Review and implications for future quality assurance. Technol Forecast Soc Chang 79(8):1525–1536
27. VQEG: Report on the validation of video quality models for high definition video content (v. 2.0). http://bit.ly/2Z7GWDI (2010)
28. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems computers, vol 2, pp 1398–1402. https://doi.org/10.1109/ACSSC.2003.1292216
29. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
30. Wierman MJ, Tastle WJ (2005) Consensus and dissention: theory and properties. In: NAFIPS 2005 - 2005 Annual meeting of the north american fuzzy information processing society, pp 75–79
31. Xu L, Lin W, Kuo CCJ (2015) Visual quality assessment by machine learning. Springer, Berlin

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Lohic Fotio Tiotsop[1] · Tomas Mizdos[2] · Miroslav Uhrina[2] · Marcus Barkowsky[3] · Peter Pocta[2] · Enrico Masala[1]

Lohic Fotio Tiotsop
lohic.fotiotiotsop@polito.it

Tomas Mizdos
tomas.mizdos@feit.uniza.sk

Miroslav Uhrina
miroslav.uhrina@feit.uniza.sk

Marcus Barkowsky
marcus.barkowsky@th-deg.de

Peter Pocta
peter.pocta@feit.uniza.sk

1 Politecnico di Torino, Torino, Italy

2 University of Zilina, Zilina, Slovakia

3 Deggendorf Institute of Technology (DIT), Deggendorf, Germany