Mimicking individual media quality perception with neural network based artificial observers

Mimicking individual media quality perception with neural network based artificial observers

LOHIC FOTIO TIOTSOP, Politecnico di Torino, Italy
TOMAS MIZDOS, University of Zilina, Slovakia
MARCUS BARKOWSKY, Deggendorf Institute of Technology, University of Applied Sciences, Germany
PETER POCTA, University of Zilina, Slovakia
ANTONIO SERVETTI, Politecnico di Torino, Italy
ENRICO MASALA, Politecnico di Torino, Italy

The media quality assessment research community has traditionally been focusing on developing objective algorithms to predict the result of a typical subjective experiment in terms of Mean Opinion Score (MOS) value. However, the MOS, being a single value, is insufficient to model the complexity and diversity of human opinions encountered in an actual subjective experiment. In this work we propose a complementary approach for objective media quality assessment that attempts to more closely model what happens in a subjective experiment in terms of single observers and at the same time we perform a qualitative analysis of the proposed approach while highlighting its suitability. More precisely, we propose to model, using neural networks (NNs), the way single observers perceive media quality. Once trained, these NNs, one for each observer, are expected to mimic the corresponding observer in terms of quality perception. Then, similarly to a subjective experiment, such NNs can be used to simulate the users' single opinions, which can be later aggregated by means of different statistical indicators such as average, standard deviation, quantiles, etc. Unlike previous approaches that consider subjective experiments as a black box providing reliable ground truth data for training, the proposed approach is able to consider human factors by analyzing and weighting individual observers. Such a model may therefore implicitly account for users' expectations and tendencies, that have been shown in many studies to significantly correlate with visual quality perception. Furthermore, our proposal also introduces and investigates an index measuring how much inconsistent would be an observer if asked to rate many times the same stimulus over time. Simulation experiments conducted on several datasets demonstrate that the proposed approach can be effectively implemented in practice and thus yielding a more complete objective assessment of end users' quality of experience.

CCS Concepts: • **Information systems → Multimedia information systems**.

Additional Key Words and Phrases: media quality, human factors, user expectations, subjects opinions, neural networks

ACM Reference Format:
Lohic Fotio Tiotsop, Tomas Mizdos, Marcus Barkowsky, Peter Pocta, Antonio Servetti, and Enrico Masala. 2021. Mimicking individual media quality perception with neural network based artificial observers. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2021), 25 pages. https://doi.org/10.1145/3464393

Authors' addresses: Lohic Fotio Tiotsop, Politecnico di Torino, corso Duca degli Abruzzi 24, Torino, Italy, 10129, lohic.fotiotiotsop@polito.it; Tomas Mizdos, University of Zilina, Zilina, Slovakia, tomas.mizdos@feit.uniza.sk; Marcus Barkowsky, Deggendorf Institute of Technology, University of Applied Sciences, Deggendorf, Germany, Marcus.Barkowsky@th-deg.de; Peter Pocta, University of Zilina, Zilina, Slovakia, peter.pocta@feit.uniza.sk; Antonio Servetti, Politecnico di Torino, Torino, Italy, antonio.servetti@polito.it; Enrico Masala, Politecnico di Torino, Torino, Italy, enrico.masala@polito.it.

1 INTRODUCTION

Effective and accurate objective media quality assessment algorithms are a key element in optimizing multimedia systems, especially considering their ever increasing share in the global Internet traffic [7]. Many works in the literature focused on proposing new approaches to estimate the average quality perceived by the end users, i.e. the so-called mean opinion score (MOS) [38, 44, 51] or improvements to the existing approaches [1, 11, 24]. Some authors went even beyond the MOS, attempting to predict the standard deviation of the opinions of the subjects (SOS), interpreted as a measure of the dispersion of the observers' opinions around the MOS [15].

Despite the MOS and the SOS are certainly useful to measure the quality of experience (QoE) of end users, they alone are not sufficient. For instance, relying just on the MOS and the SOS, the skewness of the distribution of the subjects' opinions (DSO) is disregarded. However, the skewness plays an important role in estimating the actual QoE of single users: positive values indicate that the majority of users is actually experiencing a quality greater than the mean (i.e. the MOS), and vice versa. Some works proposed to overcome the limits of the MOS and SOS, well characterized in [14], by means of additional statistical indicators proposed in the same work [14], or deriving a range in which the MOS is expected to be with a predefined probability [9].

Also, the use of statistical moments such as mean and standard deviation implies to work with a numeric score for each subject. However, with the five points absolute category rating (ACR) scale subjects are asked to rate each stimulus by choosing one among five categories ("Bad", "Poor", "Fair", "Good", "Excellent") whose mapping to a numerical scale is somehow arbitrary in terms of distance between the categories. For this reason, it would be better to work with the distribution of scores so that no arbitrary mapping is introduced. Until now, few works focused on the DSO estimation. In [20], the authors proposed a generalized linear model for the DSO estimation, proving its effectiveness on a case study. In [45] and [55], a deep NN is trained to predict a probabilistic representation of the ratings gathered from actual observers. In a recent paper [40], the authors highlighted the importance of assessing the quality directly, relying on the subjects' opinions, and thus using the DSO instead of the statistical moments.

From all the reasons mentioned above, it is clear that having the subjects' opinion would be the best option since it allows more flexibility in subsequent processing. The ability to predict individual opinions is the basis of recommender systems used nowadays in various fields [8, 16, 23]. However, for many applications, in particular those involving the visual quality of the media, the preferences of individual users are affected by a great deal of uncertainty, e.g. successive evaluations of the same stimuli by an observer typically yield different opinions. In this work we propose, for the first time to the best of our knowledge, to create, through NNs, models, which will be able to mimic individual observers' behavior in terms of quality perception while taking into consideration the stochasticity that characterizes their choices, thus yielding a complementary approach to media quality assessment.

In practice, we propose to model the quality perception of single observers through a neural network. Instead of training a single NN to predict the MOS on the basis of the averaged result of subjective experiments, as it has been already done many times in the literature, we propose to train many NNs, one for each observer, to mimic the behavior of the observers in terms of quality perception. As we think that this approach of mimicking an observer by a neural network is indeed a kind of artificial intelligence, we would like to call the result an "artificial-intelligence-based observer" or "AI-Observer" (AIO) as compared to a "Human Observer". Each AIO can take, as input, a set of features computed on a given processed video sequence (PVS) and potentially also other features considering observer characteristics as well as the interaction between the observers and the context in which the experiment is carried out,

and attempts to predict the opinion of the observer which it is trained to mimic. The results presented in this work suggest that NNs can be used to effectively model the behavior of single observers in terms of visual quality perception.

It is worth noting that the modeling of single observers allows to implicitly take into consideration human factors such as personality traits, cultural diversities, personal experience regarding multimedia content, and user's expectations that have been shown to have an impact on the quality experienced by the end users [12, 39, 48].

An important added value of the proposed approach is the possibility to quantify the ability of an observer to repeat his/her rating if he/she would be asked to evaluate the quality of a PVS several times. The inability of observers to repeat themselves is an issue that has been investigated and considered in various models [21, 27] but this paper introduces, for the first time to the best of our knowledge, an actual model being able to predict subjects' inconsistency. In our work, the NN of each observer is designed to output a probability distribution consisting of five values representing the probability of each of the five alternatives on the ACR scale, as shown in Figure 2. While for the opinion prediction we simply select the option with the highest probability, the variance of this distribution measures the likelihood that the observer would give the same score if he/she would have to assess the quality of the PVS again. Therefore, we propose to use this variance as a measure of inconsistency of the observer regarding the quality of the PVS. Numerical results confirm that such value follows the typical characteristics of a quality inconsistency measure.

Thus, the main contributions of this work are: i) to formally present a complementary approach to model media perceptual quality while taking into account end users' characteristics and thus yielding a more comprehensive estimation of end users' QoE; ii) to discuss, through a deep qualitative analysis, the advantages of the proposed model over traditional approaches; iii) to show, through extensive computational experiments, that state-of-the-art machine learning models, in particular NNs, offer suitable tools that make the practical implementation of the proposed approach feasible. We emphasize that this work does not directly aim at establishing a quantitative comparison with the existing approaches, since we believe that this topic deserves a considerable effort and future work will mainly address this issue exclusively. It should be noted here that the aforementioned contributions significantly extends our preliminary work [10], where just the basic idea has been presented without, for instance, analyzing and proving its generality over traditional approaches and addressing the modeling of subjects' inconsistency or showing extensive quantitative analysis.

The remainder of the paper is organized as follows. In Section 2 the proposed approach for media quality assessment is presented, followed in Section 3 by an in-depth analysis of its strengths over the traditional approach. Section 4 presents our methodology to design, train and test the NNs that model single subjects. Numerical results are presented in Section 5, followed by Section 6 which draws conclusions and highlights possible future directions.

2 A COMPLEMENTARY APPROACH FOR OBJECTIVE MEDIA QUALITY ASSESSMENT

Objective media quality assessment methods aim at measuring the perceptual quality as judged by the end users, which is strictly related to the users' QoE. The concept of QoE has been defined by ITU [18] as "the overall acceptability of an application or service, as perceived subjectively by an end-user". This definition implicitly underlines the importance of considering individual user expectations while measuring the QoE. A more recent and more encompassing definition has been given by the QualiNet white paper [4]: "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state". This latter definition explicitly mentions the user expectations and personality among factors that are to be considered when assessing the end user QoE. In other words, the user characteristics as well personal experiences contribute to determine the user's opinion regarding

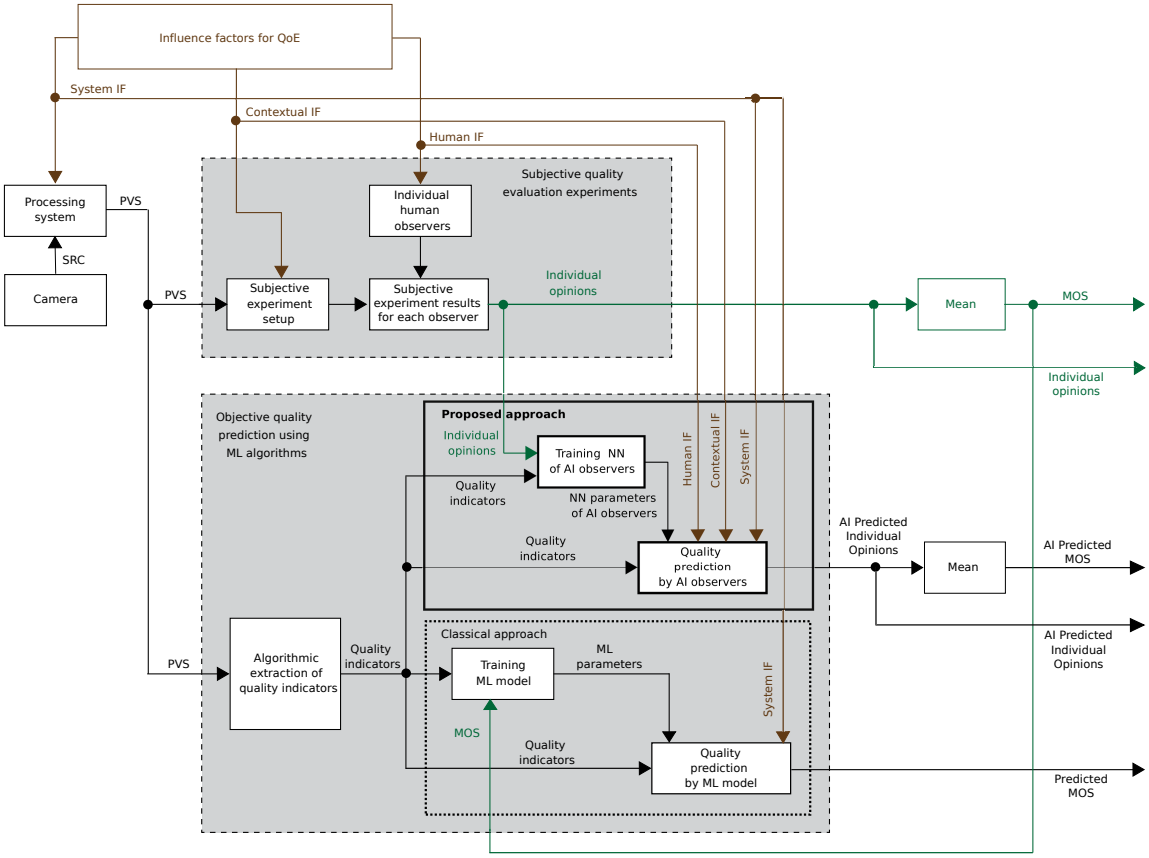


Fig. 1. Illustration of the proposed approach in comparison with the traditional approaches to media quality assessment. In particular, note that similarly to subjective experiments, the proposed approach considers also human factors and provides individual opinions yielding, in practice, more flexibility.

the perceived visual quality of a media and should therefore be considered as much as possible when designing models for QoE estimation.

While subjective evaluation methods do take into account such aspects up to some extent, it is typically neglected by the traditional approaches for an objective media quality assessment. In fact, subjective experiments are carried out by inviting a number of subjects selected according to certain heterogeneity criteria. That allows to consider a sample of observers that best represents some end-user's characteristics, e.g. the user's age [31], user's gender [17] or the user's preferences [35] that can have an impact on the subject's judgment. The effect of such human influence factors (IF) is represented by the diversity between the individual opinions gathered during the experiment. Moreover, in actual subjective experiments also context IFs, e.g. the conditions in which the test is run, can contribute to form the individual opinions [22].

On the contrary, the traditional objective approaches to media quality assessment fundamentally focus on predicting a numerical value, i.e. the MOS [6], potentially also considering system IFs, for instance using different models depending on some system parameters such as resolution. To compute the MOS value used, as a target, by the traditional objective approaches, individual opinions coming from subjective experiments are first arbitrarily mapped to numerical values

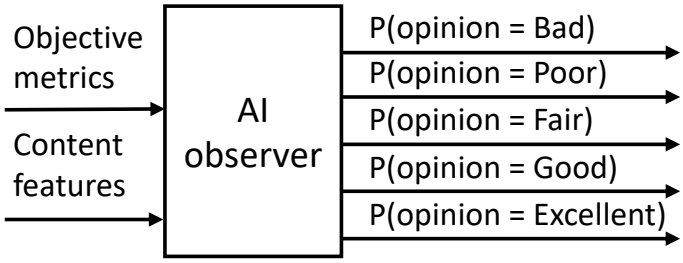


Fig. 2. The proposed artificial intelligence observer (AIO).

and then often simply averaged, thus neglecting the information related to the personality of the observers and their expectations.

For this reason and others that will be discussed in Section 3, we suggest a different approach to objective media quality assessment, illustrated in Figure 1.

More precisely, rather than designing models for direct MOS prediction, we propose to model single observer's opinions. Figure 1 presents our proposal comparing it with the traditional subjective and objective approaches. By modeling each observer, we implicitly take into account the characteristics of the individual observers (human IF) since the model of each subject (named "AI observer" in the figure) is constructed by the observer's own opinions which, as such, are only influenced by the observer's own characteristics and expectations.

To model a single observer we propose to use a NN. In fact, machine learning (ML) models have demonstrated a high accuracy in approximating rather complex processes and have been extensively used for predicting the way users judge quality in many applications [49]. They have therefore been naturally leveraged also in media visual quality assessment research [25, 52, 54, 56]. More precisely, we recommend that, starting from data collected during a subjective experiment, a NN has to be designed and trained for each observer. Such NN takes, as input, features derived from a set of PVSs, potentially also from the relative sources (SRCs) and, as ground truth data, the opinions that the observer expressed about the quality of those PVSs. Once these NNs are trained (one for each observer), they can be seen as a substitute for the actual observer in the sense that by receiving the information related to any PVS as input, they are expected to predict the opinion that the respective observers would have expressed after experiencing that PVS. Being NN-based, we will refer to them as artificial intelligence observers (AIOs) in the remainder of this work. Figure 2 illustrates the AIO.

The NN, which models each observer, is designed to output a discrete probability distribution made up by five probability values. These five values represent the probability of choosing any of the five options in the ACR scale. We estimate the predicted opinion of the observer as the mode of this discrete probability distribution, i.e. the highest probability option. On the other hand, the variance of this distribution, is expected to inform about the ability of the observer to repeat the same opinion on the quality of the PVS if asked to rate it again.

Such AIOs can then be used to run a virtual subjective experiment. In practice, the final visual quality of a PVS could be determined from the automatically generated opinions (AGOs) in a way that is most suitable for the considered purpose. For instance, classical statistical operators such as the mean and the mode can be employed. However, since the estimation of the whole distribution of opinions is available, quantiles and other more advanced statistical quantities can also be computed. This could be important for certain purposes. For instance, streaming providers could rely on quantiles since they would preferentially care about the percentage of the votes under a certain threshold to estimate

the fraction of the end users that would experience an acceptable quality. On the other hand, the mean could be used when comparing for instance the performance of two codecs. Also, in the case when providers collect informal quality feedbacks from the end users, as it often happens at the end of communication session, such users could be clustered and then potentially matched to the AIOs. Therefore each AIO becomes more representative of a set of end users with similar expectations and background. With this mapping the provider can better optimize the QoE for each cluster of users, thus enhancing the QoE of the service as a whole.

3 COMPARATIVE ANALYSIS WITH THE TRADITIONAL APPROACH

Recent works in the literature [9, 14, 20, 29, 40, 42] strongly suggest that the MOS, even when coupled with the SOS, is not sufficient to estimate the QoE experienced by the end users, and that better modeling approaches are needed. However, the reasoning behind such efforts is sometimes unclear. For this reason, this section aims to clarify the relevance of our complementary approach to video quality assessment in comparison to the traditional ones.

3.1 Issues in MOS and SOS definition

When conducting a ACR-based subjective experiment, the observers are asked to watch a stimulus and express their opinion on their perceived visual quality by choosing one among the five options of the ACR scale, which are typically indicated as labels ("Bad", "Poor", "Fair", "Good", "Excellent"). Since the arithmetic mean and the standard deviation are not defined for ordinal data, in order to be able to compute the MOS and the SOS, each of these options is mapped to an integer value starting from 1 for "Bad" and ending up with 5 for "Excellent". Such an apparently trivial mapping has strong implications that make the validity of the MOS as an effective QoE estimator highly questionable. In fact, it is not very clear whether the effort required by, e.g. reducing compression artifacts to change the opinion of an observer from "Bad" to "Poor" is equal to the one needed to make the observer change from "Good" to "Excellent". In [30, 42], the authors argued that the perceptual distance between "Fair" and "Poor" is larger than the one between "Poor" and "Bad" and it is also dependent on the language used during the subjective experiment. In other words, despite the fact that the five options of the ACR scale can be ordered, one has no guarantee that the options are equidistant. Hence the traditionally used 1 to 5 mapping might include significant bias in the evaluation process, yielding potentially large errors in the MOS and the SOS estimation (seen as indicators of the user QoE). Furthermore, the same issue, i.e. mapping ordinal data to numerical scale, has also been traditionally disregarded when designing models and algorithms for subjective quality recovering [21, 27]. We notice that the question of how to analyze ordinal data with unknown gap sizes between the categories using statistical indicators and models defined for numeric data is an issue that attracted and still attracts significant interest [28, 33, 37], hence it would be misleading to simply ignore such an issue.

In the light of the previous argument, we believe that it is important to design new media quality assessment approaches that rely directly on the ordinal data collected during subjective experiments as proposed in this work. In fact, in the proposed approach the five options of the ACR scale are *not considered as numbers* but rather as five levels of quality perception. This is reflected in the NNs mimicking each observers. Given a PVS, they simply attempt to predict the option that the related observer would have chosen.

3.2 Importance of single observer's ratings in practice

Media service providers really care about the question of ensuring that a certain percentage (hopefully high) of customers is satisfied by the quality of the service they are paying for. In other words, having an estimation of the percentage of customers that would express an opinion different than, e.g. "Bad" and "Poor", is of a paramount importance. While

such an issue can be immediately addressed by having the AGOs provided by the AIOs, it is clearly much more difficult to derive any estimation of such percentage from the MOS. Even when the SOS value is also available, an accurate estimation of the desired percentage is still very difficult since the skewness of the distribution strongly impacts on the position of its quantiles.

3.3 General applicability of the proposed approach

The proposed approach, being rather general, allows to use, if desired, many QoE estimators proposed in the literature, since they can be rather easily derived from the DSO [40], which is estimated by our approach. For instance, the MOS and the SOS can, respectively, be derived from the first and the second order statistical moments of the DSO after mapping the AGOs to a numeric scale. Furthermore, a complete estimation of the DSO allows to compute more accurate confidence intervals as well as to more accurately run statistical tests regarding an apparent higher perceived visual quality of a PVS with respect to another one [40]. In fact, it is possible to avoid the usage of the classical statistical tests designed for numerical data, on the ordinal data collected during a subjective test, since that would inflate the type 1 and type 2 errors of statistical tests, as highlighted in [28].

3.4 Pre-simulation of subjective experiments

Subjective experiments are crucial for the development and validation of ML based objective measures for media quality assessment. However, the performance of ML models is known to be strongly related to how informative and exhaustive is the underlying training set. Therefore, when designing subjective experiments, one of the major concerns is to make sure that the subjectively perceived visual quality of the PVSs to be submitted to the observers' judgment fully covers the chosen rating scale. The AIOs could be used, before running the actual subjective experiment, to simulate the behavior of the observers on the PVSs selected for a subjective test, hence gaining a preliminary insight on the heterogeneity of the chosen PVSs in terms of perceptual quality.

3.5 Human factors and QoE

It is worth noting the similarity between the proposed approach and how subjective experiments, which are usually considered the best way to evaluate the visual quality of media, are run. Given a PVS whose quality is to be evaluated, the AIOs provide AGOs as real observers would do during a subjective experiment. The diversity observed between these AGOs highlights the impact of individual characteristics and expectations in the QoE measurement. These subjects-specific factors have already been intensively studied. Several works [39, 48, 58] suggest that their consideration would improve the accuracy of models aiming at predicting the end-users' QoE. Therefore, it seems natural and appropriate to develop approaches that manage to take into account the differences between subjects in terms of sensitivity to distortion and expectations. This is an additional reason to model each single observer with a NN relying on data gathered during subjective experiments. Since each observer's ratings are implicitly influenced by his/her personality and expectations, we expect that this approach takes into consideration all the factors that contributes to each subject's judgment.

4 NEURAL NETWORK BASED SUBJECTS' MODELING

In this Section we present general guidelines on how to effectively prepare the data and train the neural network which models each observer. The feature set as well as the training process adopted in this work are then discussed and finally a measure of subject' inconsistency is introduced.

Table 1. Description of the datasets used in the experiments

Dataset	Size	Distortions	Notes
VQEG-HD1 VQEG-HD3 VQEG-HD5	~ 160 seq. each, 1080p, 10-sec long	MPEG-2 and AVC-encoded, 1 to 15 Mbps, transmission distortions due to bit errors and bursty packet losses	movies, sports, general TV material with as much variety as possible
ITS4S	514 seq, 720p, 4-sec long	AVC-encoded at either 512, 951, 1256, 1732, 2340 kbps.	Already classified into 9 categories: Broadcast, Everglades, Music&Mexico, Nature, Ocean, Public Safety, Sports, Training, and Chance (miscellaneous content)
JEG-Hybrid	59,520 seq, 1080p, 10-sec long	HEVC-encoded (0.5 to 16 Mbps + constant QP)	Not subjectively annotated

4.1 Guidelines for AIOs derivation

The media quality assessment community has long proposed learning based models aimed at predicting the MOS. However, the transition from models for MOS prediction to AIOs brings new challenges not only in the data preparation process but also in the training process. In fact, models for MOS prediction are designed in a way that the diversity of individual subjects in terms expectation and experience is disregarded in the end of the quality assessment process. The AIOs instead should be designed so that they model such a diversity as well as the subjects' inconsistency in order to capture the impact of human influence factors. Therefore, we propose the following guidelines for an effective implementation of the proposed AIOs-based approach:

- **Data preparation:** Subjective experiments represent the main tool for gathering data used by machine learning based models. The AIOs-based approach proposed in this work involves a rethinking of the methods typically used for conducting subjective tests. In fact, unlike the experiments tailored for MOS prediction where the number of subjects is very important, in our case an important parameter is the number of sequences evaluated by each single subject. The more ratings are available for a given subject, the better it is for modeling his/her quality perception. The subjective test should therefore be designed in such a way that the same subject rates a huge number of stimuli rather than having many subjects in order to yield a reliable value of the MOS. This comes with its own challenges. For instance, it is important to correctly take into account the fact that the ratings might be affected by subjects' fatigue and potentially such aspect should be modeled. Unfortunately, tools to address this issue are still lacking in the literature. Hence, how to create effective datasets for training the AIOs could be an entirely new field of research for the media quality assessment research community. In this work, as discussed later in Section 5, we propose an approach to combine existing experiments in order to obtain a training set suitable to create the AIOs.
- **Training set structure:** When rating content quality, viewers are largely more consistent at the extreme values of the quality scale, i.e., on high and low quality content. Thus, when creating or selecting the dataset to train an AIO it is really important to have a large number of observations in the middle part of the quality scale: this would allow to better model the mechanism guiding the observers' choices in terms of quality perception. Note

that this is different from the case of many machine learning-based applications for which a uniform distribution of the labels in the training set is strongly recommended.

- **Architecture of the NN representing an AIO:** The architecture of the NN modeling each observer depends on both the observer and the amount of data available in the training set. For some observers, the input features are already suitable and the network role is to determine the best way to map them to the quality scale. For other subjects, instead, the derivation of more complex features from the input ones is required. In the former case, a single-hidden-layer architecture is enough to model these observers' quality perception, whereas in the latter case more than one hidden layer is required. This aspect also allows to classify the observers on the basis of the complexity of the mechanism that guides their perception of quality. Obviously, the number of hidden layers suitable for a given observer is not known a priori, thus it should be determined through numerical experiments. The number of neurons for each layer, instead, is strongly related to the size of the training set. The larger the training set, the more the neurons that can be used in each layer. In any case, the output layer of the NN should always consist of five neurons, each predicting the probability that the AIO chooses one of the five opinions of the ACR scale. The labels in the training set must therefore be appropriately coded for this purpose. Using probabilities as output values is fundamental for modeling the inability of subjects to repeat themselves in subjective experiments.
- **Analysis of the Accuracy:** Avoiding overfitting is a major concern when using machine learning algorithms. In the case of AIOs, previous studies in quality assessment can be used to determine an accuracy threshold on the training set above which the presence of overfitting is highly probable. In fact, in the best case, we expect that the AIOs act with an accuracy similar to the one of the actual observer. Therefore, it is important to analyze what is the accuracy of a subject when he/she is used as a classifier of himself/herself. More precisely, if an observer evaluates for the second time the quality of a video sequence, what would be the expected accuracy? The experiments presented in [21] show that, when re-evaluating a set of video sequences, subjects are able to repeat their first opinions, on average, only 57% of the time and the best subject achieved 74% accuracy. Furthermore, on average, for 94% of the PVSs, each subject selected a rating that differs at most by one from the previous rating. These numbers provide indications on the upper bounds for the expected accuracy of the AIOs on the training set. More precisely, when training and testing the AIO for mimicking an actual observer, an accuracy equal or higher than 57% is already suitable. However, when the AIO accuracy is significantly higher than 74% on the training set, then the suitability of the model needs to be further investigated. In fact, being the observer not able to repeat the same opinion in correspondence to the same input, the training set is certainly noisy, and thus large accuracy would be observed only if the peculiarities of the training set are learned. On the other hand, when an accuracy close to 74% is obtained for an AIO, this does not necessarily mean that it is accurate: numerical experiments on data never seen during the training are still required to draw definitive conclusions.

4.2 AIOs' training process and subjects' inconsistency measure

To train the NN mimicking each observer involved in a subjective test, first we considered a set of features, here denoted by $\mathcal{F}$, characterizing the PVSs evaluated during the subjective experiment. In this work, we consider, as features, five video quality metrics, i.e. the PSNR[50], the SSIM[57] the MSSSIM[47], the VIF[41] and the VMAF[32], as well as six perceptual features, i.e. "Blockiness", "Blockloss", "Blur", "Contrast", "Flickering" and "Noise", which attempt to quantify how much each of the listed artifacts is presented in each PVS. The features are described in details in [26]. Finally, we

also consider the spatial activity index (SI) and the temporal activity index (TI) that we computed as in [34]. Then, for each observer, we proceed to create the corresponding AIO with the procedure described in the following.

First, for each observer, we determine the subset of features as well as the NN structure that best model the quality data of such observer.

In order to find the best set of features we proceed as follows. From all the possible subsets of features selected in $\mathcal{F}$ containing at most five features, and the ratings of the observer in the training set, we trained three different NNs having respectively one, two and three hidden layers with five neurons each, and an output layer with five neurons delivering the probability of choosing any of the five possible options of the ACR scale. Then, we tested the three NNs obtained for each possible subset of the features on the test set by comparing the predicted opinions with the actual ones. For each observer, we considered the NN structure and the related subset of features that yielded the highest accuracy on the test set as his/her final model.

The aforementioned settings of the NNs, i.e. the number of hidden layers and the corresponding number of neurons, have been experimentally determined as the most effective. Three NN structures with different depths have been examined for each observer in order to investigate what is the level of complexity required to effectively model the observer.

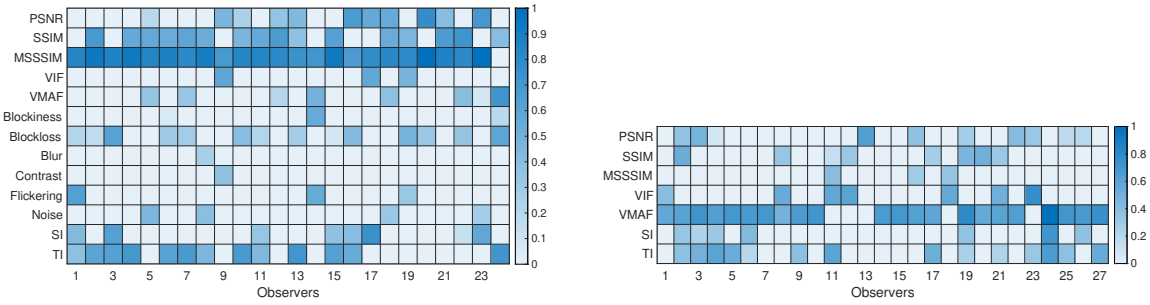
Hence, each observer is modeled by a NN in which the number of neurons on the input layer is equal to the cardinality of the subset of features that best models the observer, the number of hidden layers varies from one to three depending on the complexity of the observer and finally the output layer has five neurons that predict the probability of choosing each one of the five opinions. Therefore, once the NN, trained to mimic an observer o, is deployed on a PVS, it outputs the following discrete probability distribution $p_{oi}^{PVS} \;\; i = 1, 2, \ldots, 5$, where the index i represents the five alternatives offered to the observer. The predicted opinion is then chosen as the alternative with the highest probability.

Furthermore, denoting by $v_i \;\; i = 1, 2, \ldots, 5$, the numerical score of the five alternatives available on the ACR scale, we propose to use the variance of such a distribution, i.e.

$$\sigma^2(o, PVS) = \sum_{i=1}^{5} v_i^2 \cdot p_{oi}^{PVS} - \left(\sum_{i=1}^{5} v_i \cdot p_{oi}^{PVS} \right)^2 \tag{1}$$

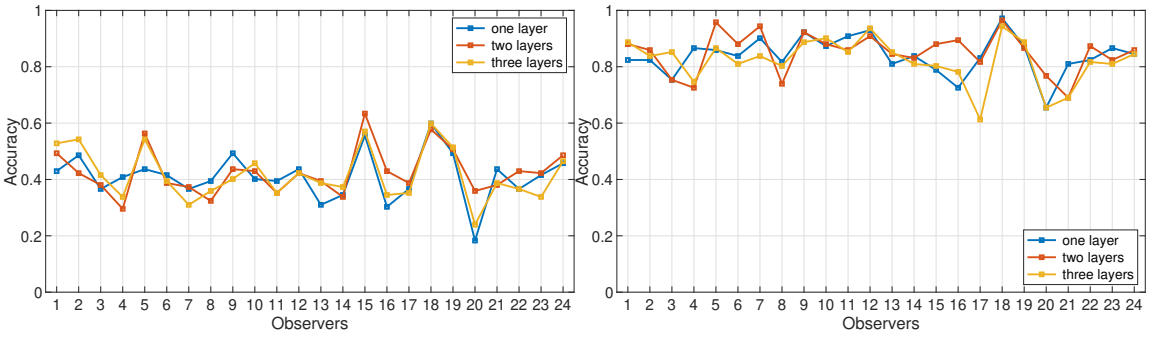
as a measure of the inconsistency of the observer o regarding the perceived visual quality of the PVS under examination. In fact, a high value of $\sigma^2(o, PVS)$ indicates that opinions different from the mode (i.e. the one with the highest probability) report a non-negligible probability value. Modeling the observer o using such a probability distribution allows to consider the fact that repeated evaluations by the same observer could naturally yield different opinions over time even for the same PVS. Hence, the $\sigma^2(o, PVS)$ value informs about how likely it is that the observer o would repeat itself in subsequent evaluations.

To understand how the measure described in Eq (1) captures the observer inconsistency, let us make the following considerations. For a consistent observer, there is a way to accurately map the features that characterize the perceptual quality of stimuli to his/her ratings. In other words, for this type of observer, the feature space can be almost perfectly partitioned and clustered on the basis of his/her ratings on the ACR scale. The AIO of this type of consistent viewers just needs to learn the mathematical expression of this partition from the training data to be able to perform a classification with high confidence. It is therefore expected that for a consistent observer, the variance of the neural network output is low. The opposite argument holds for non-consistent observers, for which it is not easy to find a subdivision of the features space in disjoint subsets, each one associated with a different quality level on the basis of the observer's ratings. The high variance of the neural network output expresses precisely this lack of consistency between the input features



(a) VQEG-HD dataset (b) ITS4S dataset

Fig. 3. Importance of the features for modeling the quality perception of each observer. Each feature's importance is obtained using the neighborhood component analysis feature selection algorithm. The heterogeneity of the columns suggests that users rate the quality on the basis of different criteria.

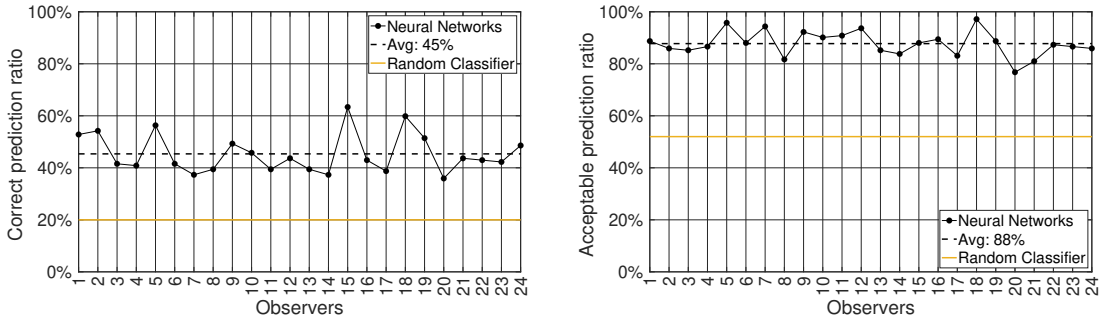


(a) Correct prediction ratio (b) Acceptable prediction ratio

Fig. 4. Accuracy of the three NN structures (characterized by different depths) when trained for each observer. In many cases the performance of the three NNs is similar but occasionally there is a significant performance difference.

that determine the objective quality of the input video sequence and the corresponding votes given by the observer. In other words, for stimuli with the same objective quality and thus with the same value of the perceptual features, an inconsistent observer is inclined to give opinions that are significantly different. During the training of the AIO of such an observer, by using only his opinions, the model learns that these different opinions given by the observer on stimuli having the same objective quality are equally probable. During the test phase, when the AIO receives a stimulus as input, the probabilistic output is equally distributed over several opinions, leading to greater variance.

A subjective experiment in which the same observer is asked to rate a significant number of times the same stimuli would be required to fully assess the accuracy of the inconsistency measure proposed in Eq (1). This is unfortunately too expensive to be carried out in practice. For this reason, in the Section 5 a different approach is adopted to show Eq (1) effectiveness as a measure of inconsistency. More precisely, we will show that the proposed measure possesses the properties that are expected from an inconsistency measure.



(a) Correct Prediction Ratio (b) Acceptable Prediction Ratio

Fig. 5. Accuracy of the AIOs. The NNs were trained on the VQEG-HD1 and VQEG-HD5 data sets, and tested on the VQEG-HD3 data set. The performance is significantly higher than that of the randomly voting observer (yellow line).

5 RESULTS

To assess the feasibility as well as the effectiveness of the proposed approach, we conducted extensive numerical experiments that are presented and discussed in this Section. In order to compare the AIOs with a random classifier and the MOS, and also to develop a data augmentation approach to cope with the lack of datasets for an effective training of the AIOs (see the guidelines in Section 4), we relied on the widely used mapping of the ACR scale labels to integers from 1 to 5, despite having pointed out the limit of this behavior. However, note that this is done only for comparisons and data augmentation purposes and does not imply that the use of the proposed AIOs-based approach is restricted to this case.

5.1 Datasets description and experimental setup

The numerical experiments were done considering four subjectively annotated datasets, i.e. the VQEG-HD1, VQEG-HD3, VQEG-HD5 [46] and the ITS4S [13, 36] dataset. In addition, the large scale JEG-Hybrid dataset [2], that has not been subjectively annotated, has also been used, since it contains many more PVSs than the former ones. The characteristics of the used datasets are summarized in Table 1. As it can be clearly seen from the table, during the VQEG-HD experiments several types of distortions have been applied to the PVSs, while only coding artifacts have been considered in the other datasets. Therefore, for our numerical experiments, we mostly relied on the VQEG-HDTV datasets since this allowed us to investigate the effectiveness of the approach for a wider range of cases.

For completeness' sake, a brief description of the VQEG-HD experiments is provided. Such a description is based on the three fundamental aspects considered in [19]. More details can be found in the VQEG-HD experiment final report [46].

Test environment: The VQEG-HD experiment took place in six different laboratories, leading to six different datasets named VQEG-HD1, 2, 3, 4, 5 and 6. The environment of each laboratory was prepared in accordance with the ITU-R Recommendation BT.500-11 [5]. In general, a test session involved only one viewer per display assessing the stimuli. The viewer was seated in front of the screen at a distance equal to three times the height of the picture. Either high-end consumer TVs (Full HD) or professional grade LCD monitors were used in all the laboratories. In all the cases, the display resolution was $1920x1080$.

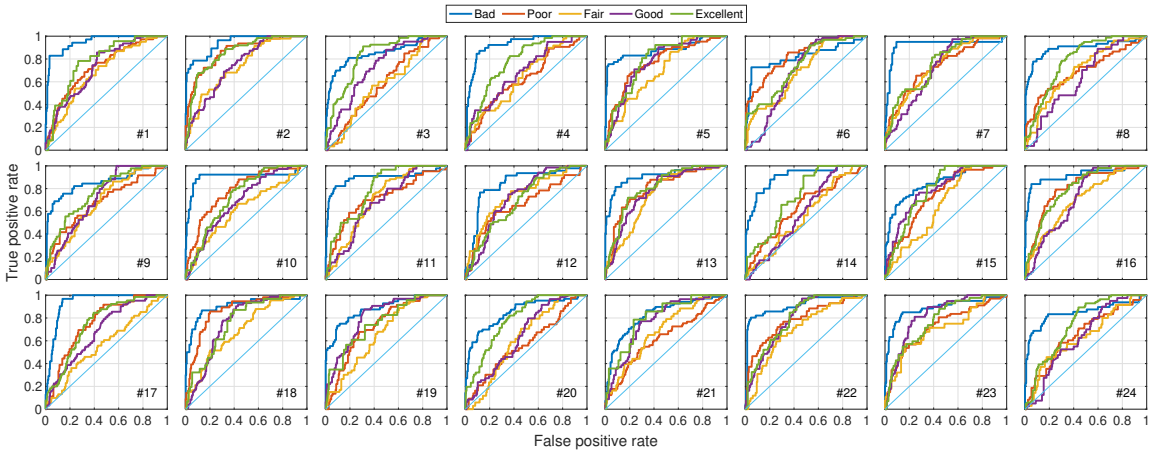


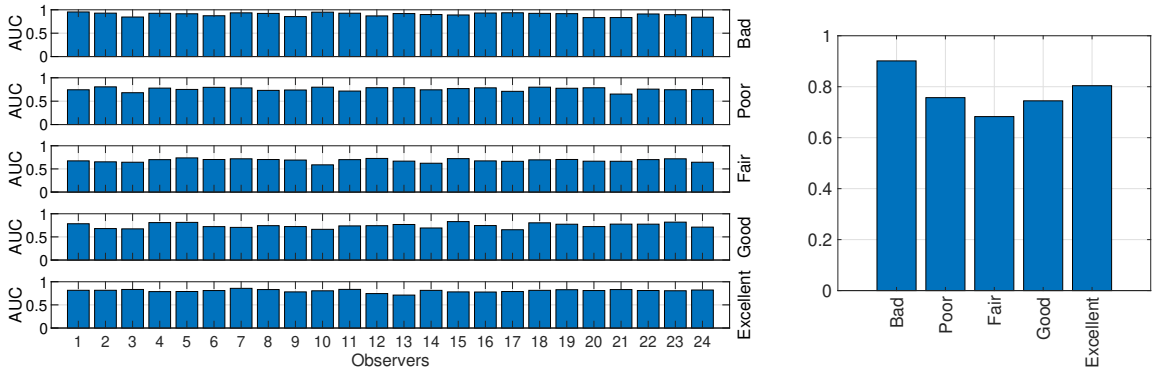
Fig. 6. The ROC curves associated with the NN , which models each observer. In all the cases, the curve is above the 45 degree line: the NN is effectively modeling some of the aspects that concur with the way how the observer perceives the visual quality.

Participant cohort: In each of the 6 laboratories, 24 viewers participated in the test. The viewers were screened for normal visual acuity (with or without corrective glasses) by means of the Snellen test [43] and for normal color vision by means of the Ishihara test [3]. After the completion of the test, a statistical criteria was used to assess whether each viewer's ratings were consistent with the average of the others viewers. If not, that viewer's ratings were rejected and a new one was invited to participate in the test (more details on the criteria in [46].)

Assessment procedure: The absolute category rating with hidden reference [34] was used. Each video sequence, also including the reference one, was shown exactly once to the subject that was then asked to rate the visual quality by choosing one among the following alternatives: "Bad", "Poor", "Fair", "Good" and "Excellent". In all the six laboratories, before starting the experiment, each subject received a short tutorial aimed at familiarizing not only with the assessment procedure but also with the software used to record the votes. After this tutorial, the stimuli were shown in a randomized order to the subjects. A break was given to each subject after evaluating half of the stimuli to minimize potential inaccuracies due to fatigue. Each video sequence was 10-second long; between one stimulus and the other the display was kept grey until the subject expressed the opinion. Table 1 summarizes the characteristics of the used stimuli.

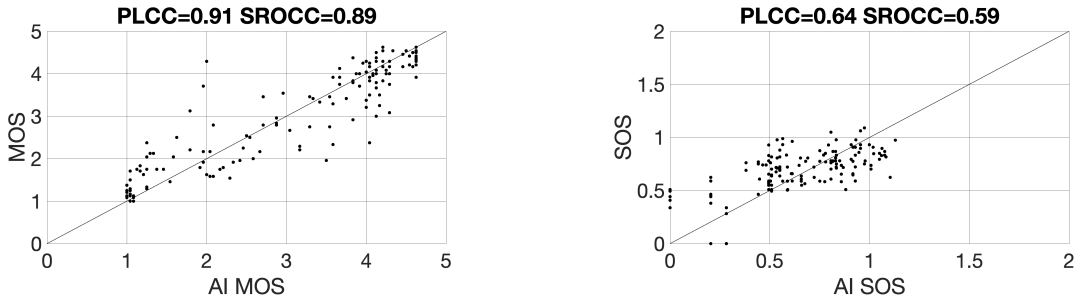
Please note that in this work we did not consider the VQEG-HD2, VQEG-HD4 and the VQEG-HD6 datasets because the first two include interlaced content that is out of the design scope of the full reference measures used as part of the features, and the latter is not publicly available. During each of the three sets of the VQEG-HD experiment considered in this study, as already mentioned, 24 observers were asked to rate almost 160 PVSs. Even if the three experiments took place in different laboratories, with different observers and different set of PVSs in each laboratory, 24 PVSs have been rated in all the three experiments and thus by all the 72 observers. In the remainder of the work, we refer to these PVSs as the "common set".

To model an observer in terms of visual quality perception, one should ideally rely only on the opinion scores related to the PVSs actually watched and rated by that observer. However, we experimentally observed that the ratings of 160 PVSs are not enough to effectively train and test a NN which models an observer. To be able to use more data for each observer, we approximated the observers' ratings, on the PVSs whose quality has not been assessed directly by



(a) AUC indexes for each observer's model (b) Average AUC

Fig. 7. The AUC indexes associated with the NN, which models each observer. The closer to 1, the better. The NNs seem to be more accurate when modeling the observer's behavior in the case of the PVSs with the very low or high quality.

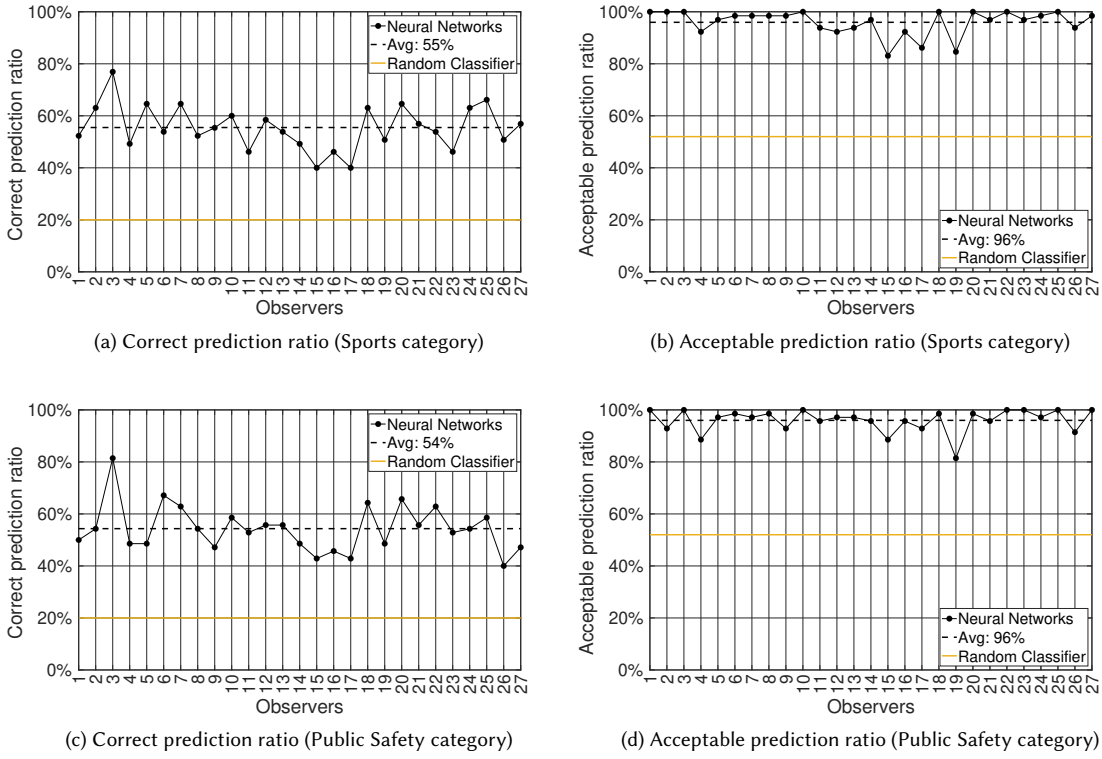


(a) (b)

Fig. 8. Results obtained when deploying the AIOs trained on the VQEG-HD1 and VQEG-HD5 datasets on the PVSs coming from the VQEG-HD3 dataset to simulate a subjective experiment. The AI MOS and SOS are computed respectively as the average and the standard deviation of the AGOs.

that observer, by using those provided by another observer that voted similarly on the common set, according to the mapping procedure explained in [10], which is briefly reported in the following for completeness' sake. Let consider an observer $\hat{O}_1$ that was involved in the VQEG-HD1 experiment. For such an observer, the votes for the PVSs used during the VQEG-HD1 experiment are readily available. As this set of data is not enough to train and validate the NN that would mimic the observer $\hat{O}_1$, we approximated the votes that he/she would have given to the PVSs considered in the VQEG-HD3 experiment by those provided by an observer $\hat{O}_3$ that participated in the VQEG-HD3 experiment and rated the PVSs in the common set very similarly to the observer $\hat{O}_1$. The similarity between the observers opinion scores is assessed through the mutual residual mean square error between the ratings given by the observers to the PVSs inside the common set. Hence the observer $\hat{O}_3$ is determined as follows:

$$\hat{O}_3 = \arg\min_{O_3} \sqrt{\frac{1}{|C_{set}|}\left(\sum_{s \in C_{set}} (V_s^{O_3} - V_s^{\hat{O}_1})^2\right)} \tag{2}$$



(a) Correct prediction ratio (Sports category)

(b) Acceptable prediction ratio (Sports category)

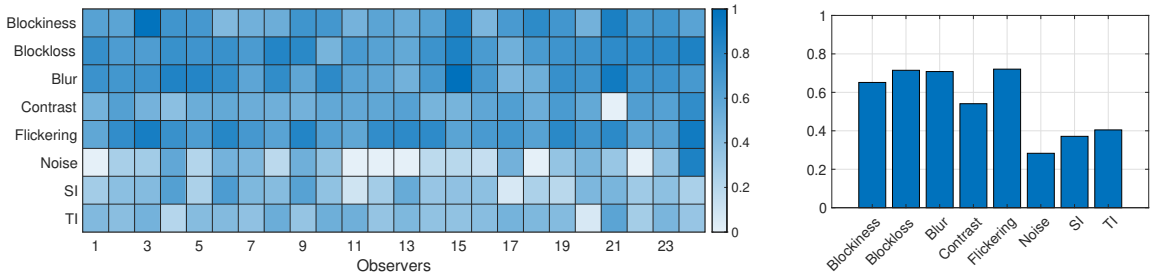
(c) Correct prediction ratio (Public Safety category)

(d) Acceptable prediction ratio (Public Safety category)

Fig. 9. Accuracy of the AIOs on the ITS4S dataset. The NNs were trained using the PVSs in all the categories except the one shown in the figures. The performance is significantly higher than that of the randomly voting observer (yellow line).

where O_3 indicates a generic observer of the VQEG-HD3 test, s is a sequence, $V_s^{O_3}$ is the vote of the observer O_3 for the sequence s and finally C_{set} represents the common set. Following the same approach, we connected the observer $\hat{O}_1$ to an observer $\hat{O}_5$ of the VQEG-HD5 experiment. Therefore, starting from 72 observers, which rated 160 sequences each, 24 triplets of the observers have been formed such that the observers in the same triplet perceive quality similarly. Each triplet of the observers was then treated as a single observer, leading to 24 observers that can be considered to have rated 480 PVSs. We then trained and validated 24 NNs to mimic those observers, thus obtaining 24 AIOs.

It is important to note that inferring the behavior of an observer on 160 stimuli starting from what was observed on the common set that contains only 24 sequences would be reasonable only if the common set is made out of appropriately selected stimuli, i.e. a subset of stimuli that reasonably summarizes the characteristics of all the other ones involved in the experiments. This is the case for the VQEG HDTV experiment [46]: the 24 sequences in the common set were carefully selected to span the full range of quality considered during the experiment. The reason behind that is that the common set was originally designed to map all the results of all the laboratories to a common scale. Therefore, the approach used in this work as a data augmentation strategy introduces some noise in the training set. However, in the light of the observations made on the structure of the common set such a noise is expected to be minimal.



(a) Importance of features for each observer

(b) Average importance of features

Fig. 10. Importance of the features for the observers' inconsistency modeling. The importance of each feature for each observer is obtained using the neighborhood component analysis feature selection algorithm.

We considered the following two ratios:

$$\textit{Correct prediction ratio} = \frac{\#(\textit{predicted OS=actual OS})}{\#(\textit{PVS in test set})}$$

$$\textit{Acceptable prediction ratio} = \frac{\#(|\textit{predicted OS-actual OS}| \leq 1)}{\#(\textit{PVS in test set})}$$

in which OS stands for opinion score. they are used together with the receiver operating characteristic (ROC) curves as well as the area under the ROC curve (AUC) indexes associated with the NN modeling each observer to assess the accuracy of the 24 AOIs. The correct prediction ratio and thus the accuracy of each NN achieved on the test set is the number of PVSs for which the rating predicted by the NN is equal to the one given by the related observer divided by the total number of PVSs in the test set. The acceptable prediction ratio, instead, represents the number of PVSs for which the NN prediction differs no more than 1 level from the rating of the related observer divided by the total number of PVSs in the test set. For a random classifier (RC), i.e. an observer which randomly selects its scores, these ratios are respectively 20% and 52%, which are the expected values considering all the possible favorable cases (2/5·1/5+3/5·3/5+2/5·1/5 for the acceptable prediction ratio).

5.2 The AIOs accuracy

In the case of the VQEG-HD dataset, the data of the VQEG-HD1 and the VQEG-HD5 were used as the training set, while the VQEG-HD3 was used as a test set. The "common set" was included only in the training set, and therefore there were no identical PVSs in the training and test set. The ITS4S dataset classifies the PVSs into nine different categories, see Table 1 for more detail. The sport and public-safety categories were considered as the test set. The respective PVSs were excluded from the training process, which was conducted on the rest of the dataset. Figure 3a and Figure 3b report the importance of each feature for modeling the quality perception of each of the observers respectively in the case of the VQEG-HD and the ITS4S dataset. We relied on the neighborhood component analysis feature selection algorithm [53] to determine the weight of each feature, i.e. its importance in modeling the quality perception of each observer. It is worth noting here that, in general, the importance of each feature changes from one observer to another. This is expected, since subjects typically judge quality on the basis of different criteria.

We remind that for each of the 24 observers, three NN structures were tested in order to handle potential diversities in terms of modeling complexity between the 24 observers. In Figure 4a and 4b, we show the correct prediction ratios and the acceptable prediction ratios of these NNs when trained using the best set of features for each observer and

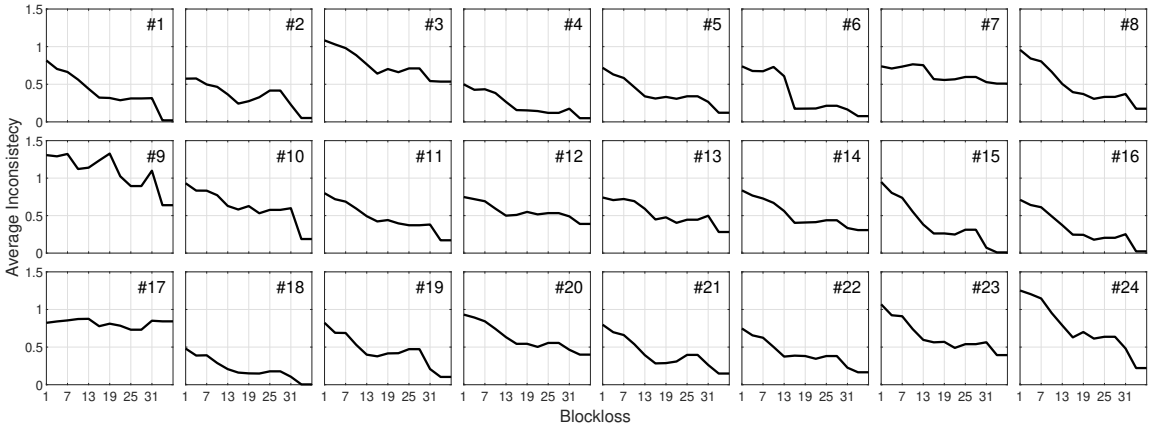


Fig. 11. The average observers' inconsistency tends to decrease as the "Blockloss" feature value increases for almost all the observers. For each value of the Blockloss feature on the x axis, the graph shows the average inconsistency of the observer evaluated on PVSs for which the Blockloss feature value is greater than or equal to the one on the x axis.

tested on the test set. As it can be seen, in general, the three curves are rather close to each other and thus none of the structures seems to be particularly suitable compared to the others. However, some exceptions are present. For instance, the observer $\#5$ seems to be rather complex to model since a NN with a single hidden layer delivered lower accuracy than those with more layers. It can also be noted that a NN with two hidden layers seems to be particularly suitable for the observer $\#20$ while the ones with one and three hidden layers lead to underfitting or overfitting, respectively.

We then chose, among the three NNs available for each observer, the one with the highest accuracy on the test set. In Figure 5 the accuracy of the 24 chosen NNs is compared to that of an observer randomly rating the PVSs in the test set. By making such a comparison, we aim at verifying whether the NN mimicking each observer did learn interesting information about the way the observer perceives quality, otherwise we would expect performance similar to that of a completely inconsistent observer that randomly rates stimuli. For each of the 24 observers both the correct and the acceptable ratio of the NN modeling their choices in terms of quality perception significantly exceeded the expected accuracy of the random classifier. In particular, an average accuracy of 45% ($> 20\%$) and an acceptable prediction ratio of 88% ($> 52\%$) were observed respectively. It is worth noting here that the aforementioned average performance also exceeded the one that a very conservative observer, i.e. an observer always judging "Fair" the quality of any PVS, would achieve. In fact, the expected correct and acceptable ratio for such an observer would be 20% and 60%, respectively.

To further investigate the accuracy of the 24 AIOs, i.e. the 24 NNs, we computed the ROC curves as well as the AUC indexes associated with each of the five alternatives (quality choices) predicted by the NN. The results are shown in Figure 6 and 7. For all the 24 observers, the curve associated with each possible alternative is above the 45 degree line, showing once more the superiority of the AIOs in terms of the perceptual quality evaluation over the observer rating at random. Furthermore, the values of the AUC index shown in Figure 7a and Figure 7b reveal that the 24 NNs are reasonably accurate since, on average, they reported AUC indexes ranging from 0.69 to 0.9. We finally notice the ability of NNs to more accurately model the observer's behavior in the context of the PVSs with the very low or high quality. In fact, higher AUC indexes are observed in the case of the "Bad" and "Excellent" opinions.

Despite we deem the MOS somehow limited as a single indicator of QoE, nevertheless it has been shown to be effective for some purposes, such as a comparison of codecs. For this reason, we also investigated the performance of

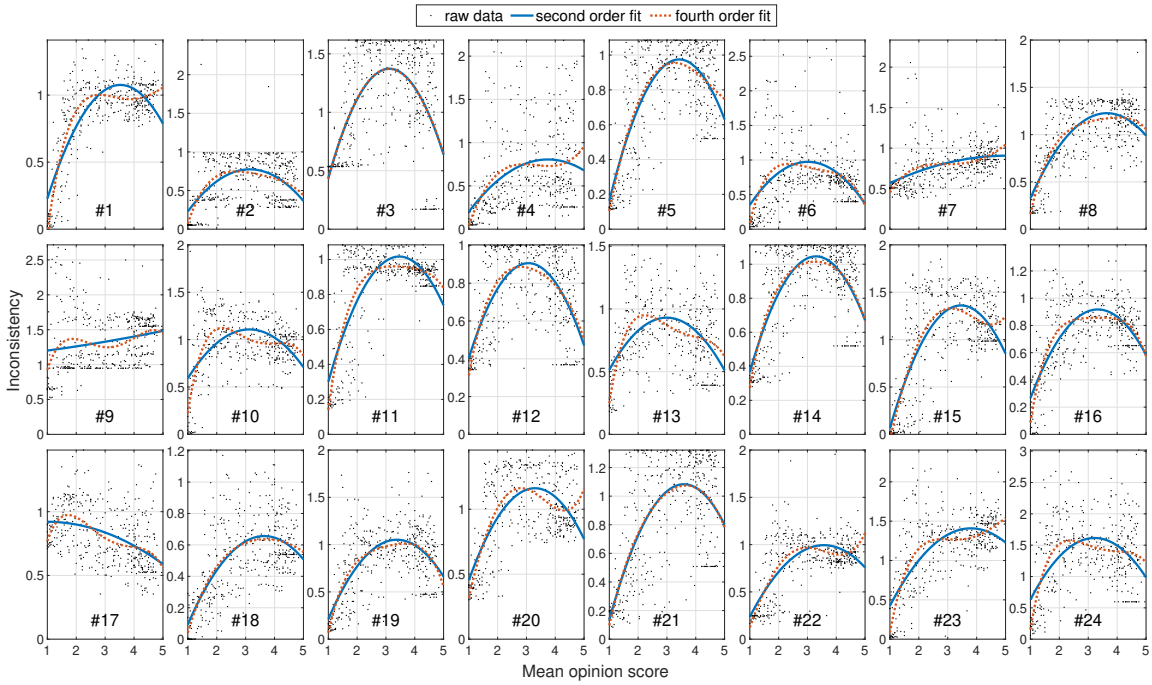


Fig. 12. Fitting of the inconsistency value with second (red) and fourth (yellow) order polynomials. Fitting functions tend to present an absolute maximum in the central part of the quality scale for almost all the observers, as expected by an inconsistency measure.

the AIOs when used to simulate a subjective experiment where the expected outcome is a MOS value for each PVS. We used the 24 AIOs, that were trained using the data from the VQEG-HD1 and the VQEG-HD5 experiments, to simulate a subjective test on the PVSs used in the VQEG-HD3 experiment. We then compared the mean of the AGOs (that we refer to as AI MOS in the following) and the standard deviation (AI SOS) to the actual MOS and SOS values. The results shown in Figure 8 are quite promising. In fact, very high correlation coefficients (0.91, 0.89) were obtained between the AI MOS and the MOS. The correlation coefficients (0.64 and 0.59) observed between the AI SOS and the SOS appear as a very promising result: in fact, we are not aware of any other algorithm, proposed in the literature, that can provide SOS predictions yielding correlation coefficients different than 0 with statistical significance.

Finally, we assess whether the deployed data augmentation method, i.e. the fusion of three similar observers into a single one, impacted negatively on the accuracy of the 24 AIOs, thus better results could have been expected/achieved by using many ratings coming from a single observer. To this aim, we trained other AIOs relying on the ITS4S dataset that provides 514 ratings for each of the 27 observers that participated in that experiment.

To train the 27 NNs, we used 7 categories and tested on the remaining two categories. Figure 9 presents the correct and acceptable prediction ratios for all the 27 AIOs. An average accuracy of 55% ($> 46\%$) and an average acceptable prediction ratio of 96% ($> 88\%$) have been obtained. Such results suggest that the performance discussed so far regarding the accuracy of NNs when used to model single observers can still be improved when a large number of ratings from the same observer is available.

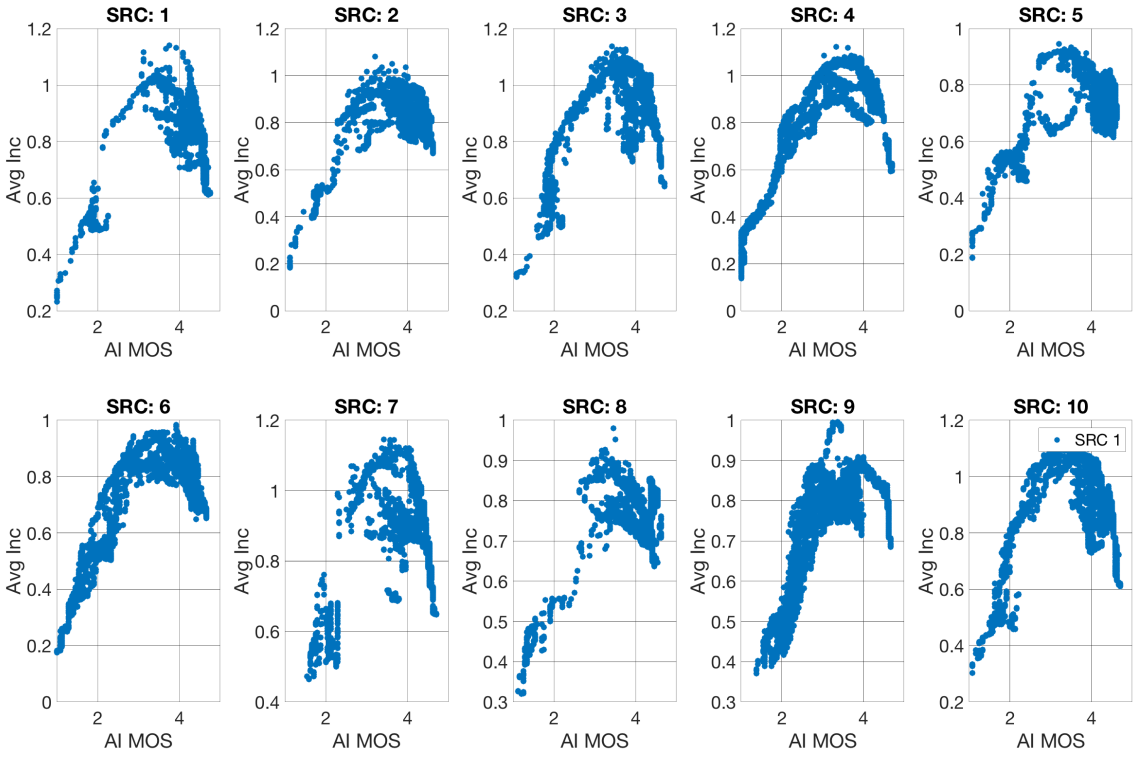
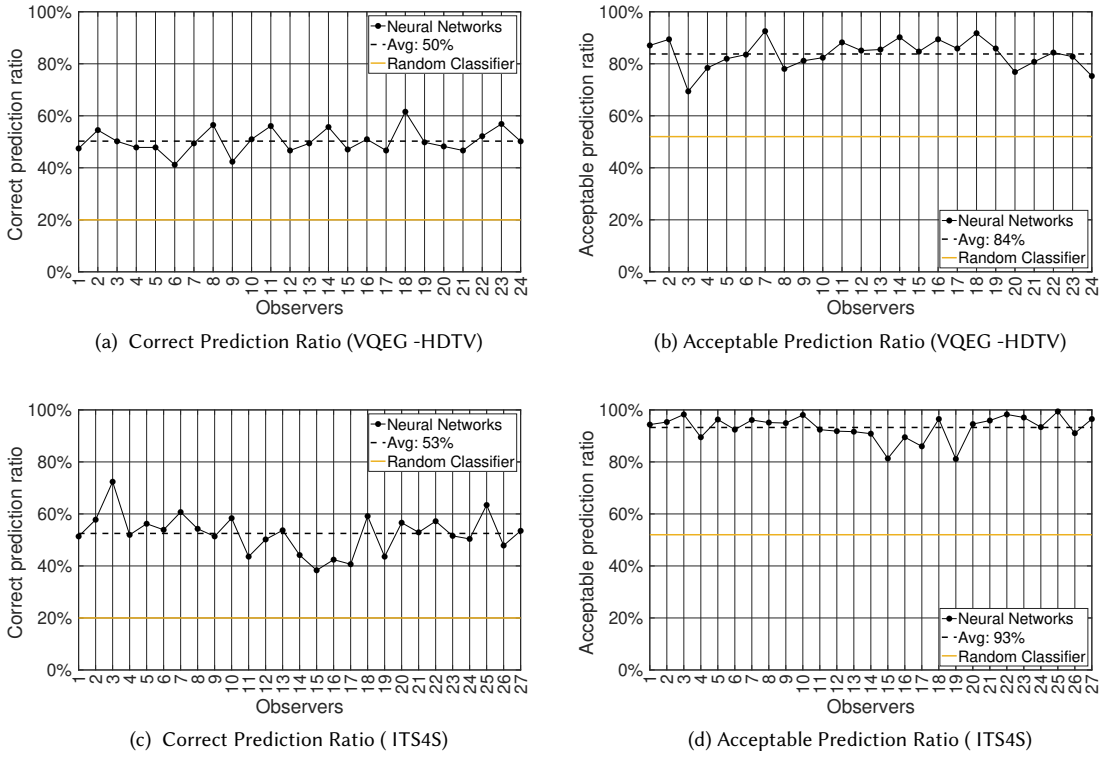


Fig. 13. The effectiveness of the proposed inconsistency measure on the large scale JEG-Hybrid dataset. The results show that low quality PVSs create less ambiguity (average inconsistency) for the AIOs independently from the SRC as it would have happened with real observers.

5.3 Subjects inconsistency

We now focus on the results related to the inconsistency measure introduced in (1). We remind that the AIO, given a PVS, produces not only a prediction of the opinion of the corresponding observer, but also a measure of its inconsistency as indicated in (1). In the experiments, we deployed the 24 AIOs on all the PVSs coming from the three VQEG-HD datasets. Hence for each observer, the value of its inconsistency for each PVS was also estimated. To analyze the properties of the proposed inconsistency measure as a function of the MOS, we employed the widely used mapping that assumes that the five alternatives on the ACR scale are equidistant. Therefore, in our analysis, with reference to the Eq 1, $v_i = i, \quad i = 1, 2, \ldots, 5$.

We then investigated the importance of each of the perceptual features considered in this work in predicting how inconsistent would be any observer rating a given PVS. The results, for each observer, are shown in Figure 10a, while the average importance of each feature is reported in Figure 10b. It can be noticed, for instance that the noise feature contribute less than others features to determine how repeatable the observer would be when rating the quality of a given PVS. The average features importance suggests that the presence of blur, flickering artifacts and the loss of blocks due to transmission has a major influence on the users' inconsistency and thus on its ability to repeat the same rating more times when evaluating the perceptual quality of a PVS. While a trivial relation was not found between the "Blur"



(a) Correct Prediction Ratio (VQEG -HDTV)

(b) Acceptable Prediction Ratio (VQEG -HDTV)

(c) Correct Prediction Ratio (ITS4S)

(d) Acceptable Prediction Ratio (ITS4S)

Fig. 14. Accuracy of the AIOs on the training sets. The performance is significantly higher than that of the randomly voting observer (yellow line). The obtained correct and acceptable ratios are not too much different from those obtained for the test sets (see Figure 5 and Figure 9 for more detail). None of the observers' behavior seems to be susceptible to have been particularly overfitted. Furthermore, the observed accuracy is close to the average expected one (57% as discussed in Section 4).

and "Flickering" features and the observers' inconsistency we observed that, on average, the inconsistency of almost all the observers decreases as the visual disturbance due to the amount of macroblocks lost due to transmission becomes more and more perceptible (see Figure 11). The Figure reports, for a given value of the "Blockloss" feature on the x axis, the average inconsistency evaluated on PVSs for which the "Blockloss" feature value is greater than or equal the one on the x axis. The decreasing trend of the curve of almost all the observers indicates that observers reliably recognize such a distortion.

We also studied the proposed measure of inconsistency as function of the perceptual quality of the PVS. Figure 12 reports the inconsistency of each observer on each PVS as a function of the MOS of the PVS. To better visualize the average trend from the points, we performed a least square fitting of the MOS to the inconsistency values using a second and fourth order polynomial function. It can be noticed, as expected, that almost all the observers are more consistent when evaluating PVSs with the very high or very low quality. Even using a fourth order polynomial function which allows the presence of local minimums, in almost all the cases the fitted curve still assumed the lower values only in correspondence to extreme values of the perceptual quality. We notice, however, that there are few observers, in particular observer #7 and #17, that tend to show higher inconsistency as the perceived quality increases.

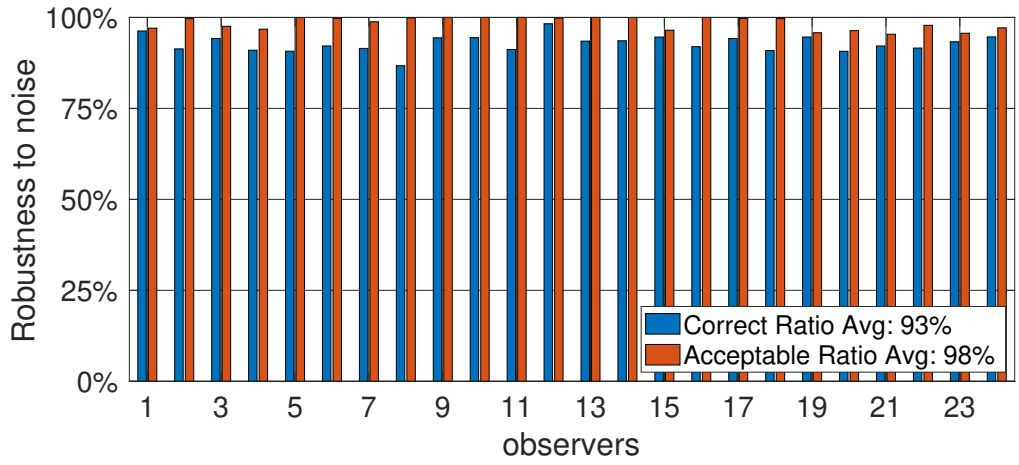


Fig. 15. Probability, for each AIO, that its output will not change (the correct ratio) or will change by at most 1 level on the ACR scale (the acceptable ratio) after adding, to each feature, a noise term which is uniformly distributed between -1% and 1% of the range of values assumed by such feature in the dataset.

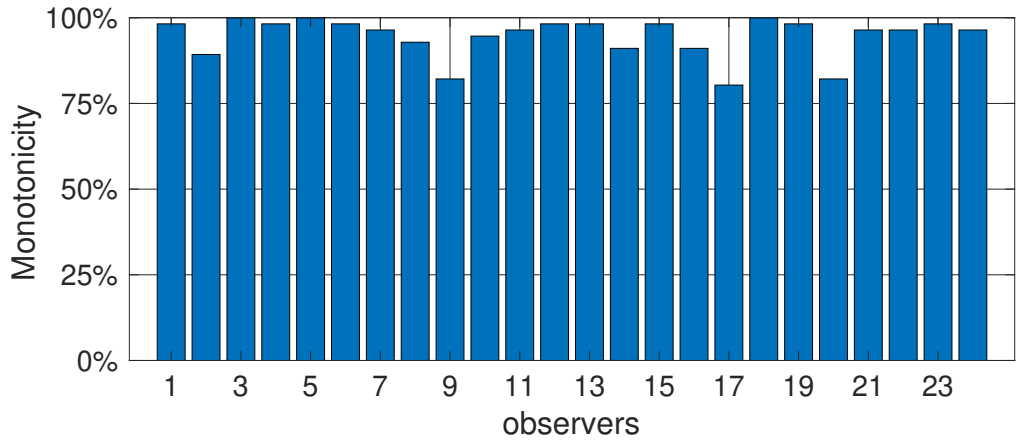


Fig. 16. Probability that each AIO would predict a higher score for the PVS encoded with a higher bitrate when assessing the visual quality of a pair of PVSs generated from the same SRC and affected by the coding artifacts only. The closer it is to 100 %, the better.

The 24 AIOs were deployed on the large scale JEG-Hybrid dataset that contains almost 60,000 (not subjectively annotated) PVSs, whose characteristics are explained in Table 1. For each PVS, the AGOs were computed and then averaged to obtain the AI MOS. The inconsistency of each AIO on each PVS was also computed. Figure 13 shows the relation between the AI MOS of each PVS and the average inconsistency of the AIOs on that PVS, separated by SRCs. The results revealed that the AIOs are able to mimic the higher consistency that characterizes real observers when rating PVSs with low perceptual quality. In fact, lower average inconsistency values were observed in correspondence to the PVSs reporting low values of AI MOS independently from the SRC.

5.4 Robustness of the AIOs

To investigate the robustness of the AIOs, we focused on three main aspects: i) Analyzing whether the performance of the AIOs on data never seen during the training is similar to that observed on the training set; ii) Studying the robustness of AIOs to the noise; iii) Assessing the ability of the AIOs to distinguish between two input video sequences with significantly different visual quality. For the second and the third aspect, the analysis was done with the 24 AIOs trained on the VQEG-HD dataset.

Figure 14 shows the performance of the AIOs on the training sets, while Figure 5 and Figure 9 report the performance on the test sets. For both the VQEG-HD and the ITS4S datasets, the performance of the AIOs on the training set is not significantly different from that observed on the test set. For instance, when it comes to the VQEG-HD dataset, the average of the correct ratios and the acceptable ratios on the training set were 50% and 84% respectively, whereas on the test set those ratios were 45% and 88%. This basically shows that the AIOs did not overfit the training set and can therefore be expected to generalize what was learned on the training set to a set of data never seen before. Moreover, it should be noted here that the performance obtained for the training sets is quite close to the expected average level of accuracy, i.e. 57%, as discussed in Section 4.

We also studied the robustness of the AIOs to the noise. For each AIO, we reported in Figure 15 the probability that its prediction will not change after adding, to each feature, a noise term which is uniformly distributed between -1% and 1% of the range of values assumed by such feature in the dataset. In practice, the noise term ranges from $-(M-m)/100$ to $(M-m)/100$, where m and M are respectively the smallest and the largest value assumed by that feature in the dataset. The probabilities in Figure 15 were obtained by simulating 10,000 realizations of the noise and counting the number of times in which the AIO does not change its prediction with respect to the noiseless case (the correct ratio) or change it at most by one level on the ACR scale (the acceptable ratio). Figure 15 shows that, on average, in 93% of the cases the AIOs provide a prediction equal to the one of the noiseless case. In 98 % of the cases the prediction changes by at most one level on the ACR scale. This result shows that the trained AIOs are rather robust to noise.

Finally, we considered the ability of the AIOs to distinguish between two stimuli, involving the same SRC, with different visual quality, that is, we assessed how monotone they are. We performed the analysis on the PVSs of the VQEG-HD dataset affected by the coding artifacts. We considered a pair of PVSs derived from the same SRC but encoded at different bitrates. For each AIO we computed the fraction of times when a lower score for the PVS encoded with a lower bitrate, as it is typically expected for PVSs based on the same SRC, was predicted. The results are summarized in Figure 16. On average, in 95% of the cases, the AIOs were, in general, able to effectively classify the input stimuli as expected, even though #2, #9, #17 and #20 seem a bit less monotone than the others.

6 CONCLUSION

In this work, we proposed a different approach to objectively evaluate media quality as perceived by the end users. In particular, we suggested to model every single observer through a neural network rather than predicting the MOS, differently from what has traditionally been done in the literature. Using the ratings of the observer gathered during a subjective experiment, we propose to train a neural network that can then be used as a substitute for that observer. In this way, we take into account and model the individual characteristics of each subject, such as personal expectations etc., which have a significant influence on the perception of quality. We illustrated the advantages of this new approach and also the flexibility it offers in evaluating the perceived media quality in different contexts. Finally, the computational results demonstrated both the feasibility and the effectiveness of the approach. In particular, we showed that neural

networks can be used to mimic with a good accuracy the choices of the single observer, allowing to estimate also how much confident he/she is in expressing his/her opinion on the quality of a PVS. Future directions include using a deep neural network approach to automatically construct, for each observer model, the set of features to be extracted from the PVSs instead of choosing among a predefined set of features as currently done in this work. Moreover, the AI observers approach could be used in specially designed extensive subjective experiments where each subject is willing to evaluate thousands of stimuli, therefore being able to create models that should be able to mimic those subjects with very high accuracy. This could be even more valuable if subjects are golden eyes: accurate models of such subjects could potentially be very valuable.

ACKNOWLEDGMENTS

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] C. G. Bampis, Z. Li, and A. C. Bovik. 2019. Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 8 (Aug 2019), 2256–2270. https://doi.org/10.1109/TCSVT.2018.2868262

[2] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnström, N. Staelens, and P. Le Callet. 2015. Objective video quality assessment-towards large scale video database enhanced model development. *IEICE Transactions on Communications* E98B, 1 (2015), 2–11. https://doi.org/10.1587/transcom.E98.B.2

[3] Jennifer Birch. 1997. Efficiency of the Ishihara test for identifying red-green colour deficiency. *Ophthalmic and Physiological Optics* 17, 5 (1997), 403–408.

[4] Kjell Brunnström et al. 2012. Qualinet white paper on definitions of Quality of Experience. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).

[5] ITU-R Rec. BT.500-11. 2002. Methodology for the subjective assessment of the quality of television pictures.

[6] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. 2011. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *IEEE Transactions on Broadcasting* 57, 2 (Feb 2011), 165–182.

[7] Cisco. 2020. Annual Internet Report: Growth in Internet users (2018–2023). https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[8] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[9] Lohic Fotio Tiotsop, Enrico Masala, Ahmed Aldahdooh, Glenn Van Wallendael, and Marcus Barkowsky. 2019. Computing Quality-of-Experience Ranges for Video Quality Estimation. In *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Berlin, Germany, 1–3. https://doi.org/10.1109/QoMEX.2019.8743303

[10] Lohic Fotio Tiotsop, Tomas Mizdos, Miroslav Uhrina, Peter Pocta, Marcus Barkowsky, and Enrico Masala. 2020. Predicting Single Observer's Votes from Objective Measures using Neural Networks. In *Proceedings of Human Vision and Electronic Imaging conference (HVEI)*. Society for Imaging Science and Technology (IS&T), Burlingame, CA, USA.

[11] Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala. 2020. Full Reference Video Quality Measure Improvement Using Neural Networks. In *Proc. Intl. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Barcelona, Spain, 2737–2741. https://doi.org/10.1109/ICASSP40776.2020.9053739

[12] Iris Galloso, Juan Palacios, Claudio Feijóo, and Asunción Santamaría. 2016. On the influence of individual characteristics and personality traits on the user experience with multi-sensorial media: an experimental insight. *Multimedia Tools and Applications* 75 (Feb 2016). https://doi.org/10.1007/s11042-016-3360-z

[13] Internet Media Group. 2019. Extension of the ITS4S Dataset. http://media.polito.it/its4s

[14] Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, and Sebastian Möller. 2016. QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Quality and User Experience* 1, 1 (Sep 2016). https://doi.org/10.1007/s41233-016-0002-1

[15] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, Mechelen, Belgium, 131–136. https://doi.org/10.1109/QoMEX.2011.6065690

[16] Qinghua Huang, Bisheng Chen, Jingdong Wang, and Tao Mei. 2014. Personalized video recommendation through graph propagation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10, 4 (2014), 1–17.

[17] Mansoor Hyder, Christian Hoene, and Noel Crespi. 2012. Are QoE Requirements for Multimedia Services Different for Men and Women? Analysis of Gender Differences in Forming QoE in Virtual Acoustic Environments. In *Intl. Multi Topic Conference on Emerging Trends and Applications in Information Communication Technologies (IMTIC)*, Vol. 281. Springer, Jamshoro, Pakistan. https://doi.org/10.1007/978-3-642-28962-0_20

[18] ITU-T Rec. G.100 Amd. 1. 2007. Definition of quality of experience (QoE).

[19] Lana Jalal, Matteo Anedda, Vlad Popescu, and Maurizio Murroni. 2018. QoE assessment for IoT-based multi sensorial media broadcasting. *IEEE Transactions on Broadcasting* 64, 2 (2018), 552–560.

[20] Lucjan Janowski and Zdzisław Papir. 2009. Modeling subjective tests of quality of experience with a Generalized Linear Model. In *International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, San Diego, CA, USA, 35–40. https://doi.org/10.1109/QOMEX.2009.5246979

[21] Lucjan Janowski and Margaret Pinson. 2015. The Accuracy of Subjects In A Quality Experiment: A Theoretical Subject Model. *IEEE Transactions on Multimedia* 17 (12 2015), 2210–2224. https://doi.org/10.1109/TMM.2015.2484963

[22] Hendrik Knoche and Martina Angela Sasse. 2009. The big picture on small screens delivering acceptable video quality in mobile TV. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 5, 3 (2009), 1–27.

[23] Jari Korhonen. 2019. Assessing Personally Perceived Image Quality via Image Features and Collaborative Filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, USA, 8169–8177.

[24] L. Krasula, Y. Baveye, and P. Le Callet. 2020. Training Objective Image and Video Quality Estimators Using Multiple Databases. *IEEE Transactions on Multimedia* 22, 4 (2020), 961–969.

[25] P. Le Callet, C. Viard-Gaudin, and D. Barba. 2006. A Convolutional Neural Network Approach for Objective Video Quality Assessment. *IEEE Transactions on Neural Networks* 17, 5 (Sep 2006), 1316–1327.

[26] Mikołaj Leszczuk, Mateusz Hanusiak, Mylene Farias, Emmanuel Wyckens, and George Heston. 2016. Recent developments in visual quality monitoring by key performance indicators. *Multimedia Tools and Applications* 75 (2016), 10745–10767. https://doi.org/10.1007/s11042-014-2229-2

[27] Z. Li and C. G. Bampis. 2017. Recover Subjective Quality Scores from Noisy Measurements. In *Data Compression Conference (DCC)*. IEEE, Snowbird, UT, USA, 52–61.

[28] Torrin M. Liddell and John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79 (Nov 2018), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

[29] Karan Mitra, Arkady Zaslavsky, and Christer Ahlund. 2015. Context-Aware QoE Modelling, Measurement and Prediction in Mobile Computing Systems. *IEEE Transactions on Mobile Computing* 14 (May 2015), 920–936. https://doi.org/10.1109/TMC.2013.155

[30] Jim Mullin, Lucy Smallwood, Anna Watson, and Gillian Wilson. 2001. New techniques for assessing audio and video quality in real-time interactive communications. In *IHM-HCI Tutorial*. Lille, France.

[31] Anja B. Naumann, Ina Wechsung, and Jörn Hurtienne. 2010. Multimodal interaction: A suitable strategy for including older users? *Interacting with Computers* 22, 6 (Nov 2010), 465–474. https://doi.org/10.1016/j.intcom.2010.08.005

[32] Netflix. 2019. VMAF - Video Multi-Method Assessment Fusion. https://github.com/Netflix/vmaf.

[33] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15 (2010), 625–632.

[34] ITU-T Rec. P.910. 2008. Subjective video quality assessment methods for multimedia applications.

[35] Joana Palhais, Rui S. Cruz, and Mário S. Nunes. 2012. Quality of Experience Assessment in Internet TV. In *Proc. Intl. Conf. on Mobile Networks and Management*. Springer, Aveiro, Portugal, 261–274.

[36] Margaret H. Pinson. 2018. ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes. NTIA, Technical Memo TM-18-532.

[37] Mijke Rhemtulla, Patricia Brosseau-Liard, and Victoria Savalei. 2012. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods* 17, 3 (2012), 354–373.

[38] M. Ries, O. Nemethova, and M. Rupp. 2007. Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming. In *2007 2nd International Symposium on Wireless Pervasive Computing*. IEEE, San Juan, Puerto Rico. https://doi.org/10.1109/ISWPC.2007.342629

[39] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea. 2016. Do Personality and Culture Influence Perceived Video Quality and Enjoyment? *IEEE Transactions on Multimedia* 18, 9 (2016), 1796–1807.

[40] M. Seufert. 2019. Fundamental Advantages of Considering Quality of Experience Distributions over Mean Opinion Scores. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Berlin, Germany, 1–6. https://doi.org/10.1109/QoMEX.2019.8743296

[41] H. R. Sheikh and A. C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (Feb 2006), 430–444.

[42] Robert C. Streijl, Stefan Winkler, and David S. Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (Mar 2016), 213–227. https://doi.org/10.1007/s00530-014-0446-1

[43] Stevens Sue. 2007. Test distance vision using a Snellen chart. *Community Eye Health* 20, 63 (2007), 52.

[44] Domonkos Varga. 2019. No-Reference Video Quality Assessment Based on the Temporal Pooling of Deep Features. *Neural Processing Letters* 50, 3 (12 Apr 2019), 2595–2608. https://doi.org/10.1007/s11063-019-10036-6

[45] Domonkos Varga, Dietmar Saupe, and Tamás Szirányi. 2018. Deeprn: A Content Preserving Deep Architecture for Blind Image Quality Assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, San Diego, CA, USA, 1–6. https://doi.org/10.1109/ICME.2018.8486528

[46] VQEG. 2010. Report on the Validation of Video Quality Models for High Definition Video Content (v. 2.0). http://bit.ly/2Z7GWDI

[47] Z. Wang, E. P. Simoncelli, and A. C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, Vol. 2. IEEE, Pacific Grove, CA, USA, 1398–1402. https://doi.org/10.1109/ACSSC.2003.1292216

[48] Ina Wechsung, Matthias Schulz, Klaus-Peter Engelbrecht, Julia Niemann, and Sebastian Möller. 2011. All Users Are (Not) Equal - The Influence of User Characteristics on Perceived Quality, Modality Choice and Performance. In *Proc. IWSDS Workshop on Paralinguistic Information and its Integration in Spoken Dialogue Systems*. Springer, New York, NY, USA, 175–186. https://doi.org/10.1007/978-1-4614-1335-6_19

[49] Xiaochi Wei, Heyan Huang, Liqiang Nie, Fuli Feng, Richang Hong, and Tat-Seng Chua. 2018. Quality Matters: Assessing cQA Pair Quality via Transductive Multi-View Learning. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*. Stockholm, Sweden, 4482–4488.

[50] S. Winkler and P. Mohandas. 2008. The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. *IEEE Transactions on Broadcasting* 54, 3 (Sep 2008), 660–668. https://doi.org/10.1109/TBC.2008.2000733

[51] L. Xu, W. Lin, L. Ma, Y. Zhang, Y. Fang, K. N. Ngan, S. Li, and Y. Yan. 2016. Free-Energy Principle Inspired Video Quality Metric and Its Use in Video Coding. *IEEE Transactions on Multimedia* 18, 4 (Feb 2016), 590–602.

[52] B. Yan, B. Bare, and W. Tan. 2019. Naturalness-Aware Deep No-Reference Image Quality Assessment. *IEEE Transactions on Multimedia* 21, 10 (Mar 2019), 2603–2615.

[53] Wei Yang, Kuanquan Wang, and Wangmeng Zuo. 2012. Neighborhood Component Feature Selection for High-Dimensional Data. *Journal of Computers* 7 (Jan 2012), 161–168.

[54] J. You and J. Korhonen. 2019. Deep Neural Networks for No-Reference Video Quality Assessment. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, Taipei, Taiwan, 2349–2353.

[55] Hui Zeng, Lei Zhang, and Alan C Bovik. 2017. A probabilistic quality representation approach to deep blind image quality prediction. (2017). arXiv:arXiv:1708.08190v2

[56] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. 2020. Objective Video Quality Assessment Combining Transfer Learning With CNN. *IEEE Transactions on Neural Networks and Learning Systems* 31, 8 (2020), 2716–2730. https://doi.org/10.1109/TNNLS.2018.2890310

[57] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (Apr 2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[58] Yi Zhu, Sharath Chandra Guntuku, Weisi Lin, Gheorghita Ghinea, and Judith A Redi. 2018. Measuring individual video QoE: A survey, and proposal for future directions using social media. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2s (2018), 1–24.