



Doctoral Dissertation
Doctoral Program in Computer and Control Engineering ($34^{th}$ cycle)

# Optimizing Perceptual Quality Prediction Models for Multimedia Processing Systems

## Going Beyond the Mean Opinion Score

**Lohic Fotio Tiotsop**
* * * * * *

**Supervisor**
Prof. Enrico Masala, Supervisor
Prof. Antonio Servetti, Co-supervisor

**Doctoral Examination Committee:**
Prof. Narciso García, Referee, Universidad Politécnica de Madrid, Madrid, Spain.
Prof. Davide Quaglia, Referee, Università di Verona, Verona, Italy.
Prof. Marco Cagnazzo, Università di Padova, Padova, Italy.
Prof. Guido Marchetto, Politecnico di Torino, Turin, Italy.
Prof. Alexander Raake, Technische Universität Ilmenau, Ilmenau, Germany.

Politecnico di Torino
March 03, 2022

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Lohic Fotio Tiotsop

Turin, March 03, 2022

# Summary

Being able to accurately predict human perceived quality of multimedia content (e.g., images and videos) through computer algorithms is an important and hot research topic, since it directly leads to several advantages: storage capacity and bandwidth savings for large organizations, and possibility to offer higher quality services when the communication channel capacity is limited. This PhD thesis focuses on using machine learning algorithms and advanced statistical methods to find new ways of addressing the problem.

Human viewers can easily distinguish between an image with a pleasant perceptual quality and another one whose quality is corrupted by artifacts introduced for instance by a codec. This task that is straightforward for a human due to his daily experiences and background, is absolutely not easy for an algorithm. For this reason, in the literature, human opinion scores on the perceptual quality of a multimedia content are considered as ground truth data. Any quality prediction algorithm (QPA) is then designed to predict the so-called Mean Opinion Score (MOS) of a multimedia content, i.e., the average of the scores that a group of human viewers would express if they were asked to watch that content and rate its perceptual quality on a given numerical scale.

Several QPAs aiming at MOS prediction have been proposed in the literature but the state-of-the-art in media quality assessment is still suffering some crucial limitations. Some of these limitations can be summarized as follows: i) none of the existing QPAs is accurate in all application scenarios, and thus there is still large room for their improvement; ii) QPAs, as estimators of the MOS, do not provide a complete measure of quality-of-experience (QoE). In fact, QPAs simply measure the quality perception of an "average viewer" and only marginally account for the individual expectation and uncertainty that characterize different users of the same service. For instance, QPAs do not provide answers to the following questions: what is the percentage of final users that would not be satisfied by the perceptual quality of a given video sequence? What are the characteristics of the final users that are expected to not be satisfied? The answers to these two questions are obviously of primary importance for any company that produces and markets multimedia

content. There is therefore a crucial need to go beyond the MOS as suggested by recent publications in the literature.

This thesis contributes to advancing the state-of-the-art by proposing some solutions to cope with the aforementioned state-of-the-art limitations. Such proposals can be summarized under three main items.

- **The perceptual quality as a random variable**: A large number of stochastic influence factors, e.g. the subject's emotional state and the way each subject interprets the quality scale, that are not under the control of the test designer, influence the scores of a human subject during a subjective test. To jointly account for the inaccuracy of existing QPAs and these stochastic influence factors, both the prediction of a QPAs and the MOS are defined as random variables. A probabilistic approach is proposed to find the quality range to which the perceptual quality of a given processed video sequence (PVS) is expected to belong with a user specified probability.
- **Measuring the reliability of a MOS prediction**: since existing QPAs are not always accurate, integrating their prediction with an index that informs on how reliable the predicted quality score is, acquires a significant importance. Two different approaches to cope with this issue are proposed in this thesis. The first one is based on machine learning and it aims at predicting the intrinsic ability of a PVS to confuse human viewers and hence the QPAs as they are trained to predict human ratings. The second approach is instead based on the level of disagreement between many different QPAs when used to assess the quality of a given PVS. In fact, an index to measure the disagreement of QPAs is proposed and it is shown that such an index allow to distinguish between the cases in which QPAs are accurate and when they are not.
- **Artificial Intelligence-based observers**: finally this thesis presents a more complete approach to quality assessment. Unlike traditional approaches, the proposed one allows to fully consider individual expectations when automatically assessing the perceptual quality of multimedia content. Instead of predicting the MOS for each content under evaluation, as it is traditionally done in the literature, a different direction is proposed. An artificial neural network is trained to mimic an individual observer in terms of quality perception. Such a neural network can then be considered as a "virtual observer", and it is called an artificial intelligence-based observer (AIO). A large number of AIOs can then be trained, each representing an actual observer with well known characteristics and expectations. The advantage of this approach is to be able to more accurately model the distribution of the opinion scores

that form the quality prediction as well as the uncertainty that intrinsically characterizes human viewers. The proposed approach therefore allows to: i) predict the percentage of viewers that would not appreciate the perceptual quality of a given processed content; ii) make inference on the characteristic of the unsatisfied viewers; iii) perform the simulation of subjective tests.

# Acknowledgements

Having reached the end of this beautiful three-year journey, I feel extremely grateful to many persons and organizations who have made it possible to achieve my goal. I would therefore like to dedicate the next few paragraphs to thank them even though I am aware that it would take a whole book to express all the gratitude I have towards them.

First of all, I would like to thank Prof. Enrico Masala, my supervisor. His acquaintances have always been a beacon that continually assured me by showing me the direction to follow in periods of darkness. I couldn't have hoped for a better guide. Despite his busy schedule, he has always made himself available, encouraging and ready to provide any advice that can help me both in the present and in general in the future. He offered me the opportunity to interact with the international scientific community and be part of an important network. This allowed me to disseminate my research activities and collaborate with international researchers with great expertise. This also helped me to contribute to several research papers and projects while improving my capacity to efficiently and effectively interact with the audience.

Secondly, I would like to express my deep gratitude to Prof. Antonio Servetti, my co-supervisor. Without his acquaintances and his unconditional willingness to help me in any situation, the results on which this thesis strongly depends would never have been achieved. I would also like to thank him for his patience and the fundamental advice he provided me with during the last three years. They will remain an asset that I will use throughout my career and that I will proudly pass on to those who are younger generations.

My gratitude is also strongly addressed to the Polito Interdepartmental Centre for Service robotics (PIC4SeR) that initially financed my PhD scholarship. I would also like to thank all my colleagues in PIC4SeR for the constructive discussions we have had on the topic of precision agriculture. This helped me to better perceive the context and allowed me to advance the state-of-the-art through the following journal paper [126].

I would then like to address special thanks to all my international collaborators. In particular, I am really grateful to Prof. Marcus Barkowsky, Prof. Peter Pocta, Dr. Tomas Mizdos, Prof. Glenn Van Wallendael, Dr. Ahmed Aldahdooh and Dr. Florence Agboma. Their expertise has helped me a lot throughout the path to this thesis.

I could not terminate these acknowledgments without mentioning the following members of the operation research and optimization group of the control and computer engineering department of the Politecnico di Torino: Prof. Roberto Tadei, Dr. Daniele Manerba and Dr. Edoardo Fadda. I sincerely thank them for their collaboration and support.

*I would like to dedicate this thesis to my wife and my sons. Their unconditional love and support have been for me an inexhaustible source of motivation.*

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The volume of multimedia content (mostly images and videos) exchanged over the internet has significantly increased in the last decades [131]. This has been made possible in part thanks to the development of more effective multimedia processing systems (MPSs). MPSs are designed to perform many tasks, i.e., capturing, compression, transmission, restoration, enhancement and reproduction of multimedia content. For any of these tasks the MPS is expected to deliver a content with the best possible perceptual quality (few visible artifacts) under the constraints imposed by the availability of resources such as the bandwidth and the storage capacity [3]. In other words, MPSs constantly solve optimization problems whose objective functions involve a measure of perceptual quality. For this reason, tools that enable the assessment of the perceptual quality represent a fundamental component for any MPS.

## 1.1 Traditional Approaches to Media Quality Assessment

The question of how to assess the perceptual quality of a processed multimedia content has been and continues to be of large interest for researchers as witnessed by these very recent papers [132, 76, 150]. In particular, two main approaches have been investigated. The first one, is known as "subjective assessment" and the second approach is referred to as "objective assessment".

### 1.1.1 Subjective Quality Assessment

Human viewers are in general the final users of processed multimedia content. The more natural approach to media quality assessment therefore consists of asking human subjects to score the perceptual quality of a processed content. This approach is known as the subjective assessment. The subjective assessment of

the perceptual quality of a multimedia content is usually obtained by designing a subjective test. A subjective test consists of inviting a group of human viewers, typically in a controlled environment, to watch a number of multimedia content and rate their perception of artifacts on a given scale. At the end of the test, the quality of each evaluated multimedia content can be rated by pooling the viewer ratings on it. The most common pooling strategy is the arithmetic average that leads to the so-called Mean Opinion Score (MOS) of the content.

A large number of ITU-T recommendations have been produced to define some formal guidelines that should be followed when designing a subjective test [92, 103]. The aim of these recommendations is to enhance the reproducibility of the results. Nonetheless, there is still not a common agreement on what is the best quality scale and the assessment method to be used, since each has advantages and drawbacks depending on the application under investigation [99, 47, 121].

For a complete discussion on the existing and standardized methods, the interested readers might refer to [125]. Here, just two methods that will later be considered in this thesis are briefly described.

- **The Absolute Category Rating (ACR)**: the ACR is the most widespread approach within the research community. The viewers are provided with a five points quality scale reporting the following options: "Bad", "Poor", "Fair", "Good" and "Excellent". They are then asked to watch each content and choose the option that matches their perception of its quality. The five points scale is usually mapped to the integers numbers from 1 to 5, i.e., "Bad"=1, "Poor"=2 and so on, in order to get numeric scores and hence compute the MOS;

- **The Degradation Category Rating (DCR)**: in this case the viewer is provided with the so-called Double Stimulus Impairment Scale (DSIS). The DSIS is a five points scale with the following options: "Very annoying", "Annoying", "Slightly annoying", "Perceptible but not annoying" and "Imperceptible". The viewer is then asked to watch first the reference unimpaired source content, then the processed one, and score how annoying he/she found the artifacts in the processed content with respect to the reference one using the DSIS. Just like in the ACR case, the five options are usually mapped to integers from 1 to 5 in order to compute the MOS of the evaluated content.

The efforts of the media quality assessment community to enhance the reproducibility of the results of subjective tests have yielded some very interesting results. For instance, when running a subjective test with the same set of stimuli in two different labs, even though the obtained MOS scores are usually not equal in absolute terms due to the potential offset/bias caused by each test' context influence factors (IFs), they typically show a very high rank and linear correlation [100]. In other words, humans can consistently order the impairment levels during a subjective

test even if they score them differently in different contexts. Therefore, the results of a subjective test conducted under the ITU-T recommendations are usually considered as reliable.

Despite the reliability of the subjective approach to quality assessment, it suffers some crucial drawbacks that prevent its deployment in many applications. For instance, human viewers assess the quality in a time frame that would not enable a real time monitoring of the perceptual quality during a live streaming section. Also, to fully compare two MPSs, it might be necessary to evaluate the perceptual quality of thousands of content processed by each one of them. Such a task would require an eternity if one relies on human viewers. In short, while human viewers are reliable, they are very inefficient, this is one of the reasons why an alternative approach to media quality assessment, based on algorithms, has been largely investigated.

### 1.1.2 Objective Quality Assessment

The objective approach to media quality assessment aims at the development of QPAs that can predict the perceptual quality as perceived by the human viewers. In this context, the result coming from the subjective assessment is typically considered as the gold standard and QPAs are designed to predict the MOS as accurately as possible. An impressive number of papers proposing different QPAs has been published in the last decades [140, 2, 11].

Depending on the information required as input, the QPAs can be classified in three main classes.

- **Full Reference (FR)**: FR QPAs require as input both the source/reference unimpaired content and the processed one. The quality of the reference content is assumed to be the desired one. The quality of the processed content is then assessed by measuring how degraded it is with respect to that of the reference one. Several approaches to measure such a degradation have been proposed. The most basic one consists of simply relying on the euclidean distance at the pixel level between the processed and the reference content; this yields to the so-called Peak Signal to Noise Ratio (PSNR) [147]. The PSNR has been shown to not correlate very well with human scores, for this reason, more elaborated ideas have been proposed. In [115, 9], authors relied on natural scene statistics; they compared the characteristics of the processed content with those of a natural scene. The authors in [35] proposed a FR QPA that measure the degradation of the quality of the processed content by modeling some characteristics of the human visual system (HVS). In [155, 156], machine learning-based approaches are investigated. In general FR QPAs are more accurate than the other ones. However, the need to have available the reference signal in order to run the algorithm represents a severe drawback, since for some applications the source content is not directly accessible. FR

QPAs are therefore more indicated for offline tasks such as codecs comparison.

- **Reduce Reference (RR)**: RR QPAs do not require the availability of the whole reference content; they make use only of some features of it. Only those features have to be accessible in order to measure the degradation of the perceptual quality of the processed content with respect to the reference one. RR QPAs therefore maintain reasonable the amount of reference information to store and/or transmit. This makes them effective for some online applications. It is however worth noting that the loss of details on the reference signal typically occurs at the expense of accuracy in terms of MOS prediction. Some examples of RR QPAs can be found in [106, 73, 137].

- **No Reference (NR)**: NR QPAs are algorithms that focus only on the processed content. The quality in this case is evaluated in absolute terms as the reference signal is not used. The lack of information on the source content makes the development of accurate NR QPAs very challenging. They are therefore usually designed for assessing the presence and level of annoyance of a specific type of artifact. NR QPAs are in general less accurate than the other QPAs, however, they offer more flexibility in practice as one does not need to handle the reference signal. With the success of machine learning, NR QPAs have received particular attention. An impressive number of authors, [80, 66, 16], proposed novel models regressing features extracted from the processed content on the quality scale to obtain machine learning-based MOS prediction. Despite the lack of large scale subjectively annotated datasets, some authors [134] succeeded by smartly using deep neural networks (DNNs) that are however known to be demanding in terms of training samples.

QPAs are obviously faster than human viewers. Furthermore, unlike human viewers that one needs to find and convince or pay so that they accept to participate in a subjective test, QPAs are directly accessible on demand. Unfortunately, the state-of-the-art in the development of QPAs is still suffering some crucial limitations that are discussed in the next section.

## 1.2 Some Limits of State-of-the-Art Approaches

This section presents some of the fundamental problems still open in the media quality assessment community. The list of problems discussed here is not intended to be exhaustive, but rather an introduction to the hot research questions to which this PhD thesis contributes to finding preliminary answers.

### 1.2.1 Accounting for Human Viewers' Inconsistency

Human viewers are not always able to repeat their previous opinion score when asked to rate again the perceptual quality of a content they have previously evaluated [55]. This characteristic of the human viewers is referred to as the subject's inconsistency in the context of media quality assessment. Because of the subject's inconsistency, if a subjective test is run twice in exactly the same conditions, i.e., with the same stimuli, the same viewers and the same environment, there is a really low probability to get for each content the same MOS scores in both tests.

The subject's inconsistency is typically caused by several IFs, some of which are not under the control of the subjective test designer. For instance, one might think about the viewer's emotional state and his/her ability to maintain concentration for a long time. In any case, these IFs make the MOS a stochastic quantity as the same test conditions would lead to different values. Unfortunately, the uncertainty that characterizes the MOS of a given content has long been disregarded. The MOS has instead been traditionally treated as a deterministic value for the sake of convenience. This directly impacts on the accuracy of QPAs as they are trained to predict a ground truth quality score that is supposed to be affected by some noise in practice.

Several publications [79, 54, 44, 122, 114] have underlined the need to design more comprehensive objective quality assessment approaches that account for the stochastic IFs that affect the human perception and judgment of the quality. However the problem still remains open since there is large room for improvement of existing approaches.

### 1.2.2 Predicting Multimedia Content Ambiguity

The complexity of a multimedia content, in terms of scene characteristics for instance, is a determining factor in the ability to accurately predict its perceptual quality. For example, at the same compression level, a video showing a scene with fast moving objects in a landscape full of details might appear to a human of better quality than another video with almost static content.

For this reason, existing QPAs usually have an accuracy that very much depends on the type of content they are requested to evaluate. They are not therefore accurate in all situations. A question of interest to researchers is therefore that of being able to determine the ability of a video sequence to mislead humans and/or QPAs on its perceptual quality. This is of crucial importance, since once identified, problematic multimedia content could benefit larger resources to avoid potential degradation of final users' quality-of-experience (QoE).

In [55, 68] approaches to estimate the ambiguity of a multimedia content, i.e., its ability to confuse human viewers, were proposed. Unfortunately these approaches can be used only to estimate the level of ambiguity of a video sequence that has

already been subjectively evaluated. This is a severe limitation, as subjective tests are time demanding and the approach can not be used for real time monitoring of the quality. Therefore the question of how to objectively predict the problematic nature of a content from the point of view of quality assessment is still open.

### 1.2.3 The Diversity in Users' Expectation

QPAs 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 [50] 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 [21]: "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 as personal experience contribute to determine his/her opinion regarding the perceptual quality of a media and should therefore be considered as much as possible when predicting the QoE.

While subjective evaluation methods account for the diversity in users' expectations up to some extent, it is typically disregarded by the traditional objective approaches to quality assessment. In fact, subjective tests 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 [86], user's gender [48] or the user's preferences [93] that can have an impact on the subject's judgment. The effect of such human IFs is represented by the diversity typically observed between the individual opinions gathered during the test.

On the contrary, the traditional objective approaches to media quality assessment fundamentally focus, for the sake of convenience, on predicting a single value, i.e., the MOS. To compute the MOS, individual opinion scores coming from a subjective test are averaged. This pooling step keeps out of the assessment process the diversity among the users' opinion scores and thus the information related to the personality of the viewers and their expectations.

The media quality assessment community is therefore still lacking effective and efficient objective approaches to assess the perceptual quality and in general the final users' QoE while accounting for the diversity in terms of expectations, background and experience. Such approaches would give answers to the following questions that are of primary importance for streaming vendor companies: i) what is the percentage of customers that would be satisfied if a given content is encoded

in a certain manner? ii) What are the characteristics of the customers that are expected not to be satisfied with the quality of a processed content?

### 1.2.4 The Need for Appropriate Data Augmentation Approaches

Obtaining large scale subjectively annotated datasets is very expensive and time consuming. This makes it difficult to effectively train deep NN-based QPAs for instance. Data augmentation approaches and algorithms have therefore been naturally adopted by the media quality assessment community.

The computer vision community has developed a number of data augmentation approaches [119]. Most of these approaches represent an implementation of a set of rules that, applied to an entity of the training set (image, video, audio), creates an additional entity that is expected to have the same label. For example, in an image classification task, a translation, rotation, and scaling of the objects in an image does not change its content and therefore its label.

As it can be seen from the latter example, the typical data augmentation approaches mainly affect the geometry of the elements present in a multimedia content. While this type of modification can generate particularly challenging samples from the point of view of the computer vision tasks, they may not constitute significant added value for the training of a model aimed for predicting visual quality. In fact, a modification of geometrical shape of the objects alone keeps constant features such as contrast, resolution, spatial and temporal activity and also quantity of motion (in case of video), which are important for visual quality assessment. The model could therefore perceive these new samples as substantially equivalent to the initial one from which they are generated.

For this reason, alternative approaches for generating more data in order to effectively train machine learning based QPAs have been proposed. In [59, 75], in addition to the subjectively annotated training samples, the authors included new training samples for which they used the output of a QPA as a substitute of the MOS. The main drawback of this approach comes from the fact that one does not know a priori on which content the score provided by the used QPA is not accurate. Hence the obtained dataset might be very noisy.

The authors in [62, 81], instead, proposed an approach to combine different subjectively annotated datasets into a single larger one, thus overcoming the issues stemming from the different contexts in which the subjective experiments have been conducted. In particular, the MOS values had to be realigned to take into consideration the context influence factors that affect the result of each experiment [98]. Therefore, the MOS values in the newly created dataset are, in practice, only an estimate of the ones that would be expected while running a single large scale subjective experiment.

Although some authors have focused on the the question of how to effectively

augment the number of training samples in media quality assessment, the question remains open as existing approaches suffers some limitations.

## 1.3 Going Beyond the MOS: the Thesis Contribution

This PhD thesis advances the state-of-the-art in media quality assessment by proposing some approaches to cope with the open problems discussed in Section 1.2. In particular, the thesis contribution can be mainly summarized under the three main points presented in this section. Also, for completeness sake, a list of all the scientific papers I coauthored and published during the PhD period is provided in Appendix A, while a short description of all the used datasets and the motivation behind their choice is provided in Appendix B.

### 1.3.1 The Perceptual Quality as a Random Variable

To account for the large number of IFs that affect the results of a subjective test and are not under the control of the test designer, I proposed in [29] to consider the MOS as a random variable instead of looking at it as a single deterministic value.

I therefore proposed to express the quality of a given video sequence in probabilistic terms. More precisely, I proposed an approach to compute an interval to which the MOS of a given video sequence is expected to belong with a user's specified probability. Note that this is very different from the concept of confidence intervals (CIs) that can be computed only after a subjective test and for which a deterministic estimate of the MOS is still required.

The proposed approach relies on well known FR QPAs. The outputs of the considered FR QPAs are considered as random variables. This is to account for the fact that existing QPAs are not accurate in all cases and might provide different quality predictions for two PVSs that are expected to have the same quality score. A Gaussian mixture model is then used to fit the joint probability distribution of each FR QPA and the MOS. From the fitted joint probability distribution, the conditional probability distribution of the MOS to the score of the considered QPAs is computed. Finally the quantiles corresponding to the user's specified probability are derived from the conditional distribution of the MOS to get the desired interval.

In addition to the fact that the approach has the advantage of taking into account the uncertainty in the quality evaluation process, as I have shown in [29], it also allows to make a more complete evaluation of the accuracy of QPAs and it can be very useful to identify peculiar stimuli in large scale not subjectively annotated datasets.

### 1.3.2 Integrating the MOS with Measures of Reliability

As pointed out in Section 1.2, one of the problems of major interest for the quality assessment community is to automatically recognize video sequences that are likely to mislead QPAs or human viewers when scoring their quality.

I have shown in [128] that some well known and widely used QPAs are likely to overestimate the perceptual quality when used on video sequences with few spatial details. Such QPAs do not therefore perfectly model the fact that the visibility of artifacts is emphasized by the absence of details. I then proposed a neural network-based improvement of these QPAs in [127].

The standard deviation of the opinion scores (SOS) of viewers is generally considered by the media quality assessment community as a measure of the reliability of the MOS. In [30], I highlighted two main sources of noise that affect the SOS directly computed from the result of a subjective test conducted with a limited number of viewers. I then introduced the so-called ground-truth SOS (gtSOS). The gtSOS of a content is intended to be the standard deviation of the opinion scores that would be observed if a very large number of human viewers (ideally infinite) was asked to score its quality using a continuous quality scale. Therefore a large gtSOS stands for higher complexity of the content in terms of quality prediction. I have shown that the gtSOS can be predicted from the outputs of many different QPAs. I then trained shallow NNs that can perform such a prediction.

Being able to predict the gtSOS allows to distinguish between the intrinsic complexity of a video sequence in terms of quality assessment, i.e., its ability to confuse human viewers and/or QPAs, and the diversity among the user's opinion scores that derive only from potential anomalies or noise in the evaluation process. I therefore also proposed in [30] a statistical model of the SOS that takes into account the gtSOS and some noise terms. I showed that such a model is very useful for identifying stimuli whose evaluation would deserve further attention after a subjective test.

In [130], instead, I proposed a more intuitive approach to identify peculiar PVSs for QPAs. In fact, instead of relying on NNs that are black box models, I introduced an index based on the level of disagreement between the quality scores predicted by different QPAs. Through the results of a small scale subjective test conducted to assess the performance of the proposed index, I have shown that it allows to effectively distinguish between PVSs on which QPAs are expected to be accurate and those on which they are likely to wrongly predict the quality.

### 1.3.3 Mimicking Individual Quality Perception

I have proposed in [32] to train a NN that can mimic an individual human viewer in terms of quality perception. Once trained, this NN can receive as input an image/video and predict with which probability the related human viewer would

choose any of the five options on the ACR scale if asked to rate the quality of the same content. Such a probabilistic output allows to also model the subject's inability to repeat his/her previous ratings when evaluating the same PVS more than once. The idea is therefore to use the trained NN as a substitute of the human viewer it is trained to mimic. For this reason I called such a NN an "Artificial Intelligence-based Observers" (AIO).

I then suggested training several different AIOs, each modeling a human viewer with particular characteristics. This allowed me to propose a more complete approach to objective quality assessment. In fact, the use of the AIOs goes beyond the prediction of the MOS. Given a PVS as input to the AIOs, they produce in output a list of opinion scores as it happens in a subjective test. One can then: i) average them to get an estimate of the MOS; ii) compute the distribution of the obtained scores on the quality scale and thus obtain an estimate of the percentage of final users that would not be satisfied by the quality of that PVS; iii) look at the characteristics of the AIOs that predicted the quality as unsatisfactory to get potential information on the type of final viewers that would not be satisfied.

The AIOs therefore offer the opportunity to objectively assess the quality while taking into account the expectations of individual human viewers, this is a fundamental feature that previous approaches missed.

It is important to note that being able to accurately train AIOs brings new challenges within the media quality assessment community. The MOS by definition is affected by less noise than the individual opinions scores; in fact the average is performed for this purpose. Nevertheless, existing subjectively annotated datasets are already limited in size for training QPAs that can predict the MOS. This situation becomes more critical when one has to deal with the task of predicting single opinion scores that are affected by more noise.

To effectively collect data tailored to the training of the AIOs, it is important to rethink the traditional recommendations that are valid for subjective tests aimed at MOS prediction. For instance, while having many viewers in the test is not fundamental for the training of AIOs, it is instead very important that the same viewer rates a large number of stimuli. One therefore needs new recommendations on how to handle the viewer fatigue, since doing the test in many different sections would still introduce some noise in the process. This issue is still open, and therefore I proposed in [129] and [31] two approaches to smartly make use of existing subjective experiments in order to train the AIOs.

In [129], I propose to form a cluster with viewers that have similar perception of quality, i.e., those that have expressed similar opinion scores on a given set of video sequences they all rated. A single AIO was then trained to represent all those viewers while exploiting all the subjective scores gathered from each of them.

In [31] instead, I propose to first create a large scale synthetically annotated dataset containing JPEG compressed images, then I trained on it a DNN that I called JPEGResNet50. The JPEGResNet50 is a DNN that is able to classify images

on the basis of their JPEG compression level. It is therefore a DNN that is able to extract useful features for perceptual quality assessment. Finally I performed a transfer learning step to convert the JPEGResNet50 into a DNN-based AIO by continuing its training process on the data gathered from the observer to be modeled. In this last training step the weights of the JPEGResNet50 were adjusted to extract features that better characterize the quality perception of the viewers to be mimicked.

Both approaches adopted to train the AIOs have led to very promising results that clearly show the feasibility and effectiveness of the AIOs-based approach.

The idea of training many different models, i.e. the AIOs, instead of designing a single model that predicts the MOS gives rise to two fundamental questions: i) how to select the subjects to be modeled when designing the AIOs? ii) How many AIOs need to be considered to ensure that the mean of their predicted opinion scores yields a robust estimate of the quality? It is important to note that these two questions come up every time one designs a subjective experiment. For this reason, part of the previous media quality assessment literature has been devoted to these questions.

By relying on AIOs to assess the quality of a PVS, one actually attempts to simulate the process of a subjective test. Therefore, recommendations that are valid for an actual subjective test should be followed when designing and using the AIOs.

In particular, the recommendations on how to select the participants in a subjective experiment [92] should be followed to identify a suitable set of actual observers to be modeled. As it should happen in a subjective test, the actual observers should be sampled in such a way that the related AIOs are representative of the whole population of potential users of the multimedia service of interest.

According to the following ITU-T recommendations [49, 92], the MOS deriving from a subjective experiment conducted in a controlled environment with at least 15 viewers is expected not to change with statistical significance when different sets of viewers are used. A quality measure deriving from a set of at least 15 AIOs could therefore be considered as robust when the related actual observers have been selected in order to best represent all potential users of the service.

## 1.4 The Thesis Structure

The rest of this thesis is structured as it follows. Chapter 2 introduces a probabilistic approach to quality assessment aiming at accounting to both the inconsistency of the human viewers and the inaccuracy of existing QPAs. In Chapter 3 and Chapter 4 two different approaches to derive measures that can inform on the reliability of a MOS prediction are presented and some related applications are discussed. The Chapter 5 will introduce the concept of AIOs and their training

with shallow NNs. The AIOs are further studied in Chapter 6, in particular an approach to train DNNs-based AIOs despite the challenges imposed by the lack of large scale subjectively annotated datasets is presented. Finally, conclusions are drawn in Chapter 7.

# Chapter 2

# A Probabilistic Approach for Computing Quality-of-Experience Ranges in Video Quality Assessment

## 2.1 Introduction

The distribution of the opinion scores gathered for a given PVS during a subjective test is generally modeled as a Gaussian distribution. Hence the subjective assessment approach typically leads to a mean opinion score (MOS) and a Gaussian-modeled confidence interval for this MOS. The assumption that the opinion scores distribution is indeed Gaussian and that a single value is sufficient, is challenged by a large number of system and human IFs.

Among the system IFs, one might list the diversity of source contents, the large number and type of degradation, especially multi-dimensional degradation where dimensions can be, for instance, image distortions and temporal distortions. All these factors contribute to shape the complexity of the stimuli, which in turn generates uncertainty in the evaluation process of each human subject [55, 68].

Human IFs, on the other hand, are all those characteristics that make viewers perceive and judge quality in a different way. For example, a person's background, experience, and expectations in terms of quality significantly influence his/her assessment. Furthermore, this influence is very variable over time, causing a human subject to evaluate the same stimuli differently under the same conditions [55]. It is evident that reducing the subjective quality of a content to a single value does not allow to take into account the uncertainty caused by all these IFs.

A possible approach to account for such uncertainty is to express the perceptual quality in probabilistic terms. This Chapter describes the approach I published

in [29]. In this approach, the perceptual quality of a PVS is expressed as a range instead of a single value. More precisely, given a PVS, the scope of the approach is to derive a range of numerical values to which the score of its perceptual quality belongs with a certain probability. Such probability should be specified by the user of the approach.

This approach should not be confused with confidence interval estimation for video quality, which starts from the assumption that an estimate of the MOS is available, and that the opinions of human subjects are distributed according to a Gaussian distribution.

To compute the score range, i.e., a minimum and maximum value, well-known objective full-reference video quality measures (FR VQMs) that can be easily computed were used. The underlying idea is that the use of several VQMs, each based on different approaches, could somehow capture the multi-dimensional degradation that may affect the PVSs quality.

The main steps of the approach presented in this chapter can be summarized as it follows. Given a PVS, its MOS is considered as a random variable. The joint probability distribution of the MOS and the scores of the considered FR VQMs is derived. The conditional distribution of the MOS to the score of the FR VQMs and the related quantiles are computed and used to obtain the desired range of quality. This approach is quite innovative in the media quality assessment community as it represents the first attempt to work in the direction of predicting ranges of quality to account for uncertainty. The motivation and examples of the usefulness of such approach are discussed in more details in Section 2.2.

To implement the approach, the results of the VQEG HDTV Phase I experiment [139] (VQEG-HD) which is one of the most extensive subjectively-annotated publicly-available datasets, with a large variety of high resolution (1920x1080) content, were used. Such dataset reasonably covers the large majority of cases in which the video quality research community could be interested.

The rest of the chapter is organized as it follows: Section 2.2 elaborates more on the usefulness of the approach while Section 2.3 presents the technical steps behind it. Some computational experiments conducted to showcase the effectiveness of the approach are presented in Section 2.4, and the conclusions are drawn in Section 2.5

## 2.2 Motivation

The probabilistic representation of the subjective quality of a PVS presented here is mainly motivated by the three points that are discussed in this section.

### 2.2.1 Accounting for Stochastic Influence Factors

In Section 2.1 several reasons that cause an intrinsic difficulty in trying to express and determine a single QoE value for each test case, e.g., each PVS, regardless of the goodness of the algorithm used to estimate such single QoE value have been highlighted.

The approach presented in this chapter is therefore strongly motivated by the meed to report the quality score under a format that clearly considers the stochastic nature of the problem. To be coherent with the human behavior in terms of quality perception, MPSs should also report the estimation of the perceptual quality in probabilistic terms. For a given PVS, it is more complete to state that its perceptual quality lies in a given range with certain probability. Each value in the predicted range would be a feasible quality score, and this would account for the fact the same PVS might be judged differently when watched under different conditions.

Finally, it is worth noting that for some applications, one might simply be interested in guaranteeing with a certain probability that the perceived quality would not go under a certain threshold. In that case, it is enough to encode the PVS such that the lower bound of the quality range derived by the approach described in this chapter matches the desired threshold.

### 2.2.2 Analyzing VQMs' Accuracy

The accuracy of Quality Prediction Algorithms (QPAs) is usually measured by relying on statistical indicators such as the Pearson linear correlation coefficient, the Spearman rank order correlation coefficient and the root mean square error. Despite the usefulness of these indicators, they do not allow a local analysis of the performance of QPAs on each part of the quality scale. Instead they provide a global idea on the performance of the QPA on the whole quality scale.

To compute the quality ranges, the approach presented in this chapter first derives the conditional distribution of the MOS with respect to the output of the considered VQMs. By drawing the curves reporting the quantiles of that distribution as a function of the VQM score, one obtains a figure of merit of the VQM. As it will be discussed later in the chapter, such a figure has the advantage that it provides local information on the accuracy of the VQMs while accounting for the uncertainty that characterizes the MOS.

### 2.2.3 Dealing with Large Scale non-Annotated Datasets

From a practical point of view, estimating a range without even attempting to compute a single QoE value can be useful also for analyzing large scale not subjectively annotated datasets. For instance, the size of the predicted quality range, seen as a measure of uncertainty, can be analyzed separately for each source

content to figure out the content that is peculiar in terms of quality assessment.

Another example could be the selection, from a large scale dataset, of a subset of stimuli to use for a subjective test. For a subjective test it is important to ensure that the used stimuli are a good choice in terms of variety of quality. This can be done by computing the expected quality range for all selected stimuli, so that the heterogeneity of the stimuli is assessed while accounting for the stochastic IFs that will undoubtedly affect the result of the test.

## 2.3 QoE Ranges Estimation

This section details the steps towards the derivation of the QoE range for a given PVS.

### 2.3.1 Problem Settings

Let denote by:

- $V = (vqm_1, vqm_2, \ldots, vqm_n)$ a vector containing the quality scores of $n$ VQMs determined for a given PVS;
- $\alpha \in [0,1]$ a user specified tolerance, i.e., the probability that the MOS might not be in the predicted range;
- $mos^{pvs}_{Min}$ and $mos^{pvs}_{Max}$ the lower and upper bound of the range of quality to be determined for the considered PVS.

The problem to address consists of computing the values of $mos^{pvs}_{Min}$ and $mos^{pvs}_{Max}$ such that:

$$\begin{array}{rcl} \Pr\left(MOS \leq mos^{pvs}_{Min} | V\right) & = & \alpha/2, \\ \Pr\left(MOS \geq mos^{pvs}_{Max} | V\right) & = & \alpha/2, \end{array} \tag{2.1}$$

where Pr means probability.

In words, given a PVS for which the scores of $n$ VQMs are known, one want to get an interval in which the MOS is expected to stay with probability $1-\alpha$. This is the reason why the parameter $\alpha$ is referred to as a tolerance. The smaller is $\alpha$, the larger is the size of the predicted range.

Mote that the problem in Eq (2.1) consists of finding the quantiles at $\alpha/2$ and $1-\alpha/2$ of the conditional probability distribution of the MOS with respect to the vector containing the scores of the $n$ VQMs. To this aim, one needs first to estimate the joint probability distribution of the MOS and the vector $V$ of the VQM scores, then derive from it the conditional distribution whose quantiles are of interest.

While jointly considering all the VQMs in the vector $V$ together is certainly the most desirable approach to the problem, their differences in terms of scale and

how to map them on the MOS make it very difficult and generate computational instability. Therefore, for simplicity sake and for easier graphical interpretation, each VQM is treated individually. A final pooling step is then implemented to account for the effect of all the VQMs.

### 2.3.2 The Dataset for the Joint Density Estimation

To estimate the joint probability distribution of the MOS and each VQM, the VQEG-HD dataset [139] was considered. The VQEG-HD dataset was designed in such a way that the evaluated stimuli cover a large set of content, conditions, and quality ranges.

The perceptual quality of the used stimuli was impaired with both coding artifacts, (MPEG-2 and AVC encoded PVSs with bit rates varying from 1 to 15Mbps), and transmission ones (bit error and bursty packet loss). The used video content involved movies, sports, general TV material with much variability as possible. Therefore, by relying on such a dataset, one might expect to get a rather good representation of most of the conditions that can be encountered in the majority of real-world applications.

While more details can be found in the VQEG-HD experiment final report [139], a short description of the experiment settings is provided here for completeness sake.

**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 [49]. 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 $1920x1080$.

**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 [123] and for normal color vision by means of the Ishihara test [18]. 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 [139].)

**Assessment procedure:** The absolute category rating with hidden reference [92] 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.

For each PVS in the VQEG-HD dataset, five well known and widely used FR VQMs were computed, i.e., the PSNR, the SSIM, the MSSSIM, the VIF [39], and the VMAF 0.6.2 [87]. Since most of these VQMs do not easily handle interlaced video, the analysis was restricted to non-interlaced sequences, i.e., the VQEG-HD1, VQEG-HD3 and VQEG-HD5 subsets. The vector $V$ of VQM scores for each PVS therefore included the scores computed with the five aforementioned VQMs.

Please note that for the VQEG-HD complete dataset, created by aligning and joining the results of the six different laboratories (see Chapter 7 of [139]), the MOS scores range in [0.82, 5.26].

### 2.3.3 Joint Probability Distribution of the MOS and a VQM

Let demote by $f(vqm, mos)$ the joint probability distribution of the MOS and a VQM. I proposed in [29] to approximate such a joint probability distribution with a 2D Gaussian mixture model (GMM) and use the VQEG-HD dataset to fit the parameters of the GMM for each VQM.

Therefore, exploiting the general expression of a GMM, the joint probability distribution of the MOS and each VQM can be written as it follows:

$$f(vqm, mos) = \sum_{i=1}^{k} \pi_i \cdot N\Big((vqm, mos)|\mu_i, \Sigma_i\Big) \tag{2.2}$$

Where $N\left((vqm, mos)|\mu_i, \Sigma_i\right)$ is the p.d.f. of a bivariate normal distribution with mean $\mu_i$ and covariance matrix $\Sigma_i$ and $k$ is the number of Gaussian components of the GMM.

The parameters ($\pi_i$, $\mu_i$, $\Sigma_i$ and $k$), for each VQM, have been estimated from the data collected during the VQEG-HD experiment using the expectation maximization (EM) algorithm [84]. When using the EM, there are many different criteria to determine which is the best number of Gaussian components to use. The Bayesian Information Criterion (BIC) was used to find out the optimal number of Gaussian components to use for each VQM, i.e., the point at which the BIC curve (as a function of $k$) becomes almost flat [6]. In practice, for the considered application,

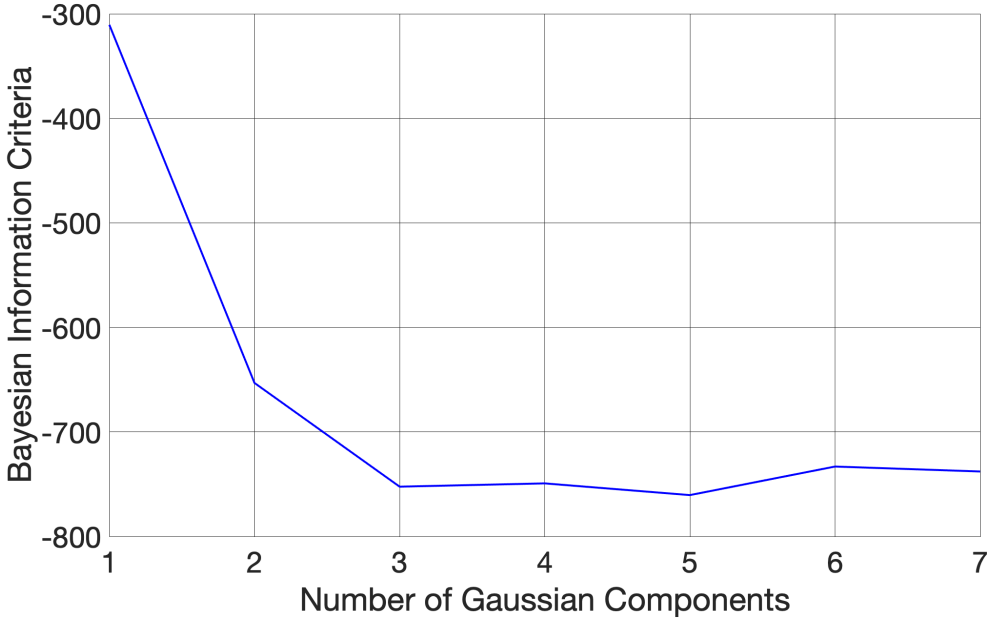



Figure 2.1: The figure shows the value of the Bayesian Information Criteria (BIC) obtained from the fitted GMM as function of the number of Gaussian components of the GMM in the MSSSIM case. As it can be noticed, using more than three Gaussian components does not yield a significant variation of the BIC and thus of the model performance. Therefore in the MSSSIM case, three Gaussian components were used.

this happens either for $k = 3$ or $k = 4$ depending on the VQM. The Figure 2.1 illustrates such a procedure in the case of the MSSSIM.

The Figure 2.2 shows how the fitted density for each VQM models the dispersion of the samples in the VQEG-HD dataset. As one could expect from a good model, there is in general a greater density of samples in the areas where the fitted distribution assumes large value and vice versa. This indicates that the assumption that a GMM can capture the relation between the MOS and each VQM scores is rather reasonable.

### 2.3.4 Deriving the QoE Ranges

Once a suitable 2D GMM is fitted for each VQM, it is possible to compute the conditional probability distribution of the MOS for a given VQM score.

In particular, let divide the useful variation range of each of the five considered VQMs in the VQEG-HD dataset into 100 equal parts. For the center $vqm_j$ of each interval, one wants to compute the function:

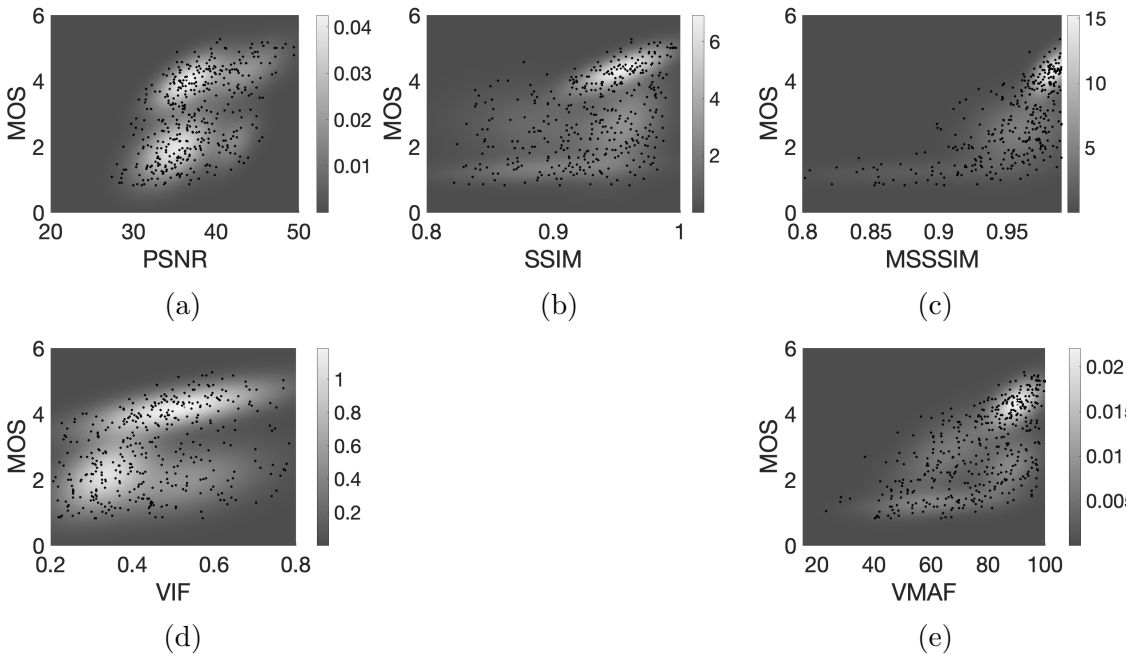



Figure 2.2: A 2D representation of the fitted GMM for each VQM. In general, the larger density of points in whiter regions highlights the GMM accuracy

$$G(x|vqm_j) = \Pr\left(MOS \leq x | vqm_j - \delta \leq VQM \leq vqm_j + \delta\right) \tag{2.3}$$

where $\delta = (\max(VQM) - \min(VQM))/100$.

Note that as $\delta$ is a small value, the function $G(x|vqm_j)$ approximates the probability that the MOS assumes a value smaller or equal to $x$ when the VQM score is equal to $vqm_j$. It is therefore the conditional probability distribution of the MOS with respect to the score of the VQM.

From the definition of conditional probability, and exploiting the properties of the joint probability distribution, the following formula for $G(x/vqm_j)$ holds:

$$G(x|vqm_j) = \frac{\int_{-\infty}^{x} \int_{vqm_j-\delta}^{vqm_j+\delta} f(r,t)drdt}{\int_{vqm_j-\delta}^{vqm_j+\delta} \int_{-\infty}^{\infty} f(r,t)dtdr} \tag{2.4}$$

where the function $f$ is the joint probability distribution expressed in Eq (2.2) and whose parameters have been fitted for each VQM using the VQEG-HD dataset. Therefore, for each VQM, given a value of $x$ and the score $vqm_j$ of the VQMs, the probability $G(x|vqm_j)$ that the MOS is smaller or equal to $x$ can be numerically computed using the formula in Eq (2.4).

For each VQM, the following equations are numerically solved respectively for $mos_{Min}^{vqm_j}$ and $mos_{Max}^{vqm_j}$ to determine the desired MOS bounds when that VQM score is equal to $vqm_j$, $\quad j = 1,2,\ldots,100$:

$$\begin{aligned} G(mos_{Min}^{vqm_j}|vqm_j) &= \alpha/2, \\ G(mos_{Max}^{vqm_j}|vqm_j) &= 1-\alpha/2. \end{aligned} \tag{2.5}$$

If one remembers the meaning of the function $g$, these equations are simply stating that $mos_{Min}^{vqm_j}$ and $mos_{Max}^{vqm_j}$ are respectively the quality scores to which the MOS is smaller respectively with probability $\alpha/2$ and $1-\alpha/2$ if the VQM score is equal to $vqm_j$.

For instance, let substitute the generic VQM in the used notation with the PSNR. It will be assumed in the experimental section that the typical range of variation of the PSNR in practice is from 20 to 50 dB. Let divide such a range in 100 equidistant intervals and call $psnr_j$ the center of the jth interval. $psnr_j$ corresponds to $vqm_j$ in the generic notation. Then the interval $[mos_{Min}^{psnr_j}, mos_{Max}^{psnr_j}]$ is the range to which the MOS of a given PVS belongs with probability $1-\alpha$ if the PSNR score of that sequence is equal to $psnr_j$. The same interpretation holds for all the other VQMs considered in this chapter.

Note that by interpolating the 100 values $mos_{Min}^{vqm_j}$ and $mos_{Max}^{vqm_j}$ $j = 1,2,\ldots,100$, for each VQM, one can obtain two curves delimiting the range to which the MOS belongs for any given value of the VQM (see Figure 2.3).

The min and max values for each single VQM obtained in Eq (2.5) and ploted in Figure 2.3 can now be combined together to obtain a global min and max MOS value for a given PVS. To perform such a pooling step, the average of the ranges obtained for each VQM was computed.

Hence, the expressions of the bounds of the final quality range are as it follows.

$$\begin{aligned} mos_{Min}^{\text{PVS}} &= \tfrac{1}{n}\textstyle\sum_i \left( mos_{Min}^{\text{VQM}_i^{\text{PVS}}} \right) \\ mos_{Max}^{\text{PVS}} &= \tfrac{1}{n}\textstyle\sum_i \left( mos_{Max}^{\text{VQM}_i^{\text{PVS}}} \right) \end{aligned} \tag{2.6}$$

where $n$ is the total number of used VQMs ($n = 5$ in this chapter's case), $mos_{Min}^{\text{VQM}_i^{\text{PVS}}}$ and $mos_{Max}^{\text{VQM}_i^{\text{PVS}}}$ are respectively the lower and the upper bound of the quality range of the ith VQM.

Therefore, at the end of the procedure, when a new PVS with unknown MOS is presented to the proposed system, the scores of the five VQMs considered in this chapter are first computed. Then the values $mos_{Max}^{\text{VQM}_i^{\text{PVS}}}$ and $mos_{Max}^{\text{VQM}_i^{\text{PVS}}}$, $i = 1,2,\ldots,5$ are computed for each VQM. Finally they are aggregated using the average to form the final MOS range for that PVS.

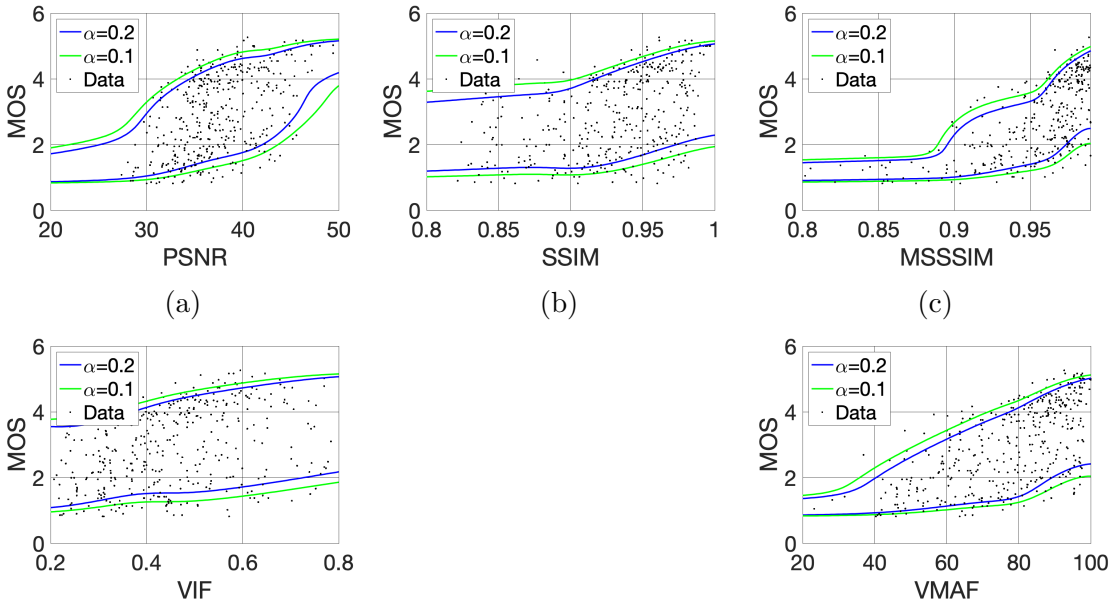



Figure 2.3: The curves determine the predicted MOS ranges as function of each VQM score. The curves are shown for two different values of $\alpha$. Each point corresponds to a single PVS in the dataset. the MOS values belong to [0.82, 5.26] due to the realignment of the six VQEG-HD subsets [139].

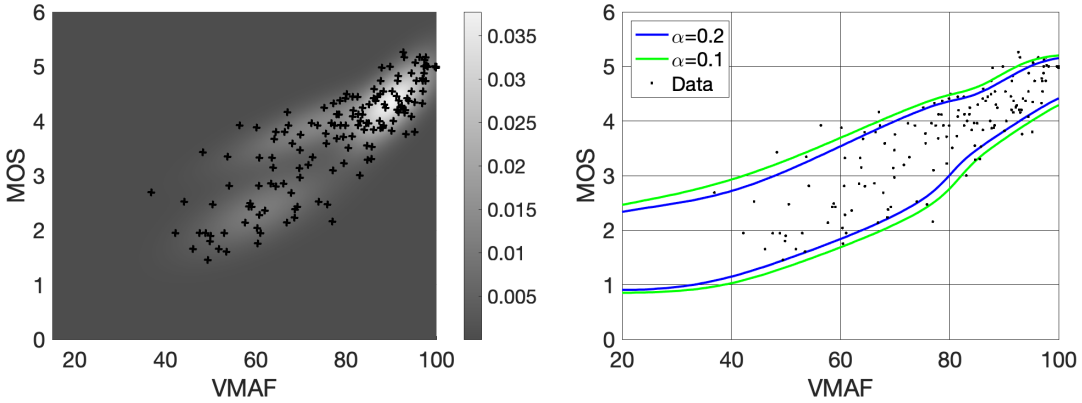



Figure 2.4: Results obtained for VMAF when considering only coding artifacts.

Table 2.1: Predicted quality range accuracy.

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

## 2.4 Numerical Experiments

### 2.4.1 Experimental Settings

To validate the effectiveness of this approach, two datasets were considered in addition to the used VQEG-HD dataset. These two datasets contain video sequences not considered by the proposed system when fitting the joint probability distributions that are used to compute the quality ranges.

Both datasets include high resolution content (1920x1080). The first is the Netflix Public Dataset [68], which includes 70 subjectively annotated PVSs covering the full MOS range. The second is the VQEG JEG-Hybrid Large Scale Database (JEG-DB)[14] which includes 19,840 1080p PVSs obtained by compressing a few source sequences in HEVC format using a large set of coding parameters, including bitrates ranging from 500 Kbps to 16 Mbps.

### 2.4.2 VQMs Figure of Merit

Figure 2.3 shows the curves obtained by interpolating the 100 $mos_{Min}^{vqm_j}$ and $mos_{Min}^{vqm_j}$ points for each VQM, for two different $\alpha$ values. The original points in the VQEG-HD dataset are also reported on the figure. Looking at the green curve in Figure 2.3a, it can be observed, for instance, that for a PVS with PSNR equal to 47 dB, the MOS is expected to be in the range [3,5] with a probability of 0.9 (1-0.1).

In addition to allowing to compute the quality range to which the MOS belongs with a certain probability, each of the Figures 2.3a, 2.3b, 2.3c, 2.3d and 2.3e can be considered as a figure of merit of the related VQM, since the distance between the curves and their shape for a given value of $\alpha$ provides interesting information on the metric behavior. The less distant the curves are, the more accurate and robust the metric is to noise that affects the MOS. It is interesting, for example, to observe how the PSNR, MSSSIM and VMAF do not seem to have a uniform accuracy on

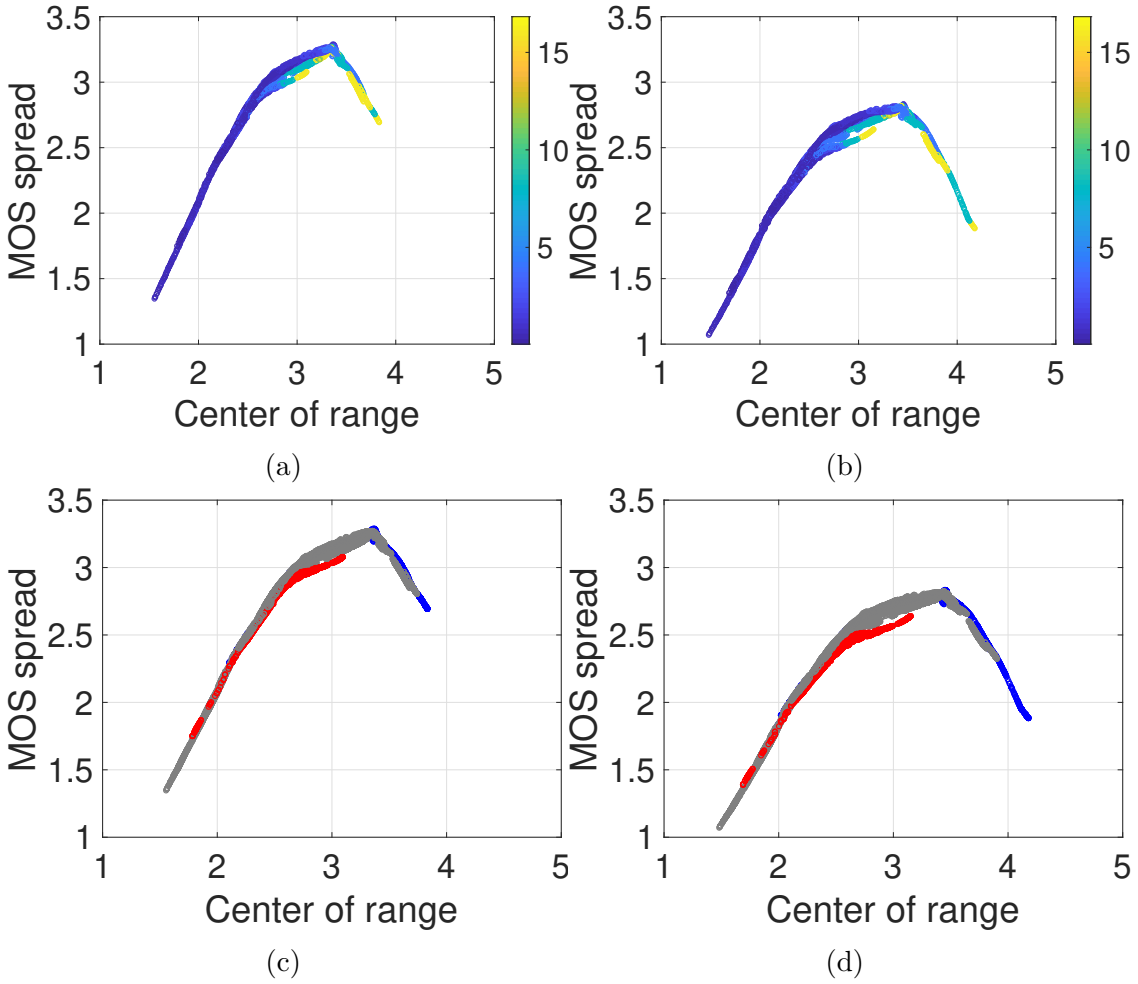



Figure 2.5: Size (MOS spread) vs center of the predicted quality range. The analysis is done on the JEG-DB. $\alpha$=0.10 (left), and $\alpha$=0.20 (right). Colors indicate the PVS bitrate (Mbps) (top), and different sources (bottom).

all their scales. While the SSIM and the VIF have almost parallel curves, and thus the same uncertainty is maintained in the estimation of the subjective quality regardless of the score they have predicted.

Note that, being derived from the VQEG-HD dataset, the curves in Figure 2.3 represent the expected ranges of quality in case the VQMs would be used to assess the quality of a PVS that might involve transmission or coding artifacts. Therefore, the obtained ranges of quality are large in size also because the information on the type of artifacts is not known a priori and the system accounts for this additional source of uncertainty.

For instance, the VMAF was originally designed to be used on PVSs whose quality has been impaired only with coding artifacts. The Figure 2.4b shows the

quality ranges for each VMAF scores, for two different values of $\alpha$, when the analysis is restricted to PVSs whose quality is affected only by coding artifacts. It can be seen that smaller quality ranges were obtained in comparison to what happened in the generic application case in Figure 2.3e. For completeness' sake the graph of the GNN fitting the joint density of the VMAF and the MOS for the particular case of coding artifacts is also shown in Figure 2.4a.

### 2.4.3 Accuracy of the Predicted Quality Ranges

The effectiveness of the approach was also assessed by checking whether the predicted interval actually contains the MOS with the user specified probability. The experiment was done on the VQEG-HD and the Netflix Public datasets.

In practice, for each PVS in these datasets, the values of the five VQMs needed to compute the quality range were first computed. Then, the quality range was computed for many different value of $\alpha$. As the MOS is expected to belong to the predicted interval with probability $1 - \alpha$, to check the effectiveness of the system, it is enough to verify how close is the actual number of PVSs whose MOS is out of the predicted ranges to the expected number given by $\alpha * N_{PVS}$, where $N_{PVS}$ denotes the total number of PVSs in the test set.

The Table 2.1 shows a comparison between the actual number of PVSs out of the predicted range and the expected one for the VQEG-HD and the Netflix Public dataset. The result shows that the proposed system can compute the MOS ranges accurately even when used on a dataset not considered when fitting the GMMs that are used to compute the quality ranges. In particular, the fraction of MOS values outside the range is close to the expected one, determined by the $\alpha$ value. In all cases, the number of PVSs falling outside the range differ from the expected one for max 8 units.

### 2.4.4 Analyzing a Large Scale Dataset

Now, let consider the JEG-DB, i.e., a dataset with a huge number of PVSs for which the MOS values are not available. The top part of Fig. 2.5 shows the distribution of the length of the ranges obtained, as a function of the center of the predicted ranges. Clearly, as $\alpha$ increases, the size of the range (the MOS "spread") decreases. Moreover, as expected, when the bitrate of the PVS is at one extreme (low or high), the range size is reduced, i.e., there is less doubt on the MOS position, respectively low or high. However, for intermediate values, the range size increases.

Despite not having MOS values, it is however possible to spot interesting peculiar behaviours. In the bottom part of Fig. 2.5 for instance, the points corresponding to two source contents are highlighted (all others are in grey): the blue shows a sequence (a cartoon) which exhibits a quite peculiar behavior in terms of MOS (less uncertainty for high quality), whereas manual inspection showed that the red

points correspond to sequences with some digital noise in the original source. This simple analysis underlines the usefulness of being able to estimate even just MOS ranges to identify interesting behaviors in a large database of video sequences.

## 2.5 Conclusion

Typically, the tool to measure the QoE for PVSs are designed to predict a single value, in most cases the MOS. Predicting this value using various algorithms has been widely studied. However, deviation from the MOS is often handled as an unpredictable error. The approach presented in this chapter allows to estimate an interval of video quality to which the MOS is expected to belong with a user specified probability. Such a probabilistic interpretation and representation of the perceptual quality allows to account for the several IFs that make the MOS a random variable.

To derive the desired interval of quality scores for a given PVS, well known and widely used video quality estimators are fused together to output a lower and upper border for the expected video quality, on the basis of a model derived from a well-known subjectively annotated dataset. Results on different datasets provide insight on the suitability of the well-known estimators for this particular approach.

While the approach described in this chapter argues that the MOS of a PVS should not be treated as a deterministic value and proposes to compute ranges of quality, many practitioners, for the sake of convenience, however continue to prefer to have available a punctual estimation of the quality. A good trade-off is therefore to account for the uncertainty by integrating any punctual estimation of the MOS, with a measure that informs on how reliable it is. This is the approach that will be adopted in the next two chapters of this thesis.

# Chapter 3

# A Neural Network-based Approach to Predict the Diversity of Users' Opinion Scores

## 3.1 Introduction

Subjective tests are considered the most reliable way to assess the perceptual quality of any type of media. However, human opinion scores are characterized by large diversity: in fact, even the same subject, is often not able to exactly repeat his/her first opinion when assessing once more the quality of a given stimulus. This makes the mean opinion score (MOS) alone, in many cases, not sufficient to get full information about the perceived visual quality [43].

It is therefore important to have measures characterizing to what extent the observed or predicted MOS value is reliable and stable. For instance, the Standard deviation of the Opinion Scores (SOS) is usually considered as a measure of the MOS reliability when evaluating the quality subjectively [43].

Unfortunately, the literature is still lacking models or algorithms that allow to objectively explain and predict how much diversity would be observed in subjects' opinion scores in terms of SOS. In this journal paper [30], I focused on this problem and proposed a machine learning-based approach to cope with it. This chapter presents and discuss the main technical steps behind such an approach.

The approach to model the users' diversity of opinion scores presented here is strongly based on a statistical analysis made on several subjectively annotated datasets. The result of that analysis revealed that on a set of processed video sequences (PVSs) for which there is large diversity among the observers' ratings, the scores predicted by different video quality measures (VQMs) are expected to be less correlated, in comparison to what happens when the same VQMs are used instead to assess the quality of PVSs on which there is low variability among users' opinions.

In light of this observation, I hypothesized in [30] that part of the variability observed among the subjects's opinion scores can be captured/modeled by exploiting the quality scores output by several VQMs on the same PVS. Therefore the SOS was modeled as the sum of two components: i) a deterministic component called ground truth SOS (gtSOS) that can be estimated through the use of neural networks (NNs) by exploiting the quality scores of several VQMs that are provided as input features to the NN; ii) a random term modeling the two main sources of error caused by subjective tests, i.e., the quantization of the quality scale and the limited number of subjects used to conduct the test.

In this way, a distinction is made between the SOS directly observed in a typical subjective test (with a finite and often very limited number of observers' rating on a discrete scale) and the gtSOS, that is to be looked at as 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 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.

By predicting the gtSOS of a PVS, one expects to measure how much reliable would be any estimation of the perceptual quality of that 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 perceptual quality is difficult to predict consistently receive higher attention, thus ensuring that all users experience a uniform and high satisfaction level.

The validity and the effectiveness of the proposed SOS model was assessed on several datasets. In particular it was shown to be a suitable tool to identify potential anomalies in the data gathered in subjective tests.

The rest of the chapter follows the following structure. The related work is briefly presented in Section 3.2. The SOS importance in media quality assessment as well as the innovativeness of the approach described in this chapter are discussed in Section 3.3. The SOS model is presented in Section 3.4, followed by the Section 3.5 where the model is validated by means of numerical experiments. Section 3.6 illustrates how it is possible to highlight potential anomalies in the data collected during a subjective test using the SOS model described in this chapter. Section 3.7 is devoted to the design and training of NNs specific for gtSOS prediction. Conclusions are drawn in Section 3.8.

## 3.2 Related Work

More and more researchers, working in different areas, are relying on machine learning (ML) techniques due to their ability to extract information from data without necessarily making assumptions about a model underlying the data [57]. Depending on the used methods, ML approaches can generate predictions on the basis of an input, e.g., regression models, or simply provide insights in the observed data, e.g., clustering techniques. The media quality assessment research community has naturally also adopted ML approaches and relied on them to propose several models aiming at predicting the subjective quality, i.e., the MOS, of a PVS starting from a number of different features extracted from it by means of algorithms [149, 17].

Unfortunately, the ML applications in media quality assessment community have been mostly restricted to the quality prediction [19, 27, 33], while the problem of predicting the deviation from the MOS, despite being a hot topic, has benefited only slightly from the success of such an approach [82]

In fact, in some recent papers, the authors highlighted the inability of the MOS to fully capture all the aspects necessary to measure the perceptual quality of a media. In [29], the deviation from the MOS is handled by determining ranges of QoE rather than a single MOS value. The authors in [114] 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 perceptual quality. The approach in this chapter therefore aims at being one of the first steps toward predicting the variability among users' opinion scores by leveraging ML-based approaches.

Finally, the analysis of data coming from subjective tests has also taken very limited advantage of ML methods to figure out potential anomalies and thus eliminate noise in the collected data [7]. Traditional techniques to identify unusual and strange behavior in the subjectively annotated datasets, makes use of standard statistical approaches (e.g. outlier detection, likelihood estimation, etc.) [49, 55, 68]. The approach of this chapter showcases the usefulness of ML-based methods also for investigating the quality of subjective data.

## 3.3 The SOS as a Measure of Users' Diversity of Opinion Scores

### 3.3.1 Computing MOS Confidence Intervals

In media quality assessment, the SOS has typically been exploited to compute 95% confidence intervals (CIs) for the MOS as follows:

$$CI = MOS \pm \frac{SOS \cdot \tau_{n-1}^{97.5}}{\sqrt{n}} \tag{3.1}$$

where $n$ is the total number of subjects that participated in the subjective test 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 has been traditionally looked at as the main tool to distinguish between PVSs whose quality has been or can be consistently evaluated (those with a small CIs) and the PVSs that caused high uncertainty (those with large CIs) during the quality assessment process.

Based on Eq. (3.1), the computation of a CI requires to first collect the subjects' opinion scores, since they are required to calculate the MOS and the SOS. CIs can therefore only be computed after carrying out a subjective test. This precludes the possibility of using them in real-time to automatically determine which PVSs need to be granted more resources in an attempt to reduce the high uncertainty that affects their perceptual quality.

This difficulty to effectively use in practice CIs as computed by the formula in Eq (3.1) can be overcome by using predicted CIs. This would however require to predict not only 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.

### 3.3.2 The SOS Hypothesis and its Limits

The most widespread approach to the SOS estimation within the quality assessment community is the one presented in this paper [43]. The authors studied the SOS in relation to the MOS, postulating that the SOS is linked to the MOS through a second order polynomial function as it follows:

$$SOS = \alpha * (-MOS^2 + 6 * MOS - 5) \tag{3.2}$$

where the parameter $\alpha$ is to be calibrated and its value depends on the application under investigation.

The crucial drawback of this postulate is that it 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 test, 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 test.

The approach in this chapter explores, for the first time, the possibility of estimating the subjects' diversity of opinions on a given PVS using only features

extracted from it, namely the VQM scores. More precisely, the main sources of errors that might affect the SOS computed from the raw data of a subjective test are highlighted. The gtSOS is introduced and shown to be predictable from the PVS' characteristics only. The gtSOS, being an estimate of the SOS not affected by the errors introduced by any specific subjective test, results in a more stable and reliable measure of the observers' opinion scores diversity.

Authors in various scientific fields have proposed several sophisticated metrics aiming at measuring the level of consensus between the opinion scores of subjects gathered in Likert scale-based studies [138, 146]. Unfortunately, the media quality assessment research community still did not adopt such measures. Instead, the SOS remains, until now, the only measure of the subjects' diversity of opinion scores.

For some of the research fields in which opinion scores are collected using a Likert scale, it is possible to re-adjust the experimental setup or the questionnaire before resubmitting it to the attention of the participants. Moreover, there is the possibility to iterate in this way until reaching a certain consensus among the subjects involved in the study. Unfortunately, the media quality assessment process is influenced by so many factors even unknown to the subjective test designer [107]. This prevents the implementation of a consensus-based process and precludes 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.

## 3.4 Modeling the SOS in Subjective Tests

This section introduces and describes the two main components that contribute to determine the SOS values observed in a subjective test. In particular, it is argued that the observed SOS value for a given PVS results from the sum of a deterministic predictable quantity (gtSOS) and a stochastic noise caused by the subjective test settings.

### 3.4.1 The Ground Truth SOS (gtSOS): Link with VQM Scores

The gtSOS of a PVS is supposed to be the systematic part of the SOS value that represents a measure of the uncertainty intrinsically associated with the perceptual quality of that PVS due to its complexity. As such it should be computed using the features extracted from the PVS itself. In the context of the approach described in this chapter, the scores of many VQMs obtained when using them to assess the quality of the PVS, were used as features.

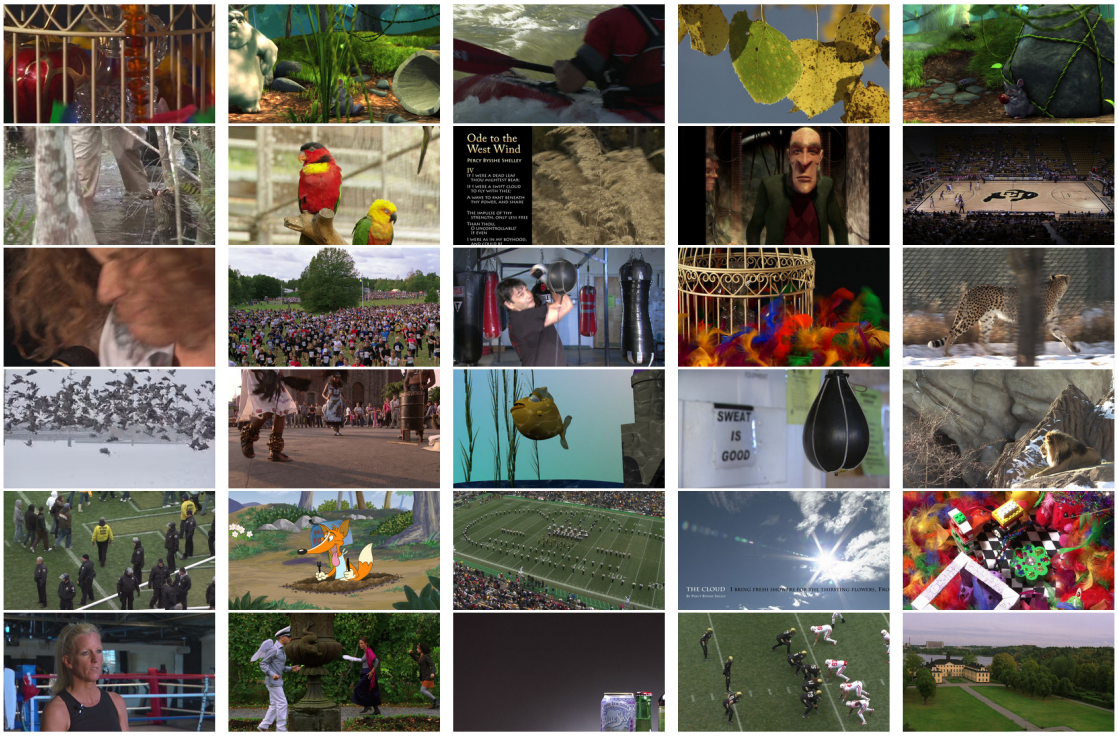



Figure 3.1: Sample video frames, each column corresponds to a single dataset and the order is: ITS4S, Netflix Public, VQEG HD1, VQEG HD3 and VQEG HD5.

The motivation of this choice is at it follows. Different VQMs are in general based on the modeling of different aspects of the human visual system. Therefore, There might be artifacts, which a certain VQM is very sensitive to, while others less so. An important and systematic part of the diversity observed among human viewers opinion scores is also devoted to that same effect, i.e., the fact that some persons see artifacts that others do not perceive, and the level of annoyance of perceptible artifacts varies from one viewer to another. The idea is therefore to capture the predictable part of the diversity of human opinion scores by leveraging the misalignment observed between the scores of several VQMs.

To verify to which extend such an idea is reasonable, a statistical analysis was performed to verify whether there is lower rank correlation between the VQM scores in the presence of greater diversity of viewers opinion scores.

The analysis was done on five subjectively annotated datasets, i.e., the ITS4S dataset [96], the Netflix public dataset [68] and three datasets released by the Video Quality Expert Group (VQEG): the VQEG-HD1, VQEG-HD3, and VQEG-HD5 [139]. Some examples of images extracted from the sources (SRCs) used to generate the PVSs in each of those datasets are presented in Figure 3.1.

The analysis was performed as it follows. For each dataset, the PVSs were sorted

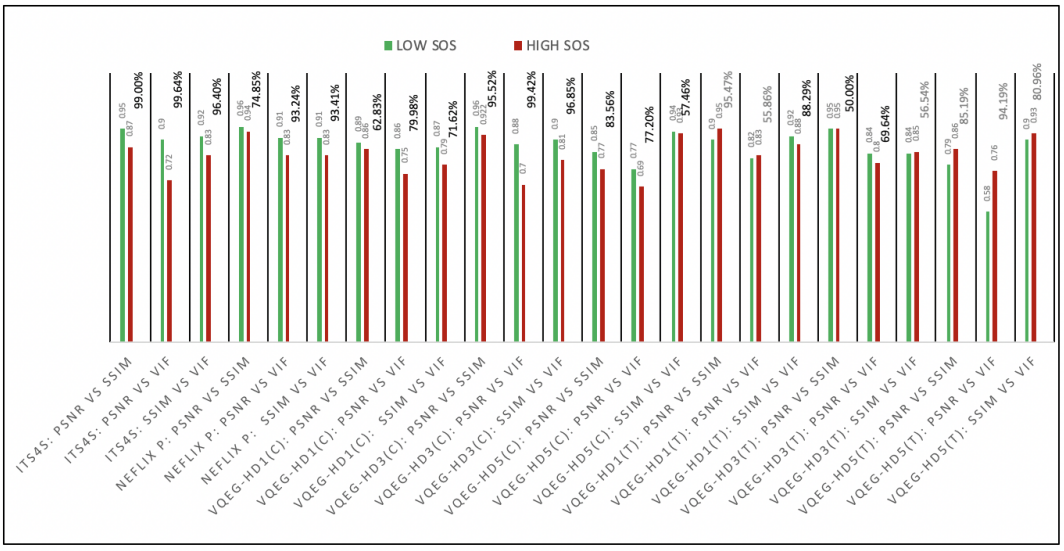



Figure 3.2: Correlation coefficient (Spearman rank order) between pairs of VQMs (PSNR, SSIM, VIF), in a given subjective experiment (the ITS4S, Netflix public dataset, VQEG-HD1, HD3 and HD5), when the PVSs with low (green) or high (red) SOS are considered. The statistical significance of the difference 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).

in ascending order of SOS values. Then, the Spearman Rank Order Correlation Coefficient (SROCC) and the Kendall Rank Order Correlation Coefficient (KROCC) were used to measure the alignment of the scores of three VQMs, i.e., the Peak Signal to Noise Ratio (PSNR) [147], the Structural Similarity Image (SSIM) [157] and the Visual Information Fidelity (VIF) [115], on the 50 PVSs having recorded the lowest SOS values as well as on 50 ones with the largest SOS values.

The three VQEG datasets used in the study contain PVSs the quality of which was impaired by both coding and transmission artifacts. While the ITS4S and the Netflix public dataset consider only coding distortion. Therefore, for the VQEG datasets, the 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 Figure 3.2 and Figure 3.3 for the SROCC and KROCC respectively. One can observe that in all the cases in which the PVSs are only affected by coding artifacts, the VQM scores show greater correlation on the set of sequences with the less diversification of opinion scores (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. This behavior might be explained by the fact that the considered VQMs

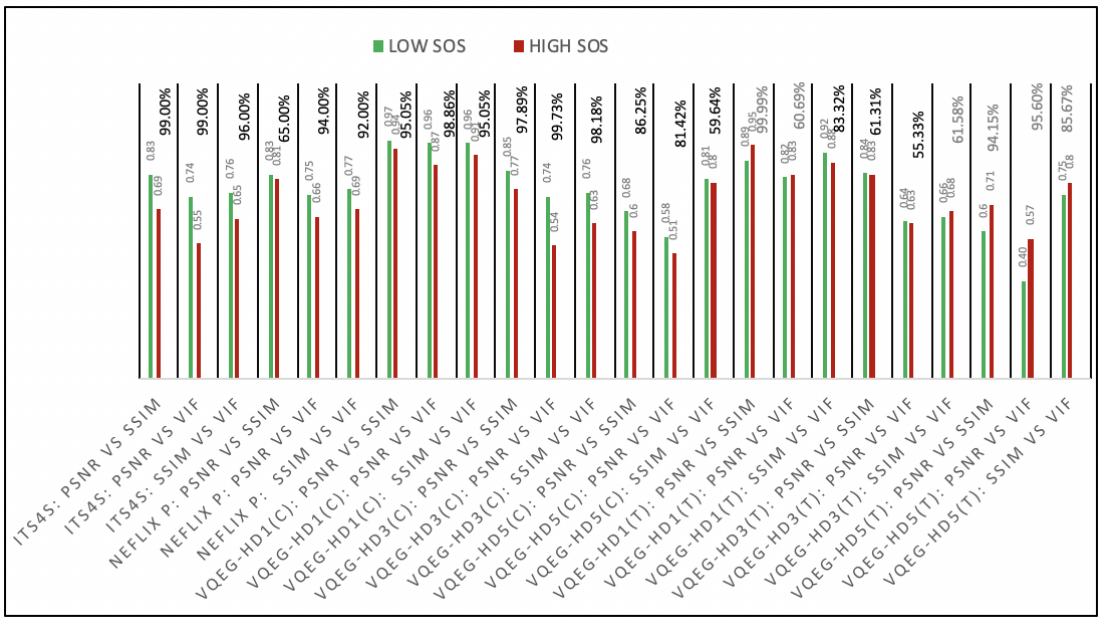



Figure 3.3: Correlation coefficient (Kendall rank order) between pairs of VQMs (PSNR, SSIM, VIF), in a given subjective experiment (the ITS4S, Netflix public dataset, VQEG-HD1, HD3 and HD5, when the PVSs with low (green) or high (red) SOS are considered. The statistical significance of the difference 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).

have empirically shown poor accuracy in handling transmission artifacts.

To make sure that the observations made from the analysis would be independent of the particular used datasets and that they are not the result of chance, statistical tests to determine how confident one should be in stating that a certain correlation value is greater than another were conducted. The percentages in Figure 3.2 and Figure 3.3 show these confidence levels for each pair of correlations under comparison.

For example, in the ITS4S dataset case, the correlation between the SSIM and the PSNR on PVSs with low SOS can be considered greater than the one obtained in presence of large SOS with 99% of confidence. Hence the difference between the two values cannot reasonably be considered as a result of chance. Similar large values of confidence are observed among all other pairs of VQMs for the ITS4S, the Netflix public dataset and the VQEG-HD3 dataset when restricting the analysis to PVSs with coding artifacts. 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 values are larger than those observed

in correspondence of large SOS values, the percentages of confidence are less than 95%.

In short, the study suggested that the degree of alignment between the PSNR, SSIM and the VIF scores, measured through the SROCC and the KROCC, is generally greater when calculated on PVSs whose quality is affected by coding artifacts and for which observers have expressed opinion scores characterized by little diversity. This supports the idea that the scores of different VQMs, when jointly exploited could provide information on the diversity among users' opinion scores at least in the case of PVSs whose quality is impaired by only coding artifacts. On the other hand, for the PVSs affected by transmission artifacts, this preliminary analysis does not allow such a conclusion. However, this does not preclude the existence of a more sophisticated measure of alignment than the SROCC and the KROCC between the VQMs that may explain the diversity of the observers' opinion scores when rating PVSs whose quality is impaired by transmission artifacts. Such a measure could be obtained for instance by fitting the VQM scores to the SOS using a highly nonlinear function as done later in Section 3.5.

### 3.4.2 The SOS Error Term

The SOS computed directly from the data gathered 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 the quality scale**: In general, the main focus of subjective tests is to assess the average perceptual quality in terms of MOS rather than the spread of opinions in terms of standard deviation [43]. When the standard deviation is needed, it is computed from opinion scores collected on a quantized Likert scale. Likert scales are useful to make sure that viewers understand what they are required to do so that they can rate the quality consistently. Unfortunately, this type of scale constitutes a source of noise when one is interested not only to the mean of the opinion scores, but also to their standard deviation. In a typical five points Absolute Category Rating (ACR) scale test, for a given PVS all viewers might select the same option, yielding to an integer MOS value and a SOS value equal to zero. This actually occurred in subjective tests even with 24 observers. The VQEG HDTV phase I test [139] can serve as a good example in this context. However, it is very likely 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 on the perceptual quality of a given PVS if a continuous scale was instead used.

2. **Limited number of viewers**: The statistics of the samples, such as the mean and the standard deviation are in general asymptotically consistent estimators. As the sample size increases, they become less unstable and converge to the exact value of the estimated parameters. Unfortunately, typical subjective tests are conducted with a limited number of subjects. In this case, the standard deviation of the opinion score gathered from a limited number of viewers 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. Therefore the inability to use a large number of raters in subjective tests generates a stochastic oscillation term that impairs the SOS value.

It is worth noting that the aforementioned two sources of error are to be taken into account when analyzing the diversity of opinion scores coming from any study that considered a limited number of subjects and used an ordinal Likert scale. Therefore, the approach presented in this chapter is not intended to be restricted to the media quality assessment community and it can be useful to explain the diversity among users' opinion scores in other research fields.

### 3.4.3 The SOS Model

In light of the discussion in Section 3.4.1 on the relation between the alignment of VQM scores and the diversity of users' opinion scores, the following hypothesis is formulated: the gtSOS of any given PVS ($gtSOS^{pvs}$) can be estimated from the quality scores of a certain number of VQMs computed on the PVS.

To that aim, the PSNR, the SSIM, the VIF [115], the Multi-Scale Structural Similarity Image (MS-SSIM) [143], and the Video Multimethod Assessment Fusion (VMAF) [87] were considered, and thus:

$$gtSOS^{pvs} = f(PSNR, SSIM, VIF, MSSSIM, VMAF) + \epsilon_{obj}^{pvs} \tag{3.3}$$

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

Now, taking into account the two sources of error discussed in Section 3.4.2, I proposed in [30] to model the standard deviation $SOS_{exp}^{pvs}$ of the subjects' opinion scores observed during a subjective experiment for a given PVS (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 limited number of subjects ($err_{exp}^{subj}$). This yielded the following SOS model:

$$SOS_{exp}^{pvs} = gtSOS^{pvs} + D_{exp}^{pvs}. \tag{3.4}$$

Summarizing, this model argues that the $SOS_{exp}^{pvs}$ observed for any PVS during a subjective test is a realization of a normally distributed random variable due to the $D_{exp}^{pvs}$ component, and has the mean equal to $gtSOS^{pvs}$ that is predictable from the scores of several VQMs modeling the characteristics of the sequence. Further insights into the validity of such a model are given in Section 3.5.

## 3.5 SOS Model Validation

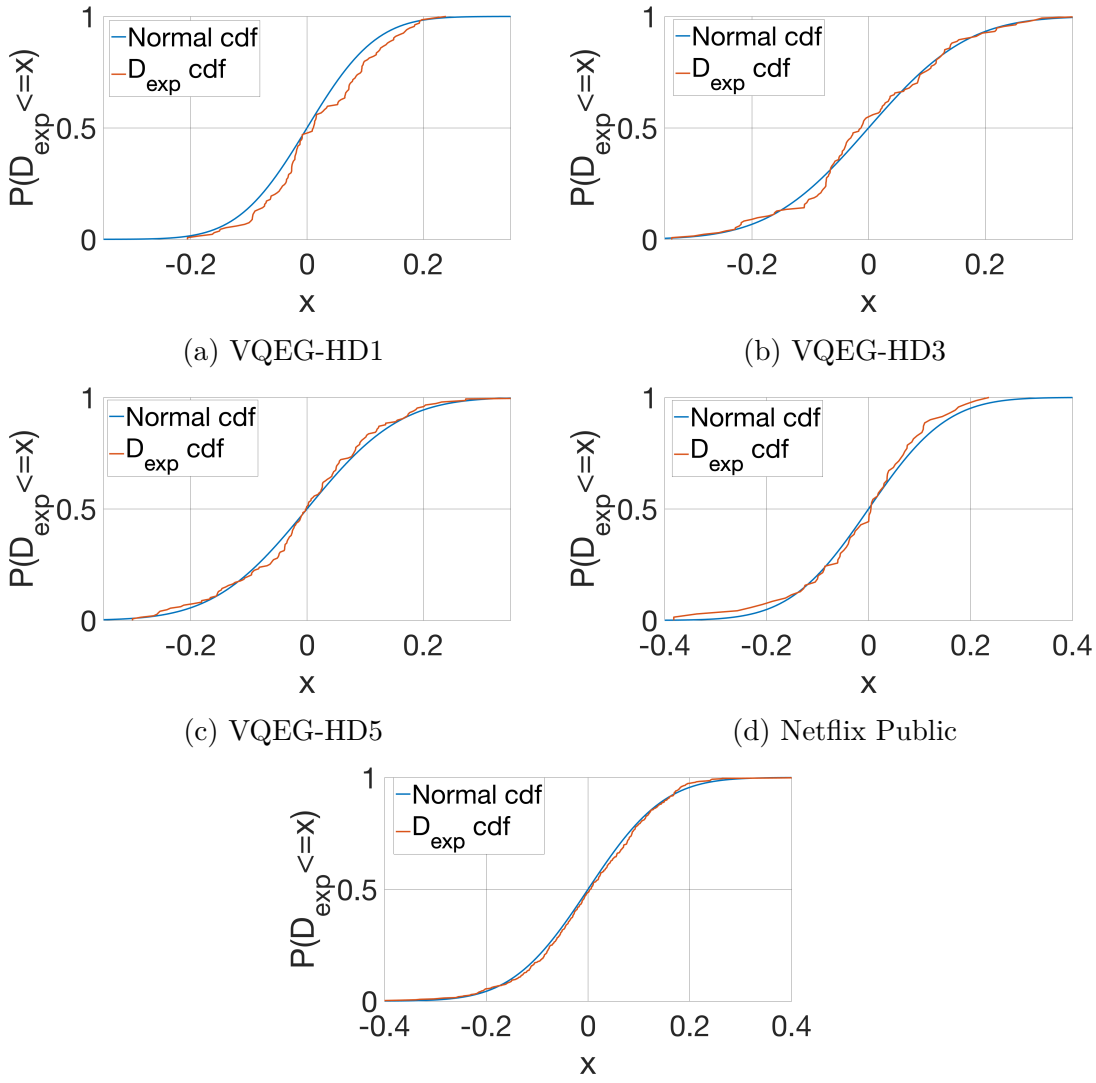
In this section, using NNs, an approximation of the function $f$ in Eq. (3.3) is derived. The validity of the SOS model in Eq. (3.4) is then shown by performing numerical experiments on several datasets.

To obtain an approximation of the function $f$, the VQM scores were regressed to the SOS observed during the subjective tests. An impressive number of ML algorithms to perform regression tasks has been proposed in the literature, however NNs-based models and support vector regression (SVR) have empirically demonstrated greater accuracy in the field of media quality assessment. To estimate the function $f$, both NN as well as SVR based models were tested. However, NNs were seen to be more effective, as they provided gtSOS predictions that showed better correlation to the SOS.

The NN that approximates the function $f$ is trained using the quality scores of the five VQMs considered in Eq 3.3 as input features, and the target is the noisy value $SOS_{exp}^{pvs}$. However, on the basis of the model in Eq. (3.4) and the Eq. (3.3), it is implicitly assumed that the stochastic component $D_{exp}$ of $SOS_{exp}^{pvs}$ is not predictable being a random error. Therefore, the NN prediction corresponds to an estimation of the the deterministic component of $SOS_{exp}^{pvs}$ and thus to the $gtSOS$.

To approximate the function $f$, only very simple architectures (single hidden layer with few neurons) were investigated because of the small size of the training sets. The use of a deep NN would have conducted to an overfiting of the the dataset. The obtained estimate of the gtSOS in that case, would therefore no longer be an intrinsic characteristic of the PVS since it would be affected by the two sources of error presented in Section 3.4.2, i.e., the noise due to the scale quantization and the limited number of viewers.

To determine the NN architecture that would perform best in approximating $f$, different numbers of neurons for the hidden layer were tested starting from two neurons. As the number of neurons increased there was more and more correlation between the predicted gtSOS and SOS on the training set. However, this was not the case when the NN was cross validated. In fact, the performance gap between




(a) VQEG-HD1

(b) VQEG-HD3

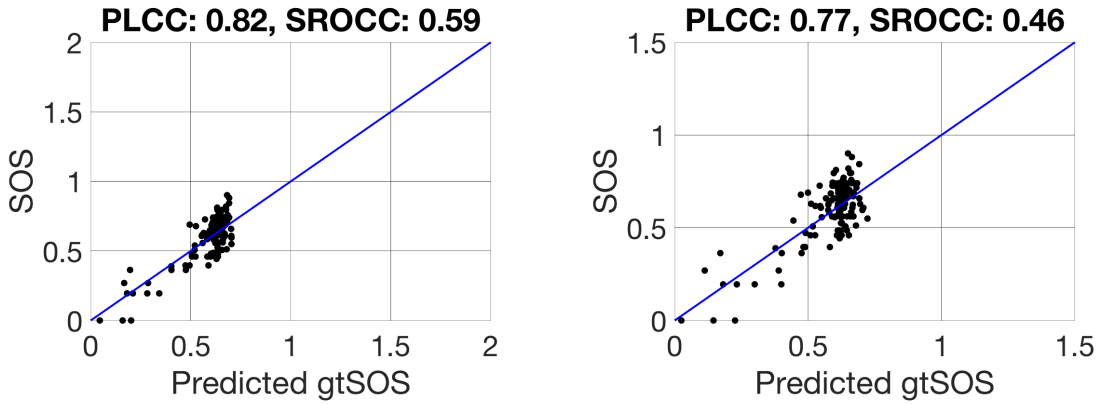
(c) VQEG-HD5

(d) Netflix Public

(e) ITS4S

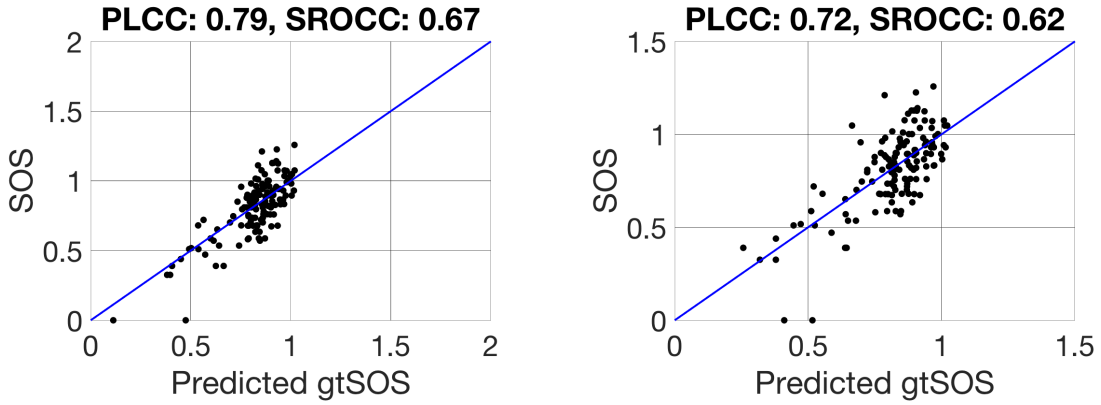
Figure 3.4: Comparison between the empirical cumulative distribution function (orange curve) of $D_{exp}$ and that of a Gaussian random variable having 0 as mean and similar standard deviation with $D_{exp}$ (blue curve). The analysis was done on five different datasets. The fact that the empirical cumulative distribution of $D_{esp}$, for each dataset, so closely approximates the cumulative distribution of the related Gaussian distribution shows that $D_{esp}$ can also be considered distributed according to this Gaussian distribution.

what was observed on the training set and what was obtained in cross validation became more and more important. However, the minimum performance gap between training and validation was achieved using five neurons on the hidden layer.




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

Figure 3.5: VQEG-HD1 dataset: the predicted gtSOS vs the SOS.



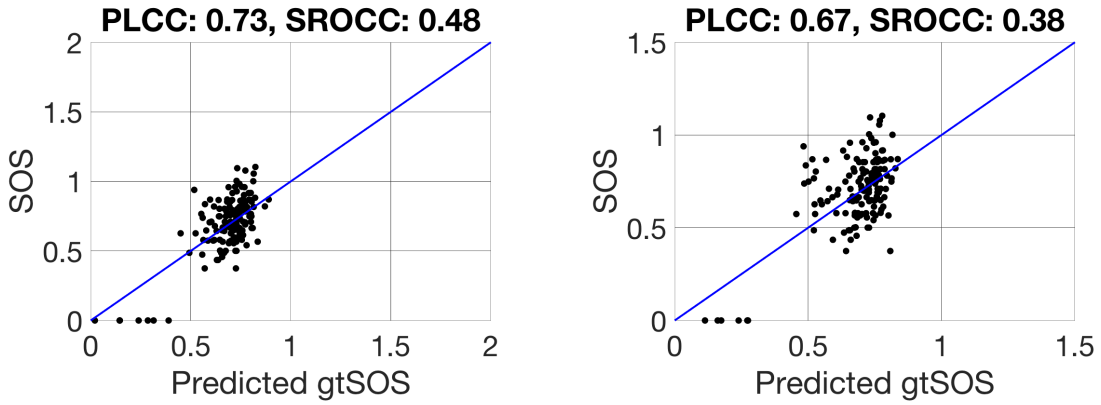

(a) Training set (b) 5-fold cross validation

Figure 3.6: VQEG-HD3 dataset: the predicted gtSOS vs the SOS.

Therefore the function $f$ was approximated with a NN having a single hidden layer with five neurons.

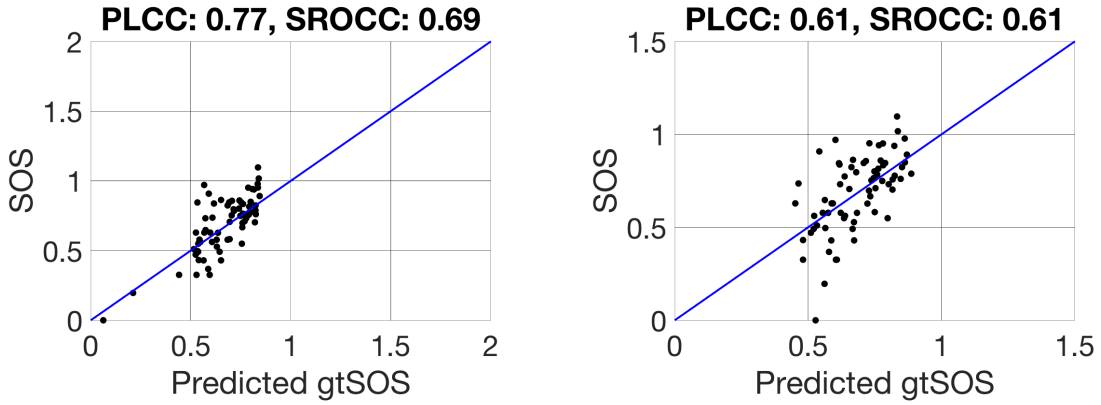
It is worth noting that the architecture of the network which approximates $f$ strongly depends on the amount of available training samples. The architecture used here is therefore not to be considered as an absolute reference, but rather as a valid architecture if one is working with a number of training samples similar to those considered in this chapter, i.e., 150 to 200 training samples.

To validate the model in Eq. (3.4), the function $f$ was estimated on five different annotated datasets, i.e. the VQEG-HD1, VQEG-HD3, VQEG-HD5, Netflix public and ITS4S dataset. Once the NN approximating the function $f$ is trained, it is possible to i) estimate the value of $gtSOS^{pvs}$ for each PVS, thus identifying content whose quality is intrinsically difficult to assess consistently (i.e., high $gtSOS^{pvs}$); ii)




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

Figure 3.7: VQEG-HD5 dataset: the predicted gtSOS vs the SOS.




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

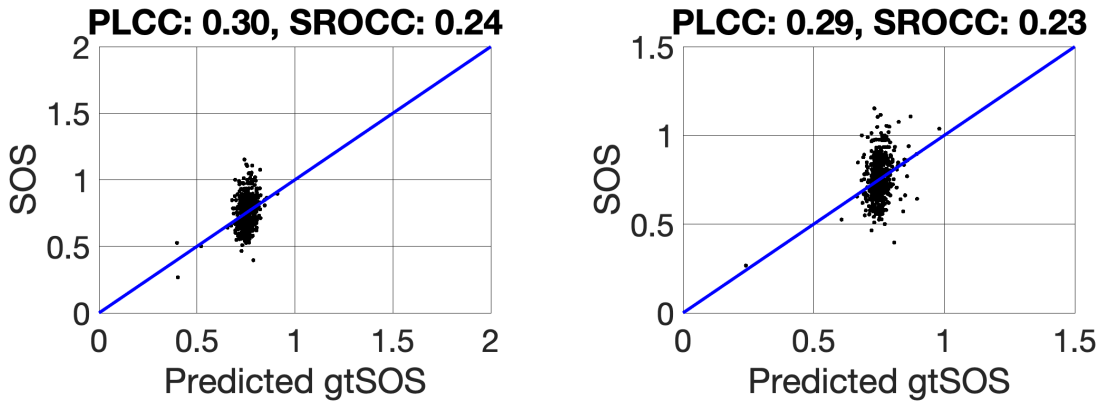
Figure 3.8: Netflix Public dataset: the predicted gtSOS vs the SOS.

deduce from Eq. (3.4) the value of the stochastic component $D_{exp}$ for each PVS.

In order to assess the fact that the stochastic component $D_{exp}$ of the SOS model is normally distributed, the following experiment was conducted. The empirical cumulative distribution of the set of $D_{exp}$ values of each dataset was computed and compared 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 Figure 3.4 presents the results of the experiment. In all the cases, the empirical cumulative distribution of $D_{exp}$ seems to be very well approximated by a Normal cumulative distribution. This is aligned with the assumptions of the SOS model proposed in Eq (3.4).

Figures 3.5, 3.6, 3.7, 3.8 and 3.9 show the correlation between the predicted gtSOS and the SOS for all the used datasets. On the various training sets, i.e.,




(a) Training set

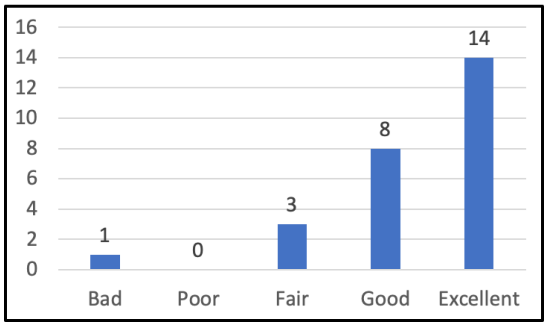
(b) 5 folds cross validation

Figure 3.9: ITS4S dataset: the predicted gtSOS vs the SOS.

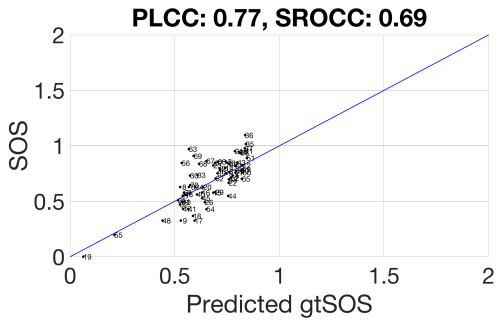
when training the NN using all the data in the dataset, the obtained Pearson linear correlation coefficient (PLCC) values range from 0.30, in the worst case, up to 0.82, whereas in cross validation the observed PLCC values range from 0.29 to 0.77. However, the SROCC values are a bit lower. In fact, on 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.

Statistical tests were performed to check whether the aforementioned PLCC and SROCC values 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 opinions expressed by different observers about the visual quality of a PVS using several VQMs scores cannot be rejected.

Lower PLCC and SROCC values were obtained in the case of the ITS4S dataset in comparison to those observed on the other datasets. This could be due to the fact that, unlike the other subjective tests considered in this chapter, the one that led to the ITS4S dataset was designed for the development of no-reference metrics. Therefore, during the experiment, the original source 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 case of the other datasets. Nevertheless, the obtained PLCC and SROCC values on the ITS4S dataset were still seen to be significantly greater than 0 with 95% of confidence.




(a) Subjective scores distribution for PVS #63




(b) Predicted gtSOS, with labelled PVSs

Figure 3.10: Analyzing the Netflix Public dataset. The value of $D_{exp}$ is large for the PVS #63. An inspection of the related distribution of opinion scores (left chart) revealed that an observer rated the quality of that PVS as "Bad" despite most of the test participants scored it as "Excellent".

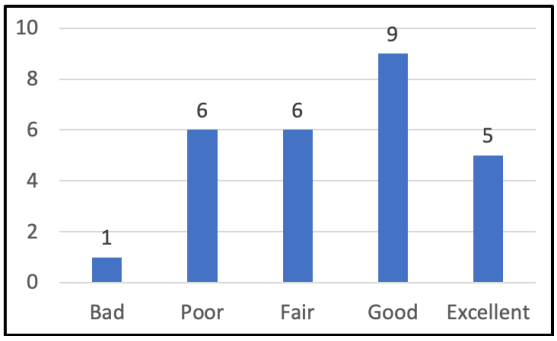
## 3.6 Application of the SOS Model to Anomalies Detection

This section illustrates an application of the SOS model. In particular, the model's capability to highlight potential anomalies in the data collected during a subjective test is showcased.

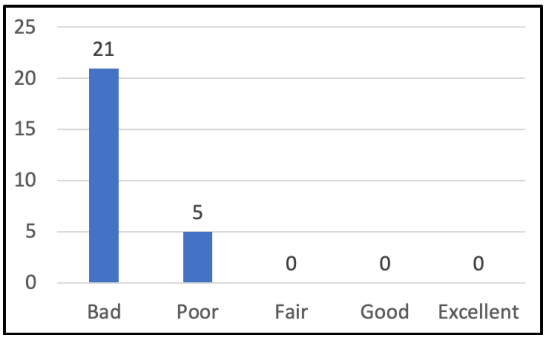
The issue of how to identify potential anomalies in the result of a subjective test is still open. These anomalies are typically caused by the use of peculiar source content or unexpected subjects' behaviors. For instance, a viewer may just assign random votes or the opinion scores gathered for a specific stimuli may be remarkably inconsistent. The presence of such anomalies negatively affect the accuracy of objective measures developed, relying on raw data collected during subjective tests. The typical approach adopted for anomalies detection is to model the observer opinion on each sequence using the normal distribution [49, 55, 68] 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 difficultly model the opinions' distribution for PVSs with very high or very low perceived visual quality as illustrated in Figure 3.10a, which shows the scores distribution for a specific PVS in the Netflix dataset.

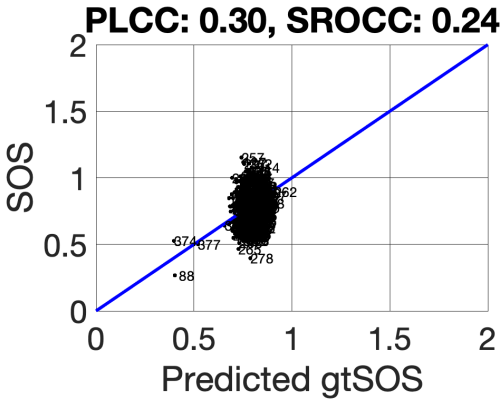
The problem can be approached differently if one relies on the SOS model described by Eq. (3.4). The term $D_{exp}$ in the model represents the part of the inconsistency in the votes introduced by the subjective test settings. As such, it also models the average inconsistency of the sample of people chosen for the test.




(a) Subjective scores distribution for PVS #257




(b) Subjective scores distribution for PVS #278




(c) Predicted gtSOS, with labelled PVSs

Figure 3.11: Analyzing the ITS4S dataset. The value of $D_{exp}$ is large for PVSs #257 and #278. The opinion scores for PVS #257 are almost uniformly distributed over the quality scale; this highlights the peculiar nature of the subjective evaluation of such a PVS. On the other hand, the analysis indicated that the low SOS value of the PVS #278 may not be a reliable estimation of its ability to generate diversity among viewers' ratings.

Therefore, an estimate of $D_{exp}$ would allow to determine the stimuli 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 a subjective test that uses a continue scale.

The approach to find potential anomalies using the SOS model can be summarized as follows. Starting from the data of the subjective test under examination, one estimates the function $f$ as discussed before, then from Eq. (3.4) and Eq. (3.3), for each PVS, one gets:

$$D_{exp} \approx SOS_{exp} - f(PSNR, SSIM, VIF, MSSSIM, VMAF) \tag{3.5}$$

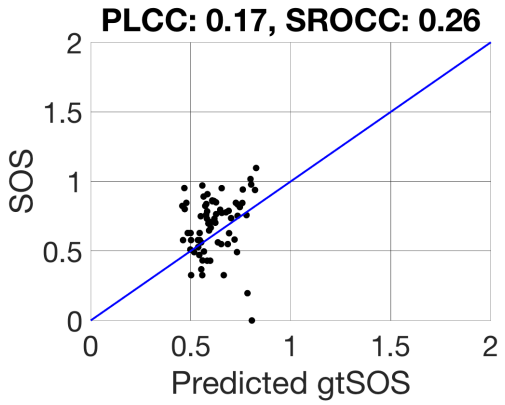
The values of $D_{exp}$ (for all PVSs) form a sample having a normal distribution with zero mean as indicated by the SOS model in Eq. (3.4). Finally, the PVSs, whose evaluation might be affected by anomalies, are those for which the estimated $D_{exp}$ value is an outlier of this distribution. In practice, denoting with $D_{exp}^{pvs}$ the value of $D_{exp}$ for a given PVS and by $std_{D_{exp}}$ the standard deviation of $D_{exp}$, it is suggested to give a closer look to the ratings of each PVS for which:

$$\left|D_{exp}^{pvs}\right| > 3 \cdot std_{D_{exp}} \tag{3.6}$$

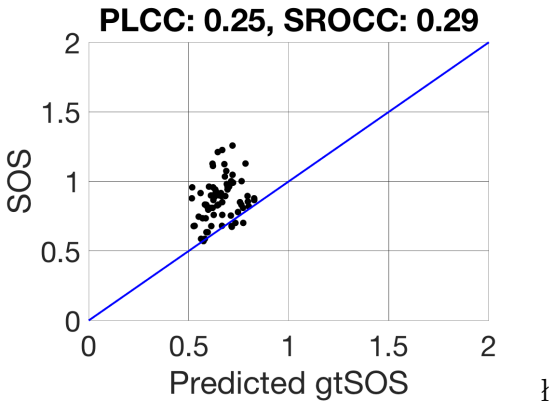
and carefully consider an examination of the opinion scores gathered for each of those PVSs before using the data.

In order to investigate the effectiveness of the method in practice, it was tested on the Netflix public dataset and the ITS4S dataset.

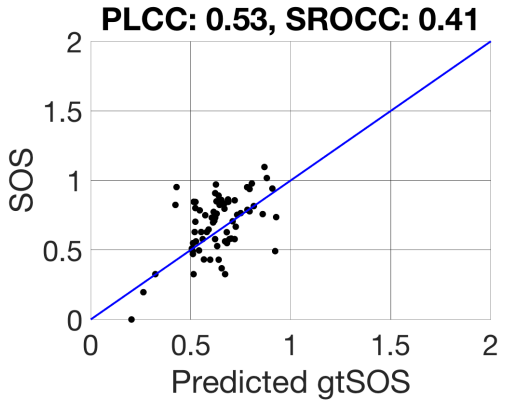
Figure 3.10, shows again the comparison between the predicted gtSOS and the SOS after determining the function $f$ on the Netflix public dataset. The PVSs were numerically labeled to facilitate the interpretation of the results. For each PVS, $D_{exp}^{pvs}$ is estimated by subtracting the predicted $gtSOS^{pvs}$ from the $SOS_{exp}^{pvs}$. Consider, for instance, PVS #63 for which the condition in Eq. (3.6) holds. The ratings collected in the subjective test are shown in Figure 3.10a. For such a PVS, even if the mode of the distribution of the subjects' opinion scores is equal to 5 ("Excellent") and 22 observers out of 26 ranked 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 one 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 Figure 3.11, the scores collected for PVS #257 and #278 that exhibit a high value of $|D_{exp}|$ were also analyzed. The individual subjects' ratings for PVS #257 (shown in Figure 3.11a) are almost uniformly distributed between "Poor" and "Excellent" leading to an observed SOS value, which is significantly larger than the predicted gtSOS. That suggests the intrinsic difficulty of evaluating this PVS should be lower than what has been observed. Therefore, its characteristics should be investigated in more details. On the contrary, for PVS #278 (shown in Figure 3.11b), 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.



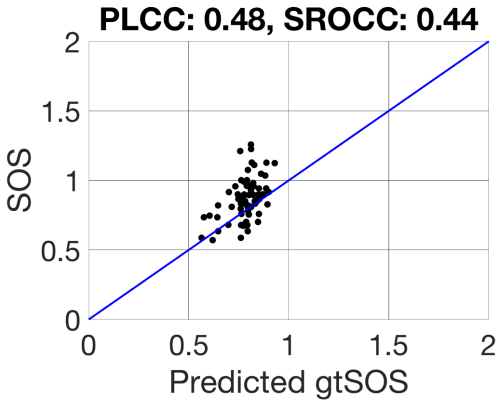

(a) Netflix public dataset, no data augmentation




(b) VQEG-HD3, no data augmentation




(c) Netflix public dataset, with data augmentation




(d) VQEG-HD3, with data augmentation

Figure 3.12: Assessing the performance of the deep NN based model when estimating the gtSOS with (bottom) and without (top) the data augmentation. The NN was trained using only the VQEG-HD1 and VQEG-HD5 datasets (coding artifacts only).

## 3.7 Deep Neural Network-based Prediction of the gtSOS

To assess the validity of the model in Eq. (3.4) for the SOS values obtained in a subjective test, the analysis has been done so far separately for each dataset. This section instead, focuses on the training of a NN that can be used to predict the gtSOS in a general context. The aim is therefore to attempt to train a model that can provide hints about the uncertainty that characterizes the perceptual quality of a PVS independently from the context in which it is rated.

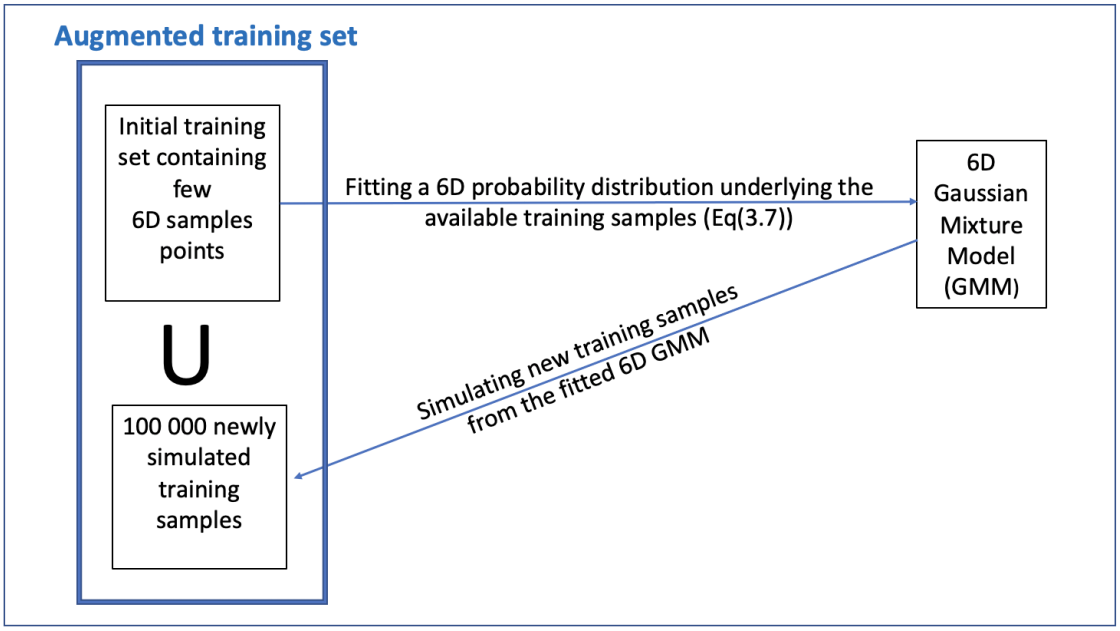



Figure 3.13: The diagram summarizes the data augmentation approach described in Section 3.7.1. A 6D Gaussian Mixture Model (GMM) is used to fit the multidimensional probabilistic distribution underlying the point cloud of the initial training samples. From the fitted GMM, 100,000 realizations are simulated. These realizations are then combined with the initial training set to obtain a greater number of training samples.

To train such a NN, the data gathered during the VQEG-HD1 and VQEG-HD5 experiments were selected as training set. Only the PVSs whose quality was impaired by coding artifacts were considered. This restriction was necessary since the VQMs used as features have empirically shown higher accuracy in assessing the quality of PVSs corrupted by this type of artifacts only.

### 3.7.1 Data Augmentation

As pointed out in Section 1.2.3, the size of freely available subjectively annotated datasets in media quality assessment does not allow to effectively train on them deep NN-based models. A data augmentation approach was therefore designed in order to get more training samples and hence to enable the use of a deep NN for the gtSOS estimation. The main steps behind such an approach are explained in the next paragraphs.

Each data point in the training dataset was considered to be a realization of a 6-dimensional random vector: $(PSNR, SSIM, VIF, MSSSIM, VMAF, SOS)$. This is in line with the model in Eq. (3.4) that explicitly considers the $SOS$ for each

PVS as a random variable. This, coupled with the potential inaccuracy of the used VQMs in some situations, suggests that data points available in the training dataset can be considered as realizations of a 6-dimensional random vector. Starting from this observation, the multivariate distribution from which the training set samples derive was fitted. Then it was used to simulate more samples for training the deep NN for gtSOS prediction.

Note that by proceeding in this ways, the influence of the settings of the specific subjective test chosen for the training is reduced. Hence, one expects to get an estimation of the gtSOS that would account only for the PVS' characteristics as desired.

The joint probability distribution $g$ of the available training samples was modeled with a 6-dimensional Gaussian Mixture Model (GMM), i.e.

$$g(VQM_{pvs}, SOS_{exp}^{pvs}) = \sum_{i=1}^{k} \pi_i \cdot N\Big((VQMs, SOS_{exp}^{pvs})|\mu_i, \Sigma_i\Big) \tag{3.7}$$

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 $\mu_i$ and covariance matrix $\Sigma_i$ and $k$ is a number of components of the GMM. The parameters ($\pi_i$, $\mu_i$, $\Sigma_i$ and $k$) of the GMM are estimated using a maximum likelihood estimation approach.

Denoting by M the number of PVSs in the training set, the following optimization problem was solved to estimate all the parameters and obtain the desired joint probability distribution of the training samples:

$$(\pi_i, \mu_i, \Sigma_i, k) = \arg\max \left(\prod_{s=1}^{M} \left(\sum_{i=1}^{k} \pi_i \cdot N\Big((VQMs_{pvs}, SOS_{exp}^{pvs})|\mu_i, \Sigma_i\Big)\right)\right) \tag{3.8}$$

where $VQMs_{pvs}$ is the vector containing the scores of the five VQMs computed on each PVS. The optimization problem in Eq. (3.8) was solved by using the Expectation-Maximization (EM) algorithm [84].

Once the joint distribution of the training samples was obtained, to augment the data for the training process, 100,000 more samples were simulated from it yielding a very large number of training samples. The diagram in Figure 3.13 summarizes the whole process.

### 3.7.2 The Network's Architecture and the Training Process

The availability of a large set of training samples enabled the exploitation of the prediction capability of deep NNs that would otherwise have led to overfitting if trained on the initially available limited size datasets.

Extensive numerical experiments to determine the deep NN architecture that best fits the task requirements were conducted. The best results were obtained using a NN with 5 neurons on the input layer, i.e. one for each VQM score, 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.

### 3.7.3 Results and Discussion

To evaluate the effectiveness of the deep NN based model, it was tested on the Netflix public dataset and the VQEG-HD3 dataset that have not been used during the training process. The results are shown in Figure 3.12 (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, the obtained correlation values are not a result of chance as they are greater than zero with statistical significance. Hence, the approach described in this section represents a promising preliminary step toward the capability to automatically predict the intrinsic ability of a video sequence to confuse viewers independently from the context in which it is evaluated.

It is important to notice that the accuracy of ML-based model for this task can be further improved if it would be possible to use data from a subjective tests designed specifically to create good predictors for the gtSOS value. This is unfortunately not the case for typical subjective tests 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 ML algorithms for gtSOS prediction, a sufficiently uniform coverage of the 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 the data augmentation approach, i.e. the simulation of more training data points by the fitted GMM, a shallow NN, having the structure presented in Section 3.5, was trained. The training was done using the VQEG-HD1 and VQEG-HD5 datasets without the data augmentation, i.e., without simulating more training data from the GMM. When testing this NN on the Netflix public dataset and VQEG-HD3 dataset, the results shown in Figure 3.12 (top part) were obtained. 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 supports the hypothesis that gathering enough data during a subjective test specifically designed for gtSOS modeling and prediction would potentially improve predictive models performance.

## 3.8 Conclusion

The study presented in this chapter 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 average perceptual quality, can also be a helpful tool in analyzing the data coming from subjective experiments in order to identify, for instance, anomalies or peculiar behaviors.

The analysis focused on analyzing and modeling the diversity observed among the subjects' opinions in subjective experiments. In particular, a model of the standard deviation of the ratings of different observers on single PVSs was provided. Such a model argues that the standard deviation of viewers' opinion scores is distributed according to a normal distribution whose mean, referred to as the ground truth SOS, can be effectively estimated by exploiting the quality scores of a set of video quality measures.

Relying on this model, it was shown 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. Furthermore, it was observed that it is possible to train neural networks that, taking the scores computed by several video quality measures as an input, can predict how much diversity would be observed among subjects' votes if the PVS would be subjectively evaluated. When trained and cross validated on the same dataset, these neural networks led to a prediction that is significantly correlated with the standard deviation observed in the actual subjective test.

Finally, by applying a data augmentation approach, it was 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 scores of several video quality measures computed on that PVS. This deep neural network provided predictions showing a 0.5 correlation with the actual SOS value. This correlation is demonstrated to be statistically significantly different from zero with 95% of confidence. This shows that the used features can explain to some extend the variability of users' opinions. Hence, the ML-based approach looks promising and future work in the same direction would contribute to refine it.

Just like any neural networks-based model, the approach to assess the intrinsic complexity of a given PVS in terms perceptual quality prediction presented in this chapter yields black box models. In fact it is not very clear how the trained neural networks make use of the scores of the video quality measures taken as input to

provide an estimate of the ground truth SOS. Therefore, in the next chapter, a more intuitive approach to figure out video sequences whose quality is difficult to assess will be presented.

# Chapter 4

# Estimating the Accuracy of Subjective Score Prediction through the Disagreement of Video Quality Measures

## 4.1 Introduction

A fundamental objective for content providers and content aggregators is to guarantee high quality of experience (QoE) to their customers. The last decades have therefore witnessed numerous publications that have proposed novel algorithms to generate video quality measures (VQMs) that can predict the mean opinion score (MOS [15, 104]. Quite often, significant differences occur between the MOS values predicted by these different VQMs, for the same processed video sequence (PVS). Unfortunately, works that investigated whether any useful information is obtainable about the accuracy of VQMs from their disagreement are still lacking in the literature.

In the following journal paper [130], I proposed an index to measure the level of disagreement between the VQMs when used to measure the quality of a given PVS. I then showed that this index is able to distinguish between PVSs for which the VQMs are expected to accurately predict the perceptual quality and those on which the VQM predictions are likely to be inaccurate. This Chapter introduces the reader to such an index and presents the results of some numerical experiments showing that index' effectiveness.

The proposed index has the potential of being very useful in academia and industry, since it can determine if predictions made by VQMs are accurate or not.

In academia, this index will facilitate the creation of effective tooling to identify appropriate subsets of PVSs to be used in subjective tests. It is useful for two

additional reasons. Firstly, it saves time and resources by excluding from subjective experiments, PVSs whose end-user scores are accurately predictable using VQMs alone. Secondly, it can be used in identifying problematic PVSs for which VQMs are poor at predicting the end-user scores. Results from experiments using such PVSs are typically of great value to researchers.

In the media industry, it is of primary importance to be able to quickly and automatically identify the PVSs on which the quality predictions provided by the VQMs could be misleading. Misleading quality predictions often result in unexpected degradation of customers' QoE through inadequate resource provisioning. The index presented in this Chapter is the outcome of a collaborative work with a global media company. This index was therefore intended to be used internally by the company to figure out peculiar PVSs from the point of view of perceptual quality assessment.

To perform the analysis that yielded the proposed index, a dataset comprising 368 industry grade PVSs was created. Industry-grade (mezzanine format) content is minimally compressed during data acquisition [109]. This dataset differs from other widely used video quality datasets, which are typically built by using pristine-quality content and acquired without any compression. In media industries, content is usually of the mezzanine format, which is of high quality but not pristine. A decision was made to work with industry grade content to closely replicate the conditions encountered in actual media industry processing chains.

The following VQMs were considered: Peak Signal to Noise Ratio (PSNR) [147], Structural Similarity Index Measure (SSIM) [157], Multi-Scale Structural Similarity Index Measure (MSSSIM) [143], Visual Information Fidelity (VIF) [115], Extended Weighted Peak Signal-to-Noise Ratio (XPSNR) [41], Video Multi-method Assessment Fusion (VMAF) [87], and two proprietary VQMs internally used by the media company, i.e., PVQM1 (the first proprietary VQM), and PVQM2 (the second proprietary VQM). Due to corporate legal considerations, the full names of the two proprietary VQMs cannot be mentioned.

There are newer VQMs than the ones listed above, some of which are presented in the ITU recommendation P.1203 [52], and others are based on deep learning approaches. The academic and industry communities have not yet adopted these VQMs on a large scale since many of them have not yet been tested in real-world environments. As such, the focus of the study was not on these more recent VQMs. Unlike the newer VQMs, the VQMs considered in this chapter are those typically used by academic researchers for designing and evaluating state-of-the-art video processing applications [71, 94, 25, 83, 67]. Therefore, an index, as the one presented in this chapter, that provides information on the accuracy of those VQMs, is of great interest for the scientific and industrial community.

To compute the value of the proposed index, all the VQM scores were mapped onto the same scale. For each PVS, it was counted the number of unique VQM pairs from the collection of possible VQM pairs, where one VQM provided a quality

prediction that was perceptually different from the other VQM of the pair. This number, expressed as a fraction, was shown to be an effective index for measuring the accuracy of the VQMs. In other words, if many VQMs disagree on the perceptual quality of a given PVS, then each VQM is also likely to wrongly estimate the MOS of that PVS. While some standardised techniques for comparing VQMs [53] already exists, it is important to note that the index presented in this chapter aims at investigating the implications of VQMs disagreement rather than comparing the VQMs.

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

To assess the effectiveness of the proposed measure, a small-scale subjective experiment was carried out on a subset of PVSs characterized by both low and high VQMs disagreement. The results showed the effectiveness of the proposed index in deducing the accuracy of VQMs.

This chapter is organized as follows. Section 4.2 presents a short review of previous works exploiting the agreement and/or disagreement of a set of VQMs. Section 4.3 provides a description of the dataset used for the analysis. Section 4.4 details the proposed VQMs disagreement index. Section 4.5 describes the subjective experiment setup. Results are discussed in Section 4.6. Section 4.7 draws final conclusions.

## 4.2 Related Work

The idea of leveraging many VQMs together in order to deliver more accurate prediction of perceptual quality has been investigated in the literature [74]. It has been shown that a machine learning (ML) model that takes, as input, a set of different VQMs computed on a given PVS, can yield improved quality predictions as opposed to using only single VQM. In [70], the authors designed a support vector regression model that jointly utilizes several VQMs to provide a more accurate MOS estimations. The work presented in [127] argues that PVSs whose sources are characterized by a low spatial activity index are challenging to work with from the point of view of objective quality assessment. In that work, a neural network-based model was proposed to address such challenges. The model relied on the scores from many full-reference metrics in addition to the spatial and the temporal activity index to mitigate the inaccuracies of single VQMs when estimating the

quality of these PVSs. By feeding a ML based model with many different VQMs, the authors aimed at exploiting the diversities and/or similarities between the VQM scores in order to reach a better MOS estimation.

The approach of using the differences between the predictions of many VQMs has not been exploited solely for accurate MOS estimations. In fact, in [30] the authors showed that the agreements between different VQMs, as measured by the Spearman and the Kendall rank order correlation coefficients, were related to the standard deviation of subjective ratings for a given PVS. They designed a neural network-based model that takes as input five VQMs and estimates the diversity among users' ratings. Still focusing on the quality scores as predicted by different VQMs, in [29], the authors proposed an approach based on Gaussian mixture models to find the range of quality values to which the MOS of a given PVS is expected to belong with a given probability.

In all the papers mentioned so far, the VQMs were studied together with ML models to enhance some aspects of the quality assessment processes. Despite the useful results reported in all these papers, their use of ML models means that they relied on black box models whose internal workings might not be trivial or easy to understand.

Instead of using ML models, some other authors have exploited the information associated with the diversity or similarity between VQM scores in a more intuitive and easier to interpret way. In [5] and [4] the authors investigated the disagreements between PSNR, SSIM and the VIF at the frame and sequence level. In both works the authors analyzed the behavior of the three metrics on a given pair of PVSs. They evaluated, for different source content, the ability of these metrics to coherently rank the perceptual quality of a pair of PVSs.

The approach presented in this chapter differs from those in [5] and [4] in that the proposed index for measuring the VQMs disagreement focuses on pairs of VQMs instead of PVSs, thus yielding an indicator that determines how difficult it is to assess the quality of a given PVS using VQMs. A small-scale subjective experiment was used in validating this concept, and the results showed that such a simple indicator could provide relevant information regarding the ability to accurately predict the perceptual quality of a PVS without resorting to a subjective experiment.

Another fundamental difference between the study presented in this chapter and many others in the literature is the inclusion of proprietary VQMs. Researchers typically use open-source tools to benchmark their proposals. As such, VQM comparison studies have mostly focused on freely available VQMs [116]. However, in some cases, open-source software are not properly optimised for effectively operating in real-world scenarios. There is just a small number of published works that have conducted studies involving proprietary VQMs [78]. Therefore, this study also contributes in shedding light on the existence of a potential gap between the accuracy of well-known and widely used open-source VQMs and proprietary ones.

## 4.3 Dataset Description

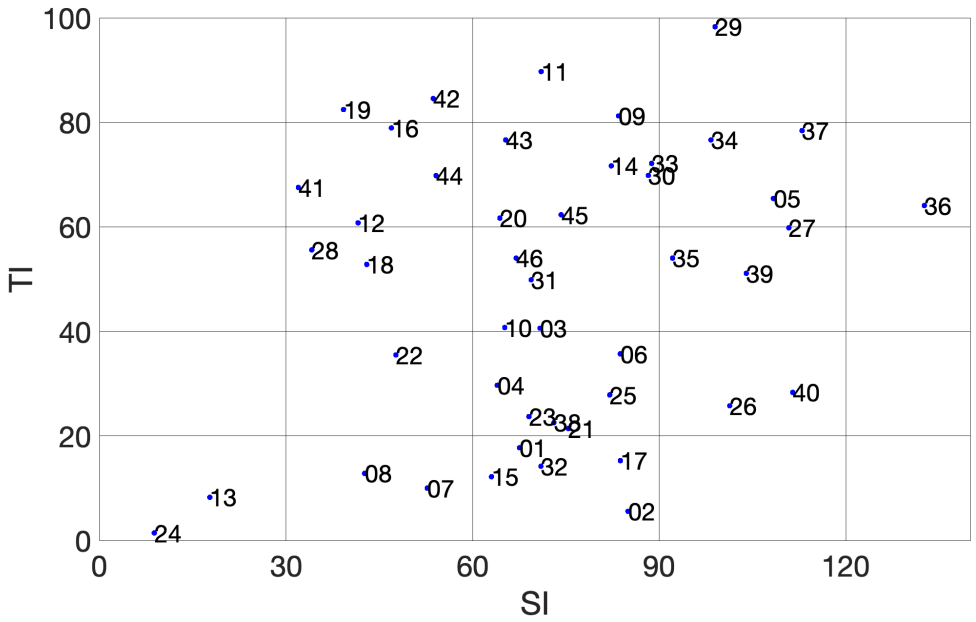



Figure 4.1: Evaluating the heterogeneity in terms of the temporal and spatial activity index of the used 46 sources. The labels identify different sources.

To create the dataset used for the analysis, a total of 46 Full HD (FHD) industry grade source videos were selected according to guidelines in [97]. These comprised a range of entertainment videos including sports, movies and animations. Depending on which region (Europe or US), the video frame rates per second (fps) were either 23.976, 25.000 or 29.970. Figure 4.1 shows the selected sources covered a wide range in terms of Spatial Information (SI) and Temporal Information (TI) as recommended in [92].

The source videos were encoded using H.264/AVC constant bit rates. The Apple's HLS authoring specification [10] was used as guidelines in producing the eight hypothetical reference circuits (HRCs) summarized in Table 4.1. Some of the key encoding configurations included one-pass encoding preset, the instantaneous decoder refresh (IDR) interval was set to two seconds, with an option of inserting an I-frame if there was a scene change within a given IDR interval. The size of the video buffer verifier was set to 5 seconds and the deinterlacing mode was set to motion adaptive interpolation. A summary of the bit rates and resolutions are given in Table 4.1.

From each of the 46 source videos, eight PVSs were created resulting in a total of 368 PVSs. The PVSs in the dataset were also divided into two main categories, namely movies and sports. For sports content in Europe, the frame rates were

interpolated from 25.00 fps to 50.00 fps. For sports content in the US, the frame rates were interpolated from 29.97 fps to 59.94 fps. This was done to reduce judder during playback, caused by camera panning movements. The frame rates for the movie content were untouched, so they were the same as the source videos.

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

The video quality of the 368 PVSs were evaluated using the eight considered VQMs - PSNR[147], SSIM[157], MSSSIM[143], VIF[115], XPSNR[41], VMAF[87] and the two proprietary VQMs PVQM1 and PVQM2.

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

All eight VQMs considered in this chapter are full reference VQMs, i.e., they evaluate the quality of a distorted signal by comparing it to the source. PSNR measures the quality of the distorted content by deriving its mean square error (MSE) with respect to the source pixels. SSIM evaluates the similarity between the source and the distorted signal by considering three main aspects, namely the luminance, the contrast and the preservation of the structures. MSSSIM implements the same steps as SSIM but at multiple scales. VIF uses natural scene statistics models to define the image information perceived by the human visual system (HVS). It then quantifies the amount of information shared between the source and the distorted signal. XPSNR is an enhancement of PSNR, which uses a *distance* between the source signal and the distorted signal considering some characteristics of the human

Table 4.1: The 8 different Hypothetical Reference Circuits (HRCs) used to generate the 368 $(46 * 8)$ PVSs of the dataset.

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

vision system which are not considered when using the MSE alone. VMAF fuses together multiple elementary full reference metrics using machine learning. The rationale behind VMAF is that each elementary metric may have its own strengths and weaknesses with respect to the characteristics of the source video, the type of artifacts, and the degree of distortion. VMAF seeks to preserve the strengths of the individual metrics and to deliver a more accurate final score.

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

Note that PSNR, SSIM, MSSSIM and VIF were originally developed for assessing the quality of still images. However, due to their analytical properties and low complexity, they are also the most used metrics for monitoring quality when designing video processing applications [71]. PSNR is even considered a kind of baseline in the context of video quality assessment. The Video Quality Experts Group (VQEG) [136] for instance, often uses PSNR as a benchmark for validation experiments, as was done during the performance evaluation of full reference VQMs in the HDTV experiment [139]. Many papers have considered the PSNR, SSIM, MSSSIM in their analysis [120, 41, 105]. The study in this chapter examines their accuracy with respect to that of VQMs used to optimize the delivery pipeline of media companies.

## 4.4 An Index for Measuring the VQMs Disagreement

This section is devoted to the definition of the proposed index for measuring the VQMs disagreement. Such an index enables the establishment of whether a VQM would accurately estimate the perceptual quality of a given PVS as it will be shown in Section 4.6.

Let denote by $D_{pvs}$ the value of the desired index for a given PVS. To formally define $D_{pvs}$, let introduce the following notation:

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

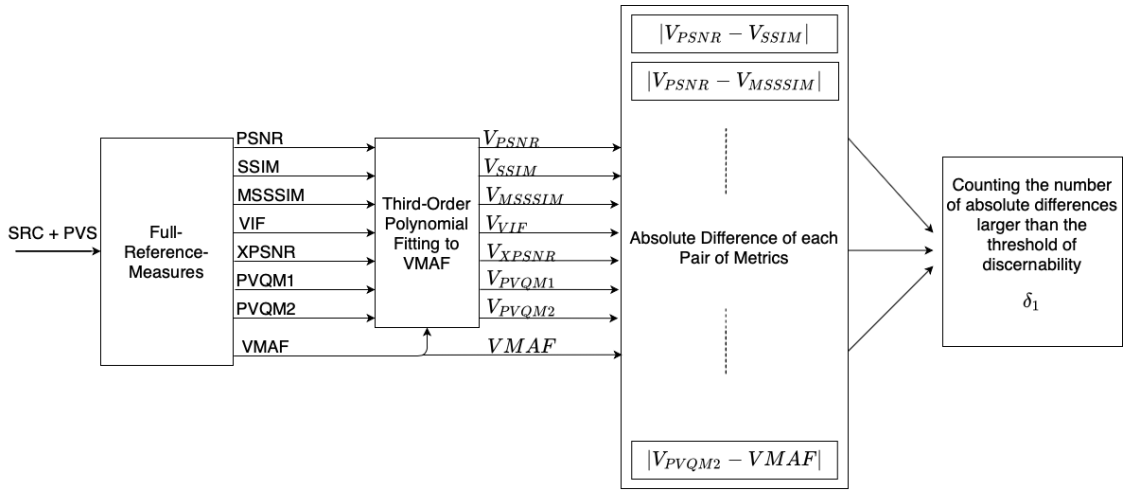



Figure 4.2: The diagram summarizes the implementation steps of the proposed disagreement index. VMAF is chosen as the reference VQM, hence, the VQM sensitivity $\delta_1$ is set to 7. $V_{PSNR}$ is the quality score obtained after performing a least square fitting of the PSNR to the VMAF scale using a third-other polynomial function. The same definition holds for all the other VQMs. By considering eight different VQMs, in total, 28 absolute differences were computed that corresponded to the number of unique pairs of VQMs that can be formed by selecting two VQMs from the eight available.

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

- $f_i \quad (i = 1,2,\ldots,n)$ be for mapping each $VQM_i$ from its original scale to the $VQM_1$ scale.
- $\delta_1$ denote the $VQM_1$ sensitivity, which is the minimum variation in quality perceptible by most human viewers if the quality were to be predicted using $VQM_1$.

For instance, it has been empirically observed that two pictures having VMAF scores that differ by less than seven points are likely to be judged as equal in terms of perceptual quality [88]. Therefore, for VMAF, the $\delta$ would be seven. The consideration of the VQM sensitivity is not a peculiarity of this study; similar approaches have already been proposed in the literature [51].

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

$$D_{PVS} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{1}_{\left(\left|f_i\left(vqm_i^{pvs}\right) - f_j\left(vqm_j^{pvs}\right)\right| > \delta_1\right)}}{\binom{n}{2}} \tag{4.1}$$

where $\mathbb{1}$ is the indicator function, whose value is 1 if the subscript proposition is true and 0 otherwise.

The denominator in Eq (4.1) is the total number of unique pairs that can be formed using the $n$ VQMs. The numerator counts the number of these pairs for which the two VQM scores that constitute the pair disagree on the perceptual quality of the PVS. Two VQMs are said to disagree when the absolute value of the difference between the predicted quality scores (using the reference metric scale) is greater than $\delta_1$.

In the context of this chapter, VMAF was chosen as the reference VQM and $\delta_1$ was therefore set to 7. Furthermore, the mapping functions were computed by performing a least square fitting of each of the other VQM scores to the VMAF scale using third-order polynomial functions [53]. The diagram in Figure 4.2 summarizes the implementation steps for the computation of the proposed index.

For any PVS, $D_{pvs} \in [0,1]$. The closer the value of $D_{pvs}$ is to one, the larger the disagreement between the VQMs regarding the perceptual quality of the PVS.

The main contribution of this chapter is the following statement. *The larger the value of $D_{pvs}$ for a given PVS, the more likely it is that VQMs will be inaccurate when assessing the perceptual quality of that PVS.* To verify such a statement, a subjective experiment whose details are provided in the next section was conducted.

## 4.5 A Small Scale Subjective Experiment

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

Since the main goal was that of investigating the implications of VQMs disagreement, viewers were shown PVSs on which the VQMs strongly agreed and those for which the VQMs strongly disagreed.

The VQMs disagreement index $D_{pvs}$, as described in Section 4.4 was computed for each of the 368 PVSs in the dataset. Afterwards, the PVSs were sorted in ascending order of $D_{pvs}$. From this, the following were found: i) at the lowest scale, 31 PVSs had $D_{pvs} < 0.2$ ii) at the highest scale, 36 PVSs had $D_{pvs} > 0.6$. These PVSs at the lowest and highest scales were selected for the subjective test dataset. In addition to these 67 PVSs $(31 + 36)$, 16 additional PVSs were added onto the dataset to ensure viewers evaluated a dataset whose perceptual qualities covered the entire quality scale, as this is a good practice in designing subjective experiments.

A total of 16 subjects (viewers) working in the media industry participated in this subjective experiment across two laboratories in Italy and Germany. The Double Stimulus Impairment Scale (DSIS) method was used (see Chapter 1). The DSIS method closely follows how most of the full reference VQMs operate; that is

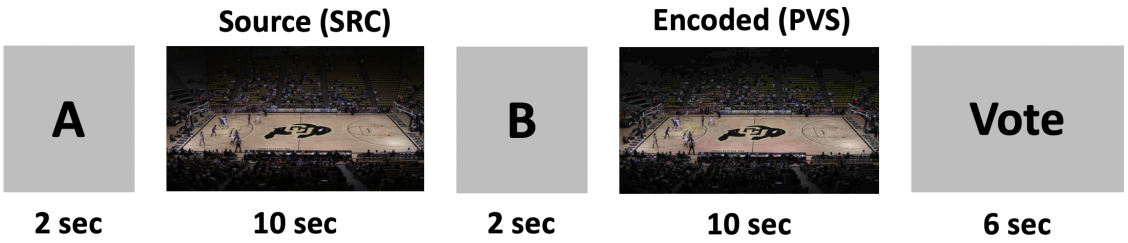



Figure 4.3: Subjective testing procedure: first, the subject was asked to watch the source video, then, after two seconds, the PVS. Finally, he was given six seconds to provide his/her rating.

by computing the perceptual differences between the original reference video and the degraded test video. The DSIS was therefore used with the aim of aligning the subjective evaluation as closely as possible to how full reference VQMs assess the quality. This was to mitigate against any extraneous sources of inaccuracies not directly related to the VQMs.

The source video was shown first, followed by the encoded one (PVS) as illustrated in Figure 4.3. After watching the source video, the PVS was shown two seconds later. The subjects were then given six seconds to rate their perception and the annoyance of artifacts within the PVS against the source video using the DSIS. To aid in the computation of the MOS values, the five options on the DSIS were assigned unique numeric scores (ratings) from 1 to 5 respectively. For each subject, the viewing distance to the monitor was fixed in accordance with the relevant ITU recommendations [92].

Figure 4.4 shows a histogram of the MOS values obtained from the subjective

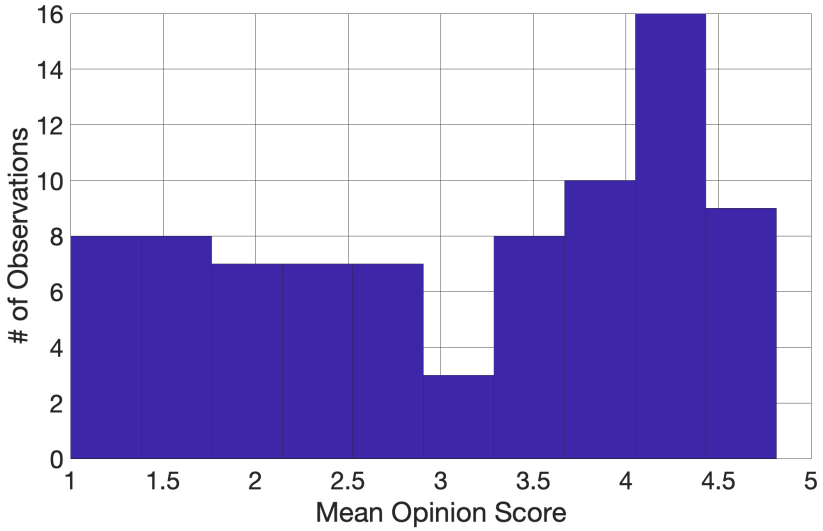



Figure 4.4: The distribution of the MOS values on the quality scale.

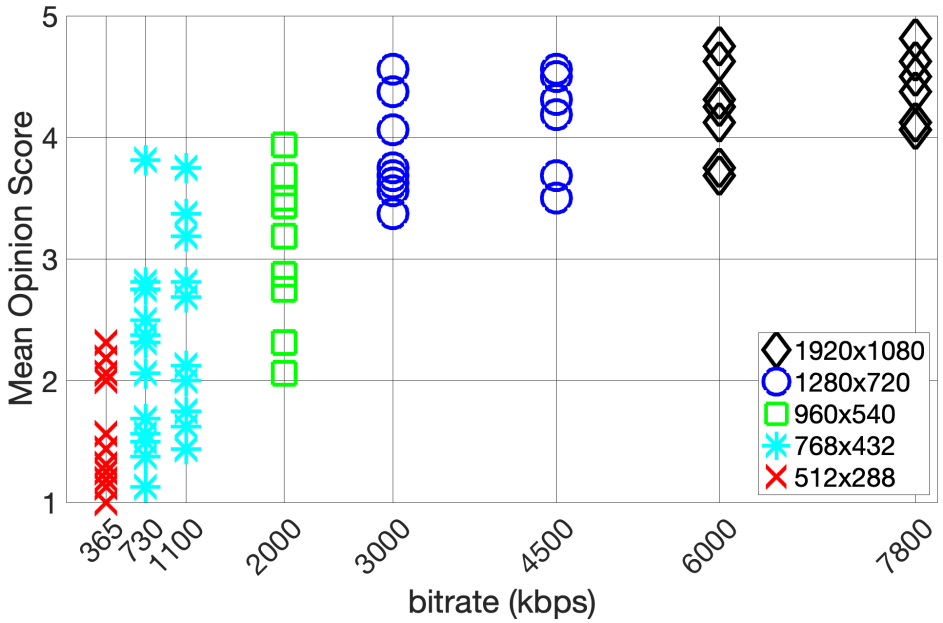



Figure 4.5: Each point corresponds to a PVS. The PVS's bit rate is shown on the x axis, while the y axis shows the PVS' mean opinion score. The color is used to highlight different resolutions. As expected, larger MOS scores were observed on PVSs with higher bit rate (kbps) and resolution.

test. The histogram shows the MOS scores span across the quality scale, and the numbers in the different bins are reasonably well balanced. This is a fundamental requirement to ensure that the conclusions of the analysis based on this dataset are valid on the whole quality scale.

Figure 4.5 presents the MOS values as a function of the bit rate and the resolution. It is evident that subjects were consistent in discerning between low and high video qualities. For example, the video quality at 512×288@365kbps and 768×432@730kbps were rated lower than those encoded at 1280×720@3000kbps. For 1280×720 and 1920×1080 resolutions, an increment in bit rate from 3000 kbps to 4500 kbps and from 6000 kbps to 7800 kbps respectively did not result in noticeable difference in perceived quality.

## 4.6 Results and Discussion

In this section, the effectiveness of the proposed index as a measure of the reliability of VQMs is shown. its robustness to the choice of the set of VQMs used to compute it is discussed. Finally, it is shown that the PVS bitstream features can be used to effectively predict it.

### 4.6.1 MOS Prediction Accuracy vs VQMs Disagreement

This section outlines, the effectiveness of the $D_{pvs}$ index as an indicator of VQM accuracy.

As typically done in the literature, the root mean square errorr (RMSE) between the MOS and the VQMs scores was considered as an indicator of VQMs accuracy.

Figure 4.6 shows the RMSE values for two groups of PVSs. The first group of PVSs is where the disagreement between VQMs, measured by the $D_{pvs}$ index, is low (Low D, $D_{pvs} < 0.2$); the second group of PVSs is where the disagreement between VQMs is high (High D, $D_{pvs} > 0.6$).

In general, in cases of high disagreement (High D), each VQM yielded a prediction affected by a larger RMSE, i.e., larger deviation from the actual MOS value. On the other hand, when the VQMs agree, (i.e., Low D) the average of the observed RMSE values was around 0.4. This is quite interesting since this value is close to the average mutual RMSE that would be observed between MOS values obtained for the same PVSs evaluated in two different subjective experiments [99]. Therefore, this result seems to indicate that, if the proposed $D_{pvs}$ index for a given PVS yields a small value, then the VQMs will provide good approximations of the perceived quality that is obtained in a subjective test for that PVS.

In short, as stated in the section 4.4, the value of the $D_{pvs}$ is able to inform

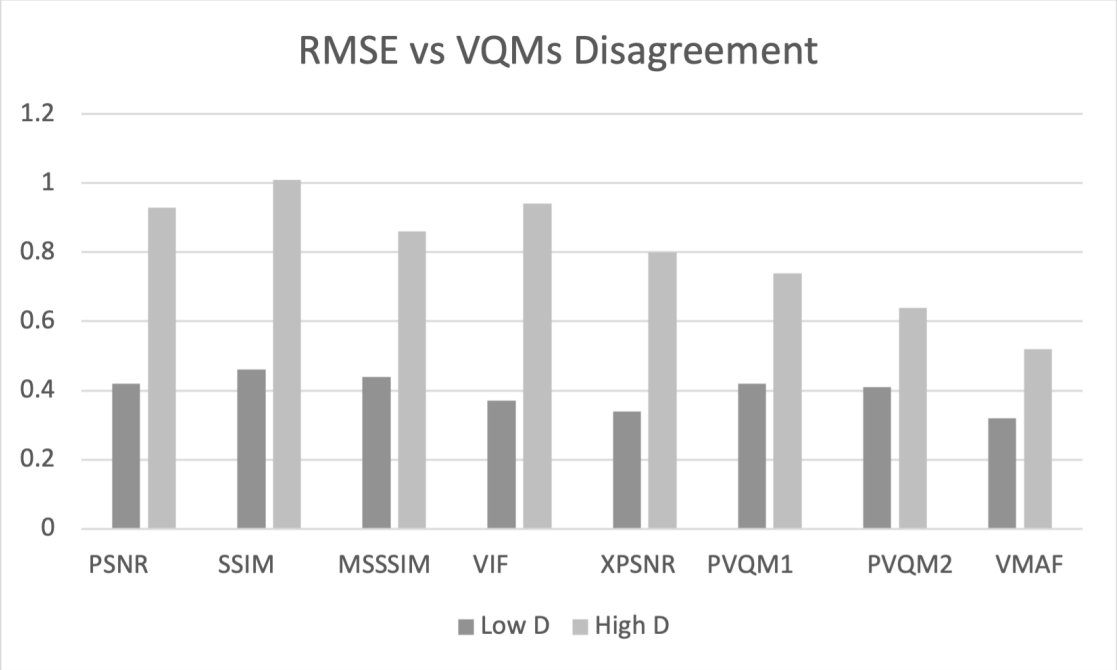



Figure 4.6: Accuracy of the VQMs, in terms of RMSE, as function of the disagreement index. When there is high disagreement, all the VQMs are less accurate.

on the level of accuracy of the VQMs when used to predict the quality of a given PVS. In particular, according to the results in Figure 4.6, a high value of the index ($D_{pvs} > 0.6$) indicates that VQMs predictions are expected to be affected with larger error and vice versa.

### 4.6.2 MOS Prediction Inconsistency vs VQMs Disagreement

The inconsistency of the VQMs in predicting the MOS was also studied as a function of the $D_{pvs}$ index. In measuring the VQMs inconsistency, the variance of the predictions errors (on all PVSs) for each VQM was used. The prediction errors are the differences between the quality score predicted by the VQMs and their corresponding MOSs.

The variance was chosen in order to run statistical tests of significance. These tests (F-test) were performed to check whether VQNs are more inconsistent when predicting the MOS in case of large disagreement with statistical significance.

Table 4.2 reports on the variance of each VQM's prediction errors for PVSs with low and high VQMs disagreements, as well as the p-value of the F-test. In all cases the the test's p-value is smaller than 0.05. Hence, with a more than 95% of confidence, it might be claimed that VQMs are more inconsistent in predicting the MOS of PVSs for which the $D_{pvs}$ index assumes a large value.

### 4.6.3 Open-source vs Proprietary VQMs

Table 4.3 shows a comparison of the performance drop of the different VQMs when used on PVSs whose quality assessment is challenging rather than on those that are easy to evaluate. The results in Figure 4.6 and Table 4.2, show that the challenging PVSs are those corresponding to a high value of the $D_{pvs}$ index, and

Table 4.2: The variance of the VQMs' prediction error is larger, with statistical significance, in case of high VQMs disagreement.

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

Table 4.3: Analyzing the performance drop of the VQMs when used on challenging PVSs. The performance drop ($\Delta$) for each statistical indicator was determined by performing the difference between the value observed when the VQMs are likely to be very accurate, i.e., in case of low VQM disagreement (Low D), and the one observed when there is high VQMs disagreement.

| **Metric** | **$\Delta$ PLCC** | **$\Delta$ RMSE** | **$\Delta$ Var** |
|---|---|---|---|
| PSNR | -0.44 | +0.51 | +0.91 |
| SSIM | -0.61 | +0.55 | +0.84 |
| MSSSIM | -0.35 | +0.42 | +0.6 |
| VIF | -0.48 | +0.57 | +0.69 |
| XPSNR | -0.27 | +0.46 | +0.52 |
| PVQM1 | **-0.18** | **+0.32** | **+0.38** |
| PVQM2 | **-0.10** | **+0.23** | **+0.23** |
| VMAF | **-0.07** | **+0.20** | **+0.20** |

vice versa. Therefore, for each statistical indicator in Table 4.3, the drop $\Delta$ was calculated by taking the difference between the values obtained respectively on the PVSs with high VQMs disagreement ($D_{pvs} > 0.6$) and those with low disagreement ($D_{pvs} < 0.2$). It is very interesting to note that excluding VMAF, all open-source VQMs had a higher accuracy drop than the proprietary ones when moving from less to more challenging PVSs.

Specifically, PVQM1, which is a proprietary metric, had the greatest drop in accuracy, it showed a +0.32 RMSE increase and a -0.18 MOS correlation decrease. On the other hand, the lowest performance drop observed among open-source metrics (excluding VMAF) was +0.42 and -0.27 for RMSE and PLCC respectively. Similar considerations can be made for the variance of the MOS prediction errors $\Delta Var$. These results showed that VMAF and the PVQMs are more robust to the intrinsic ability of a PVS to confuse or mislead VQMs.

It is worth noting that VMAF has a different history and circumstance to the other open-source VQMs considered in this study. Open-source VQMs, in general, mainly originate from academia where access to resources is often constrained in terms of funding and the availability of large libraries of test PVSs. However, VMAF was the result of extensive R&D efforts aimed at optimizing the delivery pipeline of a major media company - Netflix.

The results in Table 4.3 therefore highlight a gap of accuracy between widely used open-source VQMs within the research community and three metrics used to measure the perceptual quality in media industry.

### 4.6.4 Effective Selection of PVSs in Subjective Experiments

This section presents some results aiming at showcasing the usefulness of the $D_{pvs}$ index when selecting the PVSs to be used in a subjective test in order to get the most from it.

The lack of accuracy observed in cases where VQMs disagreed was not actually caused by subject inconsistency. It can be seen, for instance, that the proposed VQM disagreement index is poorly correlated to the subject opinions' standard deviation (SOS) as shown in Figure 4.7. This means subjects did not experience any less or any more difficulty in rating the perceptual quality for cases of high VQM disagreements.

To further investigate how difficult it is to humans to evaluate PVS, labeled as challenging for VQMs on the basis of the $D_{pvs}$ index, the Netflix's SUREAL software that implements the model proposed in [68] for subjective quality recovering was applied. Such a model was chosen because there has been some evidence of its superiority over traditional approaches such as BT.500 [49] and Z-score normalisation [113]. See [68] for more details. The model recovers the so called "true subjective quality" for each PVS while automatically estimating and removing subjects' biases and inconsistencies.

Figure 4.8 shows comparisons between the MOS obtained from the subjective test and the recovered quality (the "true subjective quality") values by the SUREAL software. As seen in Figure 4.8, there was a very good agreement between the two sets of values. This suggests that there were no PVSs whose evaluation had been particularly problematic to the subjects.

The inconsistency that affected the ratings of each individual subject who participated in the test, as computed by the Netflix's SUREAL software is shown in Figure 4.9. The analysis was dome separately for PVSs with low and high disagreement of the VQMs. It can be seen in Figure 4.9 that each subject's inconsistency did not seem to be consistently larger in cases of high VQM disagreements.

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

### 4.6.5 Robustness of the VQMs Disagreement index

The dependencies of the VQM disagreement index on the number and the types of VQMs (i.e. open-source or proprietary) was examined. The value of the $D_{pvs}$ index obtained by using in the Eq (4.1) all the eight VQMs involved in this study

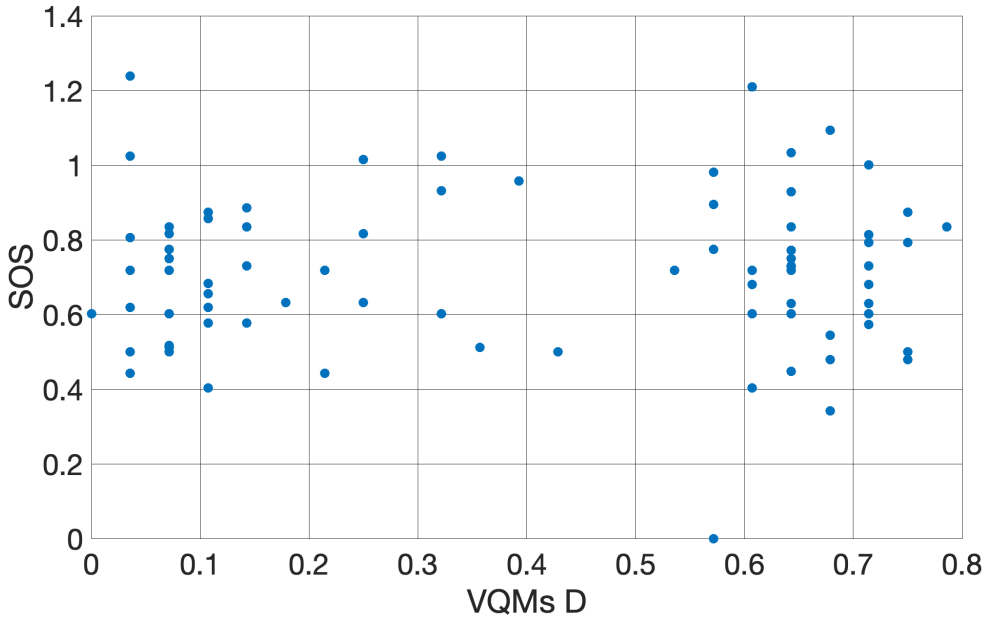



Figure 4.7: The SOS vs the proposed VQMs disagreement index. The subjects' diversity of opinion scores seems to not be correlated with the disagreement index.

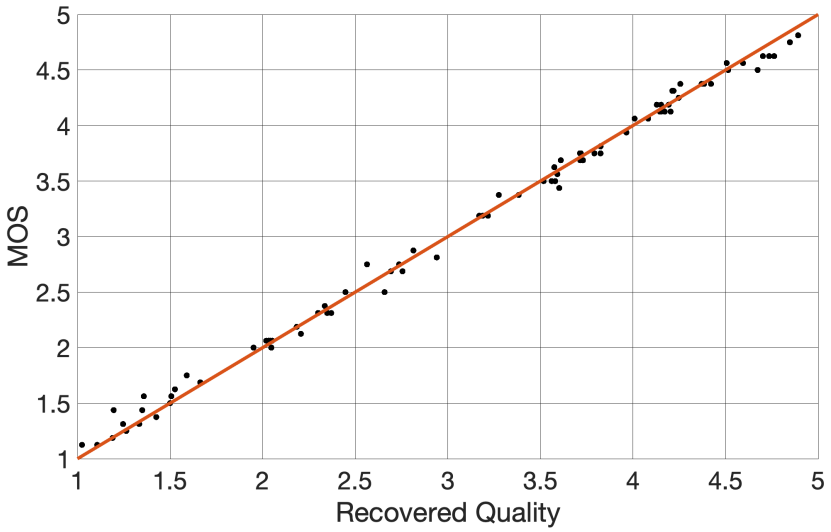



Figure 4.8: The results show that, on average, the subjects consistently evaluated the quality of all the sequences used during the subjective test since the so called "Recovered Quality" of each processed video sequence does not differ significantly from the MOS.

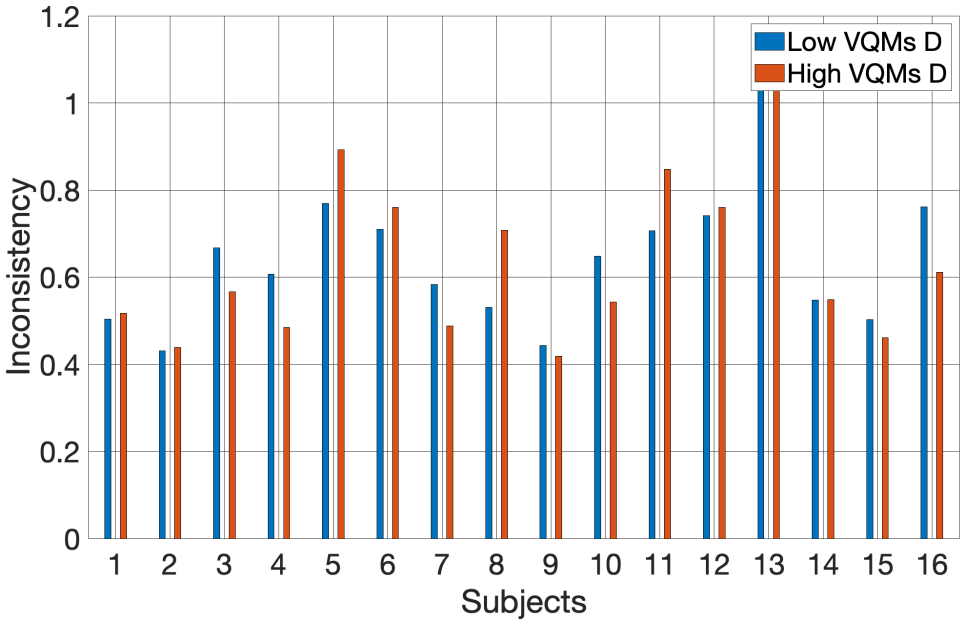



Figure 4.9: Individual subjects' inconsistency as function of the proposed VQMs disagreement index. Subjects seem to experience the same difficulty in assessing the quality of a PVS independently on the disagreement index value.

was considered as the reference value. Then, the disagreement index was computed using only $n$ VQMs (e.g., $n = 5, 6, 7$) chosen from the eight available VQMs, each time considering all possible combinations of the $n$ VQMs out of eight. For example, for $n = 5$, there were 56 distinct combinations. For each combination, the RMSE between the obtained values and the reference values of the index was computed. So, for $n = 5$, 56 values of RMSE were obtained. Note that by considering all possible combinations of VQMs for each value of $n$, this experiment also accounted for the impact of the VQM type used to compute the VQMs disagreement index.

Figure 4.10 shows the minimum, the average, and the maximum values of RMSE for each value of $n$. When all combinations of five VQMs were considered, the average of the RMSE values was 0.12. For combinations where $n$ was greater than five VQMs, an average RMSE of less than 0.08 was observed. This is less than 10% of the range [0, 1], which represents the range of variation of the VQMs disagreement index. This average RMSE value can therefore be considered very reasonable. For the minimum and maximum RMSE values, one can note that the difference between them did not exceed 0.07 for any combination of $n$ VQMs. This difference of 0.07 represents 7% of the variation range of the disagreement index. So, using any combination of VQMs to estimate the reference disagreement value would not vary the average estimation error by more than 7% of the variation range of the VQMs disagreement index.

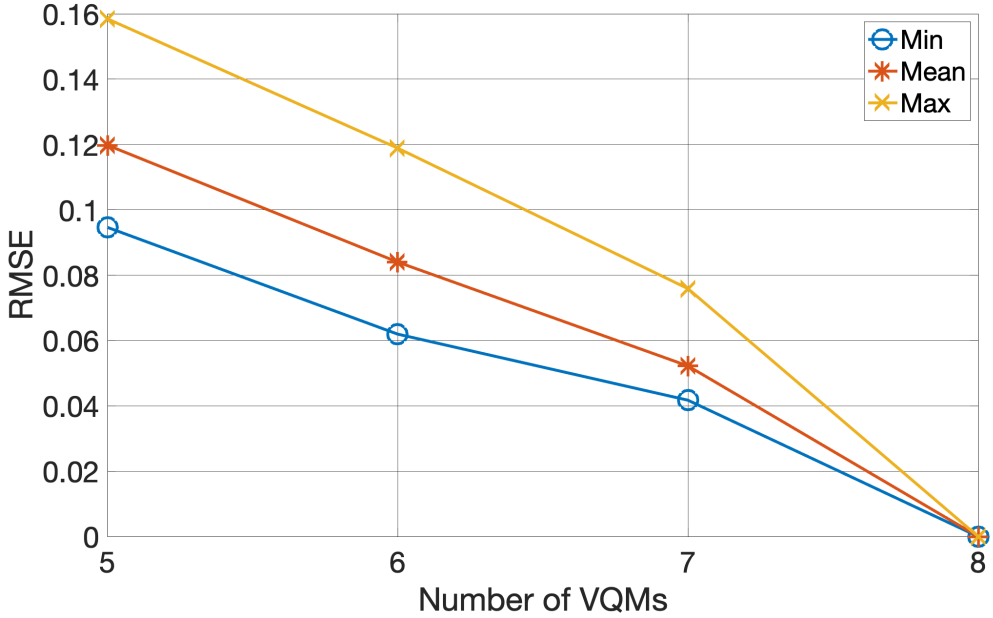



Figure 4.10: Study of the effect of the number of VQMs on the introduced disagreement index. The disagreement index obtained by using all the eight metrics considered in this chapter is looked at as the reference or ground truth. The Figure shows the RMSE between the reference value and the one obtained by using n (n=5, 6 and 7) metrics. For each n, all possible combinations of n metrics out of 8 were used to perform the disagreement index. The minimum, the mean and the maximum RMSE values observed for each n is reported on the Figure.

The results obtained for the RMSE show that the $D_{pvs}$ index is not very sensitive to the number and type of VQMs used to compute it.

To further study the impact of the VQM type on the proposed VQMs disagreement index, it was computed using only open-source VQMs and then checked whether it still remains a good indicator of the accuracy of the VQMs. The results are shown in Figure 4.11. As one can notice, the results were very consistent with those shown in Figure 4.6 where the disagreement was obtained considering all eight VQMs. In other words, when there was high disagreement from the open-source VQMs, a lower accuracy was observed in predicting the MOS. This result is quite interesting because even if the two PVQMs were not considered, the obtained VQMs disagreement index still provided significant indications on the accuracy of all VQMs. This suggests that the $D_{pvs}$ index could be used to deduce the accuracy of any VQM in the literature that had the same design scope as those considered in this chapter.

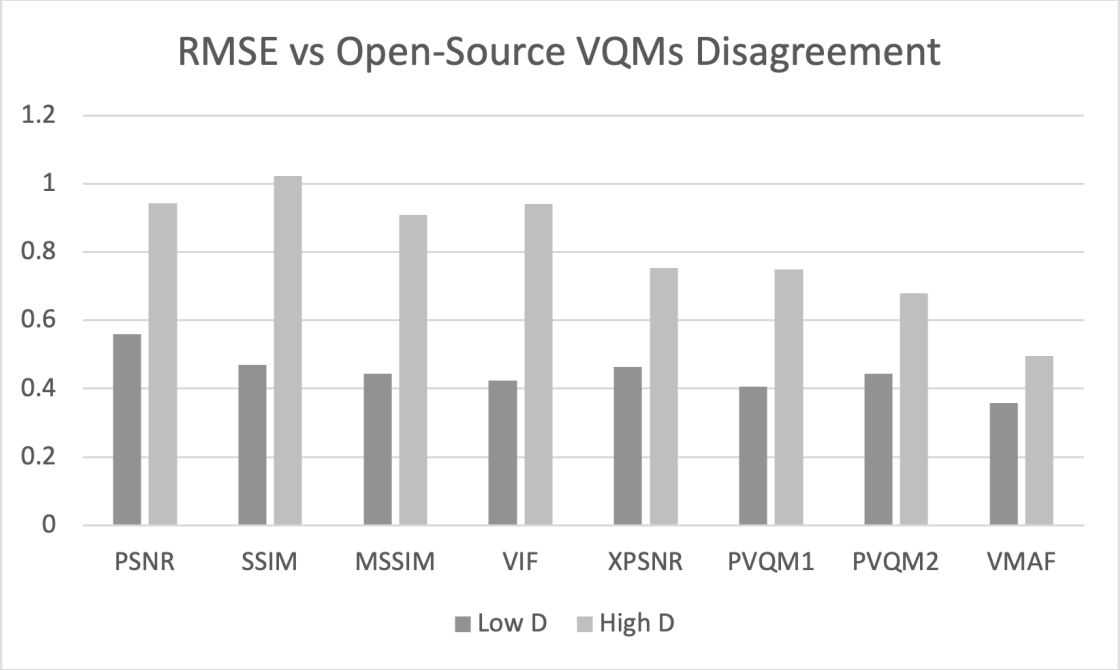



Figure 4.11: The VQMs' accuracy, in terms of RMSE, for low and high values of the disagreement index computed only with the open-source VQMs. For all the metrics, when there is high disagreement of open-source VQMs, the predicted quality score is likely to be affected by larger error.

### 4.6.6 Towards Modeling the VQMs Disagreement with Bitstream Features

The scope of the analysis conducted in this section was to identify the PVS bitstream features that contribute to determine the value of the $D_{pvs}$ index and show to which extend such features can be useful to predict it without having to compute the scores of many VQMs.

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

A backward sequential feature selection algorithm [1] was then used to find the bitstream features that were important in predicting the VQM disagreement index. The features that were seen to have major importance were the average QP, the

Table 4.4: PLCC scores observed between the predicted disagreement index and the actual one using several different machine learning-based regression methods.

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

Table 4.5: SROCC scores observed between the predicted disagreement index and the actual one using several different machine learning-based regression methods.

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

average MV in each direction X and Y, the percentage of Intra blocks in a slice and the percentage of 2Nx2N Intra coded blocks. Furthermore, the aforementioned features were pooled to reach a single value for each PVS by using the Minkowski norm with the exponent set to $p = 1.3$. In fact, by using the arithmetic average as pooling strategy, lower correlation scores between the values of the predicted disagreement index and the actual one were observed as compared to those shown in Table 4.4 and Table 4.5. Therefore, the Minkowski norm was preferred to the average as pooling strategy.

After determining the best set of features, they were regressed to the VQMs

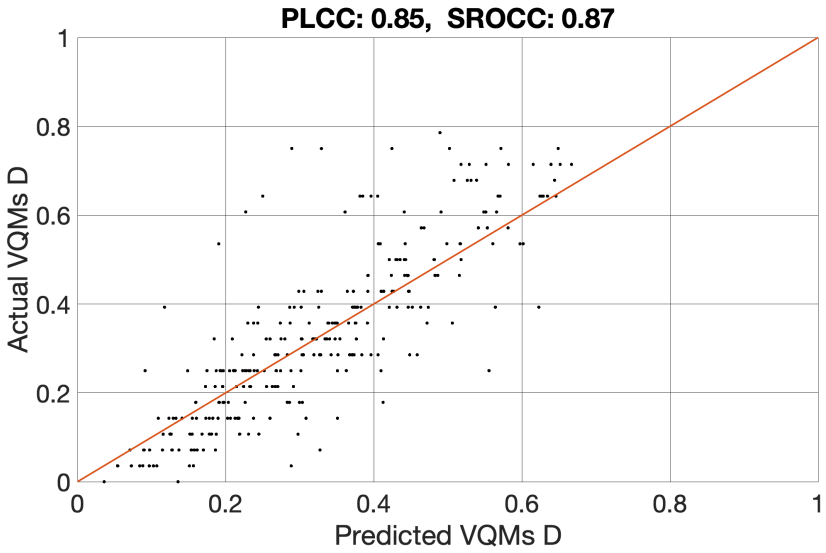



Figure 4.12: The final SVR model's accuracy on all the dataset. Despite the presence of some outliers, the model has been in general able to satisfactory predict the VQMs disagreement index, yielding high PLCC (0.85) and SROCC (0.87) scores.

disagreement index using different machine learning (ML) algorithms. A few regression methods such as linear regression model (LM), regression tree (RT), neural network (NN) with a single hidden layer having four neurons, support vector regression model with a Gaussian kernel (SVR Gaus) and support vector regression model with a radial basis function kernel (SVR rbf) were considered.

The 368 PVSs were divided into 10 folds, and all the models were trained on 9 folds and tested on the one left out. The results are shown in Table 4.4 and Table 4.5. The overall performance was determined by computing the inverse transform of the average Fisher's Z transformation of single correlation scores as recommended in [13].

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

The final SVR model (with an rbf kernel) was trained using all the data available in the dataset. The scatter plot in Figure 4.12 illustrates the performance of the final SVR model on the whole dataset. In general, its predictions correlated quite well with the actual value of the VQMs disagreement index.

The proposed VQMs disagreement index in Eq (4.1) was related to the VQMs accuracy through the results in Figure 4.6 and Table 4.2. The final SVR model (with

an rbf kernel) was also able to accurately predict the proposed VQMs disagreement index using only the PVS bitstream features as shown in Figure 4.12. This suggests that it is possible to determine the accuracy of any VQM on a given PVS, by just relying on its bitstream features, without the need to compute many full reference VQMs (particularly the proprietary ones). In other words, there is a link between the ways a PVS is encoded and the difficulty in accurately evaluating its quality with VQMs.

## 4.7 Conclusion

In this chapter, an index to quantify the VQMs disagreement was proposed. A dataset comprising 368 PVSs was created for the analysis. A subset of those PVSs was selected for subjective evaluation based on the proposed VQMs disagreement index.

Unlike many studies in the literature that analyzed only open-source video quality metrics, the analysis presented in this chapter considered two proprietary metrics used in the content delivery chain by some media industries to optimize their content preparation and delivery pipeline. A comparison analysis between some well-known and widely used open-source VQMs and the proprietary metrics was conducted on the basis of the introduced VQMs disagreement index. The results showed that metrics used by the media companies, i.e., VMAF and the two proprietary VQMs, are more robust to the uncertainty caused by the intrinsic complexity of a PVS.

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

The small-scale subjective experiment that was carried out in the context of this analysis showed promising results in the direction of designing indexes that can measure the reliability of a MOS prediction. Despite this sort of indexes account for the uncertainty that characterize humans' perception of quality, it does not enable an objective quality assessment process that highlights and considers the individual expectations of final users. The next chapter will introduce a more complete approach to objective quality assessment that allows to cope also with this issue.

# Chapter 5

# Mimicking a Single Viewer' Quality Perception with an Artificial Neural Network

## 5.1 Introduction

Different users of the same streaming platform or any multimedia service in general have different expectations in terms of the perceptual quality of the content offered to them. For example, a customer who follows fashion and regularly buys devices at the cutting edge of technology is probably much more demanding than one who only makes use of multimedia tools from time to time. Even the culture and the place where a customer grew up can make his expectations different from those of another one [112, 144, 36].

This diversity in expectations introduces two fundamental questions whose answers are of significant value to any company that markets multimedia content. These questions are as follows: i) what percentage of customers would be satisfied with the quality of a processed video sequence (PVS) if it is encoded in a certain way? What are the characteristics of the customers that would not be satisfied?

Unfortunately, the mean opinion score (MOS) that has been largely studied and predicted does not provide answers to the aforementioned questions. To cope with the first question, some authors proposed approaches to estimate the distribution of the opinion scores (DOS) of the users on the quality scale [54, 124]. While the DOS allows for the estimation of the percentage of users that might not be satisfied, it does not give information about the characteristics of those specific users. Thus the second question, before the development of this PhD thesis, was still suffering a lack of attention in the literature.

This chapter presents a more complete approach to quality assessment that allows not only to predict the MOS, but also represents a preliminary step toward addressing both aforementioned questions at the same time. The approach described

in this chapter derives from the analyses published in the following scientific papers [32, 129].

The idea behind the approach is that of modeling the quality perception of a single subject through a neural network (NN). 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, I proposed to train many NNs, one for each subject, to mimic the behavior of the subjects in terms of quality perception.

This approach of mimicking a subject with a NN is indeed a kind of artificial intelligence. For this reason, from now on, the trained NN for each subject will be referred to as 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 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 discussed in this chapter 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 [112, 144, 36]. In fact, the traits of each observer influence his/her opinion scores that in turn determine the values of the weights of his AIO during the training process.

Once trained, the AIOs enable a more complete approach to objective quality assessment. Given a PVS as input, the AIOs predict the opinion scores the related actual observers would have expressed after evaluating that PVS. These predicted opinion scores can be used to: i) compute the MOS, ii) compute any other statistical indicator of interest, e.g., standard deviation or quantiles; iii) estimate the DOS and thus the percentage of unsatisfied users; iiii) make inference on the class of customers that would not be satisfied by analyzing the characteristic of the actual observers whose AIOs predicted a low quality score.

An important added value of the AIOs-based 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 the same PVS several times. The inability of observers to repeat themselves is an issue that has been investigated and considered in various models [55, 68] to recover the so called "true quality" from subjective data. However the approach in this chapter allows to measure the subjects' inconsistency on any PVS without resorting to a subjective test. Each AIO is designed to output a probability distribution consisting of five values representing the probability of each of the five options on the ACR scale, as shown in Figure 5.1. While for the opinion score prediction, it is enough to 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, I proposed to use this variance as a measure of inconsistency of the observer regarding the quality of the PVS. Numerical results confirmed that such a value follows the typical characteristics of a subject' inconsistency measure.

The remainder of the chapter is organized as follows. In Section 5.2 the related work is briefly presented, followed by Section 5.3 where an in-depth analysis of the strengths of the AIOs-based approach over the traditional approach is performed. Section 5.4 presents the methodology behind the design and training of the AIOs. Numerical experiments and the related results are presented in Section 5.5, followed by Section 5.6 which draws conclusions.

## 5.2 Related Work

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 [22]. Many articles in the literature focused on proposing new approaches to estimate the average quality perceived by the end users, i.e. the MOS [110, 133, 148] or improvements to the existing approaches [12, 127, 62]. 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 [42].

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 DOS is disregarded. However, the skewness plays an important role in estimating the actual QoE of final users: positive values would indicate that the majority of users is actually experiencing a quality greater than the mean (i.e. the MOS), and vice versa. Some papers proposed to overcome the limits of the MOS and SOS, well characterized in [44], by means of additional statistical indicators proposed in the same paper [44], or deriving a range in which the MOS is expected to be with a predefined probability [29].

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 options ("Bad", "Poor", "Fair", "Good", "Excellent") whose mapping to a numerical scale is somehow arbitrary in terms of distance between the options. For this reason, it would be better to work with the DOS so that no arbitrary mapping is introduced.

Until now, few papers focused on the DOS estimation. In [54], the authors proposed a generalized linear model for the DOS estimation, proving its effectiveness on a case study. In [135] and [152], a deep NN is trained to predict a probabilistic

representation of the ratings gathered from actual observers. In [114], the authors highlighted the importance of assessing the quality directly, relying on the subjects' opinions, and thus using the DOS instead of its statistical moments.

The prediction of the DOS however does not allow to make inference on the characteristics of the unsatisfied users, and thus to avoid loosing important customers. It is clear that predicting individual subjects' opinion scores 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 [46, 24, 61]. 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 opinion scores. The approach presented in this chapter suggested for the first time to create, through NNs, models, which are able to mimic individual observers' behavior in terms of quality perception while taking into consideration the uncertainty that characterizes their choices, thus yielding a more complete approach to media quality assessment.

## 5.3 Comparative Analysis of the AIOs-based Approach

Let's suppose Bob is invited to participate in a subjective test. At the end of the test, according to the AIOs-based approach, the ratings collected from Bob are not to be pooled with those of other raters to obtain a MOS. Instead, a NN, i.e, Bob's AIO, has to be trained using his provided scores as ground truth data. The same should be done for the other participants in the test in order to yield many different AIOs modeling human subjects with different characteristics.

The Figure 5.1 shows the generic structure of Bob's AIO. The idea is that of approximating the Bob choices in terms of perceptual quality with a mathematical function represented by a NN. Such a NN, once trained, can receive as input a number of features computed on a PVS and predicts the probability that Bob would have chosen any of the five possible options on the ACR scale if he were asked to score the perceptual quality of that PVS. Further details on the feature set and the NN's architecture will be given in Section 5.4.

The AIOs-based approach to quality assessment aims at being a step towards an objective assessment of the perceptual quality that more resembles a subjective test. In a subjective test, a set of viewers with different characteristics rates the perceptual quality of a stimuli. The AIOs based approach to quality assessment aims at designing NNs (AIOs) that can be used as substitutes of actual viewers with different characteristics. The diagram in Figure 5.2 illustrates the similarity between the AIOs-based approach and a typical subjective test, while highlighting some limits of the traditional approach based on MOS prediction. A deep analysis

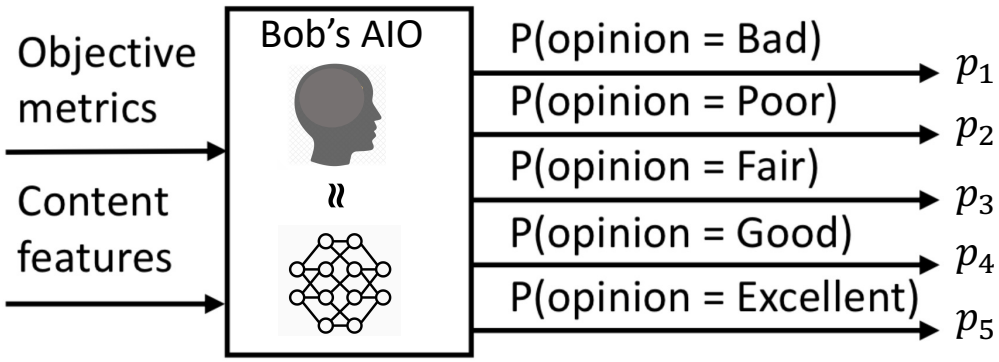



Figure 5.1: The Bob's AIO. A NN is trained to mimic Bob's quality perception. This NN can then predict the Bob's choice probabilities on the ACR scale for a given PVS

of this diagram allows to figure out the main advantages of the AIOs-based approach over the traditional approach. These advantages are summarized from Section 5.3.1 to Section 5.3.4

### 5.3.1 Accounting for Individual Expectations and Inconsistencies

The authors in [55, 68] have shown that the raw opinion scores of any individual observer hide two main characteristics, i,e., the subject's bias and inconsistency. The bias is up to a certain extent an indicator of the subject's expectation. For instance, a subject that has high expectation in terms of perceptual quality tends to give lower quality scores than those of other subjects, this results in a negative bias. On the other hand, subjects with positive bias might also be those that are less demanding. The subject's inconsistency instead translates his/her ability to remain coherent when evaluating content with the same visual quality. Both the bias and inconsistency observable from individual scores are lost when doing operations such as the mean to get the MOS. In fact, the average leads to less noisy data at the expense of the information on individual characteristics of the subjects. This is the reason why in Figure 5.2, it was highlighted that, unlike the subjective tests and the AIOs-based approach, traditional approaches based on MOS prediction do not

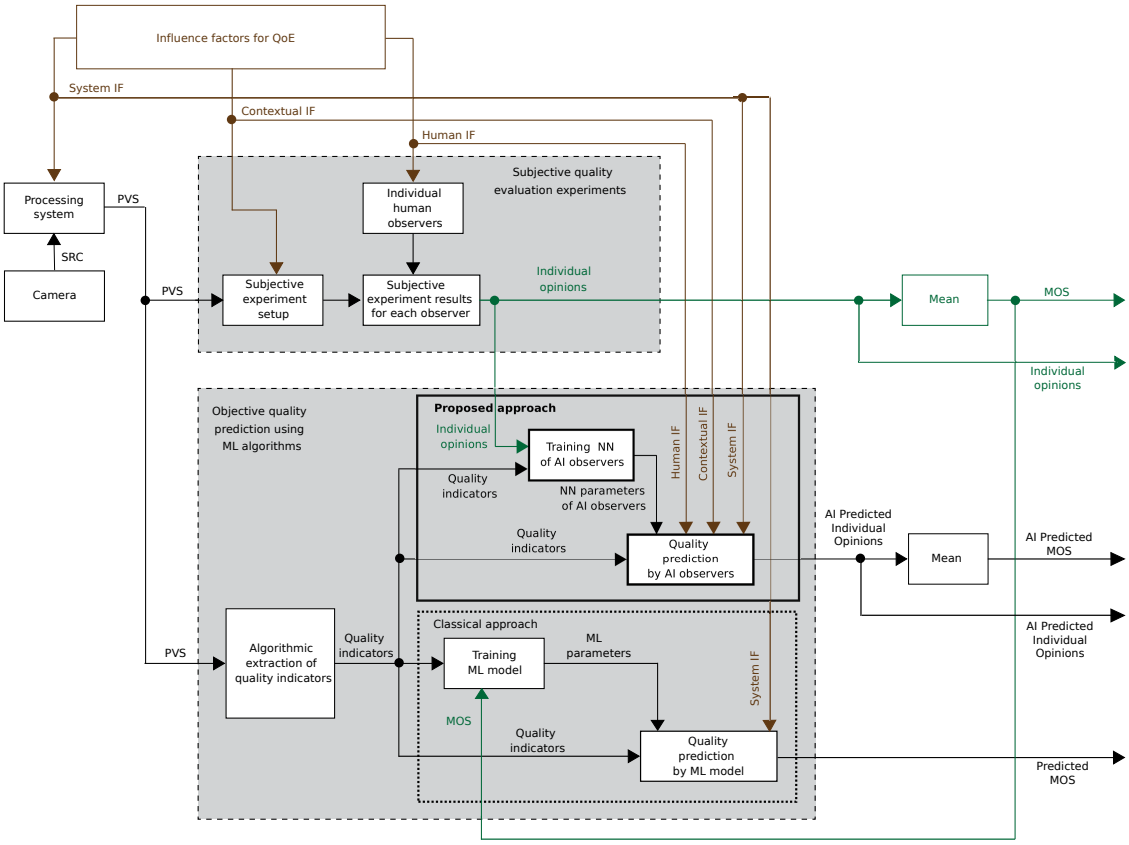



Figure 5.2: Illustration of the proposed approach in comparison with the traditional approaches to media quality assessment. In particular, note that similarly to subjective experiments, the proposed approach considers also human factors and provides individual opinions yielding, in practice, more flexibility.

account for human influence factors (IFs).

The AIO of each subject is trained on his/her raw opinion scores prior to the application of any pooling operator, e.g., the average to get the MOS. The weights of the AIOs, during the training process, capture the subjects characteristics when trying to replicate his/her opinion scores. Given a PVS whose quality is to be evaluated, the trained AIOs provide automatically generated ratings (AGRs) as real observers would do during a subjective experiment. The diversity observed between these AGRs highlights the impact of individual expectations in the QoE measurement. On the other hand, as it will shown later, the probabilistic output of each AIO yields a good measure of inconsistency.

The subjects-specific factors have already been intensively studied. Several papers [112, 144, 158] 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 thus expectations. This is an additional reason to model each single observer relying on data gathered during subjective experiments.

From a practical point of view, when a subject is invited to a subjective test, nothing prevents the test designer from collecting any of his/her characteristics that may in some way impact his/her judgment of perceptual quality. In this way, the characteristics of the subjects whose AIOs predict a critical opinion score for a given PVS would consent to make inference on the profile of customers who would not be satisfied with the visual quality of that PVS.

### 5.3.2 Generality of the AIOs-based Approach

The AIOs-based approach to quality assessment is much more general in the sense that it allows an estimation of almost all the statistical indicators used so far in the media quality assessment literature. In fact, from the AGRs, an estimate of the DOS can be derived. Almost all the QoE statistical indicators proposed in the literature can then be rather easily derived from the DOS [114]. For instance, the MOS and the SOS can, respectively, be derived from the first and the second order statistical moments of the DOS after mapping the AGRs to a numerical scale.

Furthermore, a complete estimation of the DOS 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 [114]. This is because, using the DOS, it is possible to avoid the usage of the classical statistical tests designed for normally distributed 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 [69].

### 5.3.3 The Issues with the MOS Definition

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 after a subjective test, each of the options of the quality scale is typically 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 change from ”Good” to ”Excellent”.

In [85, 122], 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 [55, 68].

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 [91, 108, 69], hence it would be misleading to simply ignore such an issue.

In the light of the previous argument, it is important to design new media quality assessment approaches that rely directly on the ordinal data collected during subjective experiments as proposed by the AIOs-based approach. In fact, 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 different observers. Given a PVS, they simply attempt to predict the option that the related observer would have chosen.

### 5.3.4 Simulation of Subjective Experiments

Subjective experiments are crucial for the development and/or validation of machine learning-based objective measures for media quality assessment. However, the performance of machine learning-based 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 observers with different characteristics 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.

## 5.4 Implementation of the AIOs-based Approach

This section describes the main steps towards the derivation of the AIO mimicking an actual subject in terms of quality perception.

### 5.4.1 Dealing with the Data's Noisy Nature

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, raw opinion scores of individual observers are much more noisy than the MOS. Indeed, several models have been proposed to recover the actual subjective quality from raw opinion scores [55, 68] while getting rid of noise caused by the subjects' inconsistency. This observation justifies why training NNs that can effectively mimic individual subjects results in a tricky learning task.

Different approaches to cope with noisy labels have been proposed in the machine learning community [34, 90]. It has been shown that more training samples as well as more complex learning based models are required when dealing with noisy labels [34]. In fact, overfitting a clean dataset might yield a model that can still perform well up to a certain extent on the test sets. Instead when learning on noisy data, one should absolutely avoid an overfitting of the training set, since the trained model would memorize patterns coming from the noise and will not be useful on a different dataset. For this reason an effective data augmentation strategy could really help to get models able to learn only useful features when dealing with noisy labels.

To train the AIOs, the data gathered during the VQEG-HDTV Phase I test [139] were used. This dataset was chosen to ensure that the trained AIOs could be used for a wide range of applications (A detailed description of this dataset was provided in Chapter 2).

In particular, as mentioned in Chapter 2, the VQEG HDTV Phase I test was done in six different Labs. In each Lab, 24 viewers participated in the test and rated 168 PVSs. Each of the six Labs used a different set of stimuli. However, there are 24 PVSs that had been rated by all the $24*6$ participants. In other words, these 24 PVSs were used in all the 6 Labs. From now on, these 24 PVSs will be referred to as the "common set". For the analysis, only the data collected in the Lab 1, 3 and 5 were considered since the tests in the other Labs involved interlaced content that is out of the design scope of some of the objective video quality measures (VQMs) that were used as part of the features to train the AIOs.

The 168 raw opinion scores collected from each subject during the VQEG HDTV Phase I test were seen not to be enough to effectively train and validate the AIO of that subject. In fact, when splitting these 168 observations in training and testing sets, independently from the used NN architecture, the trained model was not able to perform well on both the training and the test set at the same time. This could have been expected, since training an effective model for the prediction of a quantity such as the MOS, that is less noisy than individual opinion scores, is already challenging when working with a hundred of training samples. A data augmentation strategy to cope with this issue was therefore designed.

The analysis in this chapter aimed at training 24 NNs, each mimicking a single viewer of Lab 1. The main idea behind the proposed data augmentation approach was that of finding an estimate of the opinion scores that each viewer of Lab 1

would have expressed on the set of PVSs used in Lab 3 and 5. This estimated opinion scores allowed to augment the data available for training the AIOs of each of the 24 viewers of Lab 1.

The estimated opinion scores of each viewer of Lab 1 for the PVSs used in Lab 3 and 5 was obtained by finding the viewers of Lab 3 and Lab 5 that voted similarly to him/her on the "common set". To this aim, as a measure of dissimilarity on the "common set", the "mutual" root mean square error between the opinion scores of two viewers was used.

Formally, let denote by:

- $i$ and $j$ two generic viewers taken from two different Labs;
- $\mathcal{C}_{set}$ the "common set" of PVSs;
- $V_{pvs}^{i}$ the opinion score of the subject $i$ on a generic PVS in the "common set".

The dissimilarity between the subject $i$ of Lab 1 and the subject $j$ of Lab 3 or 5 in terms of quality perception was expressed as it follows:

$$d_{ij} = \sqrt{\frac{1}{|\mathcal{C}_{set}|}\left(\sum_{pvs \in \mathcal{C}_{set}} (V_{pvs}^{i} - V_{pvs}^{j})^2\right)}. \tag{5.1}$$

Each of the 24 viewers in Lab 1 obviously rated only 168 PVSs. For each viewer $i$ of Lab 1, one therefore needs to find the viewers $j$ and $k$ respectively from Lab 3 and Lab 5 whose ratings can reasonably approximate those that the viewer $i$ would have expressed. At the same tine, the viewers $j$ and $k$ should also be similar in terms of quality perception. For each viewer $i$ of Lab 1, the viewers $j$ and $k$ were chosen such that the total "mutual" root mean square error between the ratings of the triplet of subjects $i$, $j$ and $k$ on the "common set" is minimized.

More formally, for each viewer $i$ in Lab 1, let denote by $x_i^{j,k}$ a binary decision variable equal 1 if and only if the ratings of the viewer $i$, on the PVSs used in Lab 3 and 5, will be approximated by the ratings expressed by the viewers $j$ and $k$ respectively from Lab 3 and 5. The values of the decision variables $x_i^{j,k}$ were found by solving the following optimization problem:

$$\begin{aligned} &\min_{x} \sum_{i \in Lab1} \sum_{j \in Lab3} \sum_{k \in Lab5} (d_{ij} + d_{ik} + d_{jk}) x_i^{j,k} \\ &s.t. \sum_{j \in Lab3} \sum_{k \in Lab5} x_i^{j,k} = 1 \ \forall i \in Lab1 \\ &x_i^{j,k} \in \{0,1\} \end{aligned} \tag{5.2}$$

Mote that the first constraint simply expresses the fact that the ratings of each single viewers of Lab 1 should be augmented by those of exactly one viewer of Lab

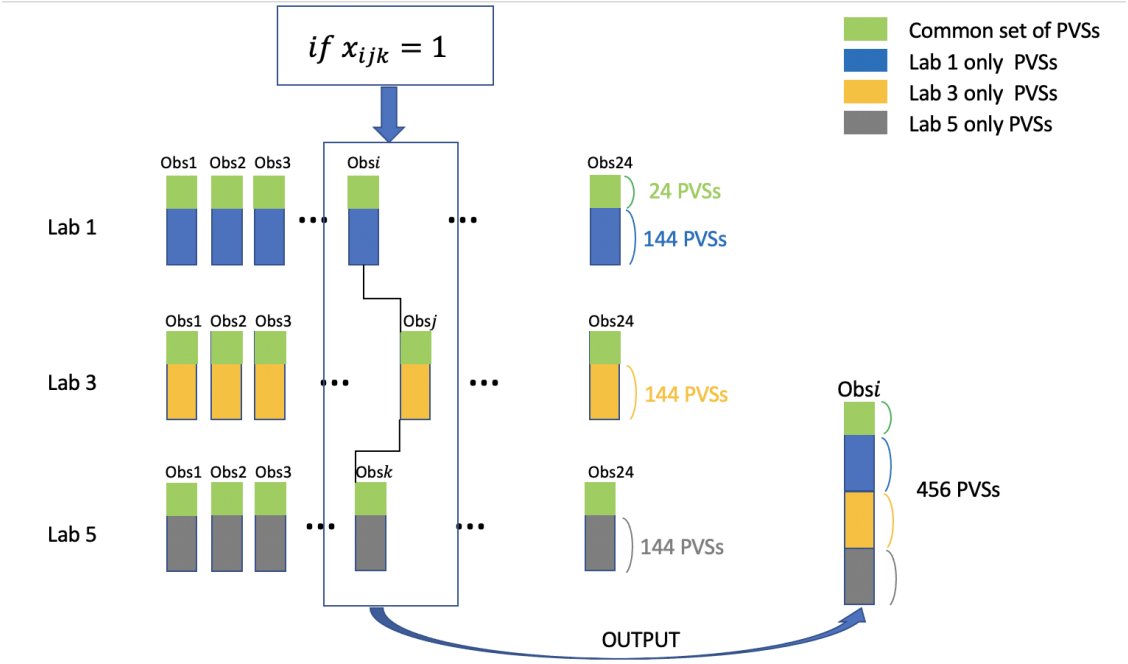



Figure 5.3: Proposed data augmentation approach. Each viewer of the Lab 1 was put together with a viewer of the Lab 3 and a viewer of Lab 5 based on the solution of the optimization problem. This yielded 24 viewers that were considered to have rated 456 PVSs instead of 168.

3 and one of Lab 5. The second constraint instead expresses the binary nature of the decision variables.

Note that the solution of the optimization problem in Eq (5.2) could allow to approximate/augment the ratings of two different subjects of Lab 1 by those of a single subject of Lab 3 and/or Lab 5. Thus there might be some ratings from Lab 3 and Lab 5 that are not used. The primary scope of the approach is to make sure that the approximated scores, for each of the 24 participants of Lab 1, are highly consistent with the scores he/she actually expressed. Putting a constraint that enforces the use of all ratings might yield a sub optimal solution connecting together viewers that do not have similar perception of quality.

Once the problem in (5.2) is solved, exactly 24 among the decision variables $x_i^{j,k}$ are equal to 1. The solution indicates how to augment the ratings of each of the 24 viewers of Lab 1. For instance, if the variable $x_2^{4,7}$ is equal to 1 in the optimal solution, then the ratings of viewer #2 of Lab 1 are augmented by using those of the viewer #4 of Lab 3 and those of the viewer #7 of Lab 5.

The Figure 5.3 summarizes the data augmentation procedure. Based on the optimal solution of the optimization problem in (5.2), the ratings of each viewer of Lab 1 are augmented with those of two viewers, one from Lab 3 and the other from

Lab 5 to form a unique viewer. Therefore the whole data augmentation procedure combines the original data to generate 24 "observers" such that each of them can be considered to have rated 456 $(168 * 3 - 24 * 2)$ PVSs. An AIO was then trained for each of these 24 observers.

It is important to note that inferring the behavior of an observer on 168 stimuli starting from what was observed on the "common set" that contains only 24 PVSs 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 Phase I experiment [139]: the 24 PVSs 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 consent the alignment of the results of all the six Labs.

### 5.4.2 Network Architectures and the Training Process

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 observers, 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 consists of five neurons, each predicting the probability that the AIO chooses one of the five options 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.

In the context of the analysis presented in this chapter, to train the NN mimicking each of the 24 observers, a set of hand-crafted features, here denoted by $\mathcal{F}$, characterizing each PVS was first computed.

The features set $\mathcal{F}$ included the following VQMs: PSNR[147], SSIM[157] MS-SSIM[143], VIF[115] and VMAF[87], 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. These features are described in details in [65]. Finally, the spatial activity index (SI) and

the temporal activity index (TI) as defined in [92] were also computed.

Note that a set of hand-crafted features were used since the 456 ratings available for each observer do not allow to directly train a deep convolutional NNs that can automatically figure out important perceptual features from the PVSs. Small scale NNs were instead used to regress the extracted features on the quality scale.

For each observer, to create the corresponding AIO, the procedure described in the following was implemented. First, for each observer, the subset of features as well as the NN architecture that best model his/her opinion scores were experimentally identified.

In order to identify the optimal set of features and the best architecture for each observer, the following procedure was adopted. 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, three different NNs were trained, 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, the three NNs obtained for each possible subset of the features were tested on a test set by comparing the predicted opinions with the actual ones. For each observer, the NN architecture and the related subset of features that yielded the highest accuracy on the test set were considered 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 were examined for each observer in order to investigate what is the level of complexity required to effectively model the observer.

Summarizing, each of the 24 observer was 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 varied from one to three depending on the complexity of the observer and finally the output layer had five neurons that predict the probability of choosing each one of the five possibile options offered by the ACR scale. Once trained, during the testing phase, the predicted opinion score was obtained by selecting the option with the largest probability.

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, one expects 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 several times the quality of a PVS, what would be his/her expected accuracy in repeating previously expressed opinions? The results of the subjective experiments presented in [55] 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 quality level on the ACR scale 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 always 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.

### 5.4.3 A Measure of Subjects Inconsistency

Once the AIO, trained to mimic an observer $o$, is deployed on a PVS, it outputs the following discrete probability distribution $p_{oi}^{PVS} \quad i = 1,2,\ldots,5$ , where the index $i$ represents the five options offered to the observer on the ACR scale.

Denoting by $v_i \quad i = 1,2,\ldots,5$, the actual numerical score of the five options of the ACR scale, the variance of such a predicted distribution, i.e.

$$\sigma^2(o, PVS) = \sum_{i=1}^{5} v_i^2 \cdot p_{oi}^{PVS} - \left( \sum_{i=1}^{5} v_i \cdot p_{oi}^{PVS} \right)^2 \tag{5.3}$$

was defined 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 opinion scores 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 of the same PVS.

To understand how the measure described in Eq (5.3) captures the observer inconsistency, lets 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 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/her 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 in Eq (5.3). This is unfortunately too expensive to be carried out in practice. For this reason, in the Section 5.5 a different approach was adopted to show Eq (5.3) effectiveness as a measure of inconsistency. More precisely, it was shown that the proposed measure possesses the properties that are expected from an inconsistency measure.

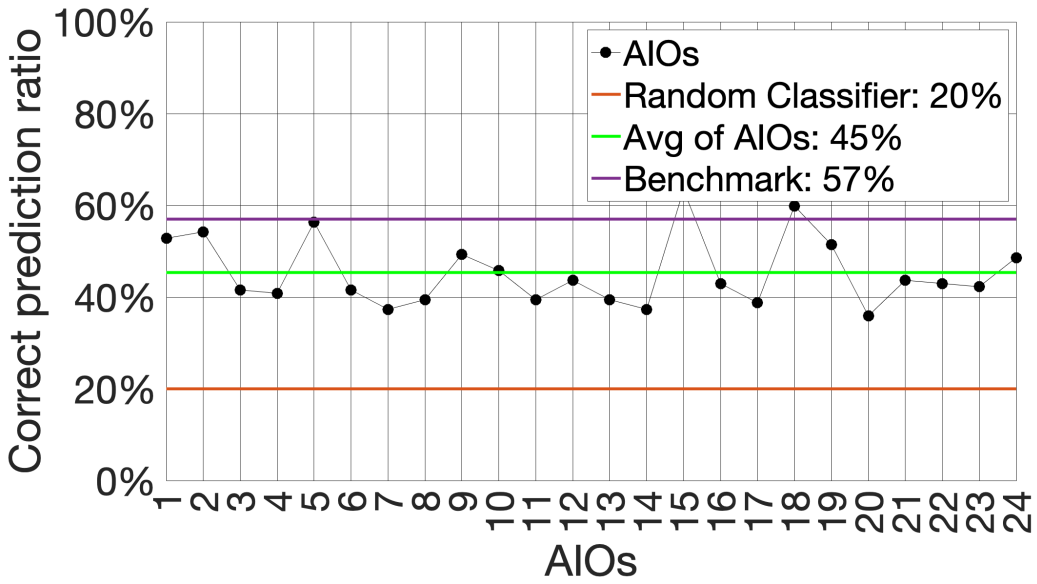
## 5.5 Numerical Experiments

To assess the feasibility as well as the effectiveness of the AIOs-based approach to quality assessment, extensive numerical experiments were conducted. The related results are presented and discussed in this section.

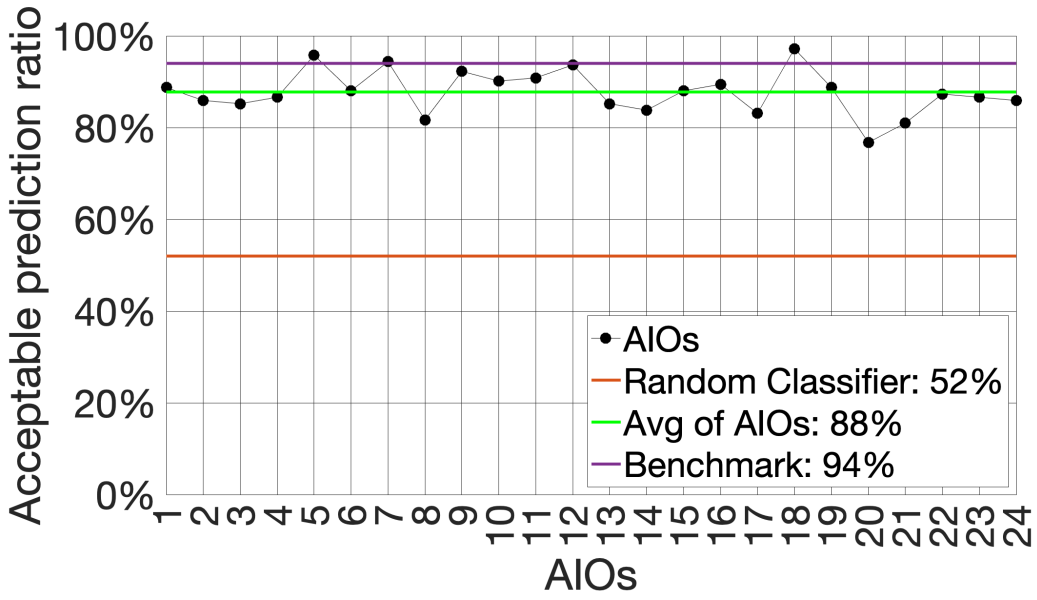
### 5.5.1 The Experimental Setup

In order to compare the AIOs with a random classifier and the MOS, and also to implement the data augmentation approach described in Section 5.4, the widely used mapping of the ACR scale options to integers from 1 to 5 was employed, 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 it is a necessary step when using the AIOs-based approach.

The numerical experiments were done considering four subjectively annotated datasets, i.e. the VQEG-HD1 (Lab 1), VQEG-HD3 (Lab 3), VQEG-HD5 (Lab




(a) Correct Prediction Ratio




(b) Acceptable Prediction Ratio

Figure 5.4: Accuracy of the AIOs. The AIOs were trained on the VQEG-HD1 and VQEG-HD5 datasets, and tested on the VQEG-HD3 dataset. The average performance ratios of the AIOs (green lines) are significantly higher than those of a randomly voting subject (orange lines) and do not differ more than 12% from the benchmark values (violet lines).

5) [139] and the ITS4S [96] dataset. In addition, the large scale JEG-Hybrid dataset [14], 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 5.1. As it can be clearly seen from the table, during the VQEG-HDTV experiments several types of distortions have been applied to the PVSs, while only coding artifacts have been considered in the other datasets. Therefore, most of the experiments were done relying on the VQEG-HDTV datasets in order to investigate the effectiveness of the AIOs-based approach for a wider range of cases.

Table 5.1: Description of the datasets used in the experiments

| Dataset | Size | Distortions | Notes |
|---|---|---|---|
| VQEG-HD datasets | 168 PVSs, 1080p, 10-sec long | MPEG-2 and AVC-encoded, 1 to 15 Mbps, transmission artifacts due to bit errors and bursty packet losses | movies, sports, general TV material with as much variety as possible |
| ITS4S | 514 PVSs, 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 PVSs, 1080p, 10-sec long | HEVC-encoded (0.5 to 16 Mbps + constant QP) | Not subjectively annotated |

To assess the ability of each AIO to mimic the corresponding subject, the following two ratios were considered:

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

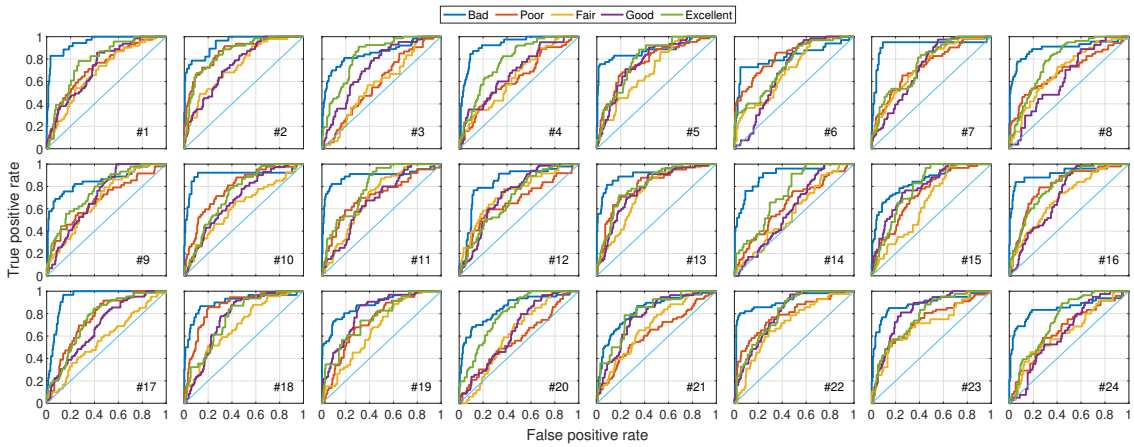
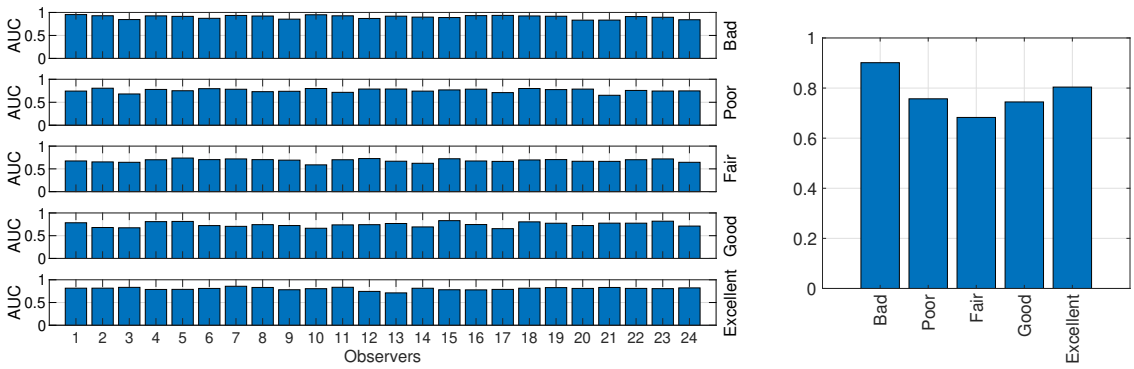



Figure 5.5: The ROC curves associated with the AIOs, which models each observer. In all the cases, the curve is above the 45 degree line: the AIOs is therefore effectively modeling some of the aspects that concur with the way how the observer perceives the visual quality.



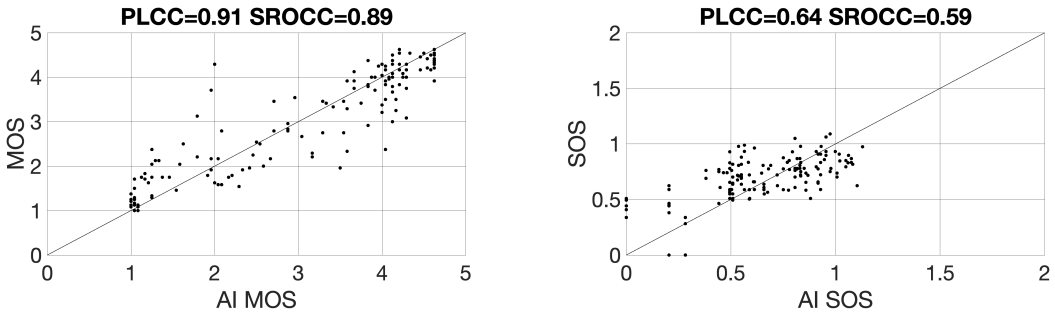

(a) AUC indexes for each observer's model

(b) Average AUC

Figure 5.6: The AUC indexes associated with the AIO, which models each observer. The closer to 1, the better. The AIOs seem to be more accurate when modeling the observer's behavior in the case of the PVSs with the very low or high quality.

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

in which OS stands for opinion score. The correct prediction ratio and thus the accuracy of each AIO achieved on the test set is the number of PVSs for which the rating predicted by the AIO 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 AIO prediction differs no more than 1 level on the ACR scale from the rating of the related observer divided by the total number of PVSs in the test set. For a random classifier (RC),




(a) (b)

Figure 5.7: 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 AIOs' opinion scores.

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 ($2/5 \cdot 1/5 + 3/5 \cdot 3/5 + 2/5 \cdot 1/5$ for the acceptable prediction ratio).

Since each AIO can be looked at as classifier, the correct and acceptable ratios were used together with the receiver operating characteristic (ROC) curves as well as the area under the ROC curve (AUC) indexes associated with the AIO modeling each subject.

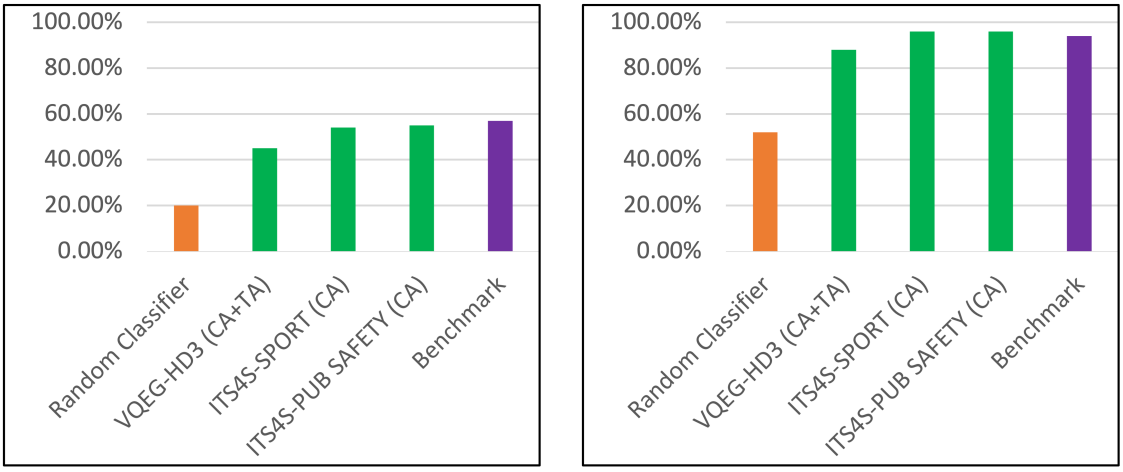
In Table 4 of [55], the overall correct and acceptable ratios obtained when asking a subject to rate again a PVS he/she already rated were respectively 57% and 94%. These values will be used in this section as a benchmark when analyzing the AIOs performance. The point is that, one should not expect the AIOs to perform better than what actual observers would do.

### 5.5.2 The AIOs Accuracy in Mimicking Actual Observers

For the experiments on the VQEG-HDTV datasets, 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.

In Figure 5.4 the accuracy of the 24 AIOs is compared to that of a model randomly rating the PVSs in the test set and the benchmark value. By comparing the AIOs with an observer voting randomly, one aims at verifying whether the NN mimicking each observer did learn interesting information about the way the observer perceives quality. For each of the 24 AIOs both the correct and the acceptable ratio significantly exceeded the expected accuracy of a model voting at random. In particular, an average correct ratio of 45% ($> 20\%$) and an acceptable

prediction ratio of 88% ($>$ 52%) were observed respectively. There were some AIOs, with an accuracy very close to the benchmark value. The average correct and acceptable ratios of the AIOs differed from the respective benchmark values by no more than 12% and 6%.




(a) Correct ratio

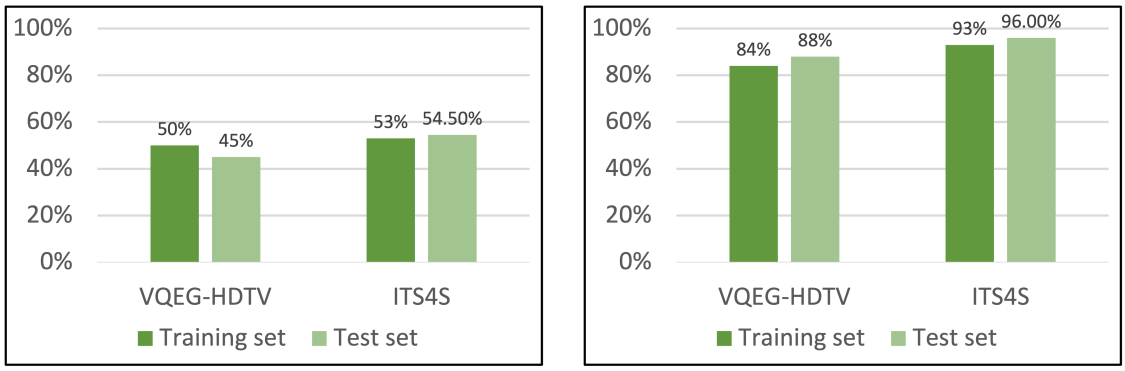
(b) Acceptable ratio

Figure 5.8: Average correct and acceptable ratios of the AIOs (green bars) in comparison to those of a random classifier and the benchmark values. CA and TA stand respectively for coding artifacts and transmission artifacts. The analysis suggests that higher performances might be expected from the AIOs when focusing only on coding artifacts.

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 ratios for such an observer would be 20% and 60%, respectively.

To further investigate the accuracy of the 24 AIOs, the ROC curves as well as the AUC indexes associated with each of the five options of the ACR scale predicted by the AIO were computed. The results are shown in Figure 5.5 and 5.6. For all the 24 AIOs, 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 5.6a and Figure 5.6b reveal that the 24 AIOs are reasonably accurate since, on average, they reported AUC indexes ranging from 0.69 to 0.9.

It is important to notice in Figure 5.6b the ability of AIOs to more accurately model the subjects' 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" options on the ACR scale.




(a) Correct ratio (b) Acceptable ratio

Figure 5.9: Comparison of the average correct and acceptable ratios of the AIOs prediction on the training and the test set.

In Section 5.3, the MOS was presented as a limited indicator when used alone to measure the QoE. Nevertheless, it has been shown to be effective for some purposes, such as codecs comparison. For this reason, the performance of the AIOs when used to simulate a subjective test where the expected outcome is a MOS value for each PVS was also investigated. The 24 AIOs, that were trained using the data from the VQEG-HD1 and the VQEG-HD5 experiments were used to simulate a subjective test on the PVSs used in the VQEG-HD3 experiment. The mean of the AIOs' opinion scores (that is referred to as AI MOS in the following) and the standard deviation (AI SOS) were then compared to the actual MOS and SOS values.

The results shown in Figure 5.7 are quite promising. In fact, 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, since models for SOS prediction are still at their early stage.

The results presented so far show that the NNs-based AIOs can reasonably model actual subjects, even when trained on a dataset such as the VQEG-HDTV that involves a wide range of distortion types. Still, one wonders whether the obtained accuracy could have been higher if the analysis was restricted to a specific type of artifacts. In particular, a closer look was given to the case of coding artifacts that tends to require more attention than transmission ones in this era characterized by a fair availability of effective network security protocols.

To this aim, other AIOs were trained relying on the ITS4S dataset that provides 514 ratings for each of the 27 observers that participated in that experiment. To train the 27 AIOs, 7 categories of PVSs out of the 9 available in the dataset were used and the models were then tested on the remaining two categories, i.e., the sport content category (ITS4S-SPORT) and the public safety content category (ITS4S-PUB SAFETY).

Note that in the case of the ITS4S dataset, no data augmentation technique was used since each subject had evaluated many more PVSs than the 168 PVSs rated by each participant in the VQEG-HDTV experiments.

Figure 5.8 presents the average correct and acceptable prediction ratios of all the 27 AIOs on the two test sets in comparison to the results obtained in the more general case. An average correct ratio of 54.5% (> 45%) and an average acceptable ratio of 96% (> 88%) have been obtained. One can also note that the obtained performance indicators are much more close to the respective benchmark values. Such result suggests that the performance of the AIOs can be further improved when they are designed to deal with specific applications.

### 5.5.3 The AIOs Robustness

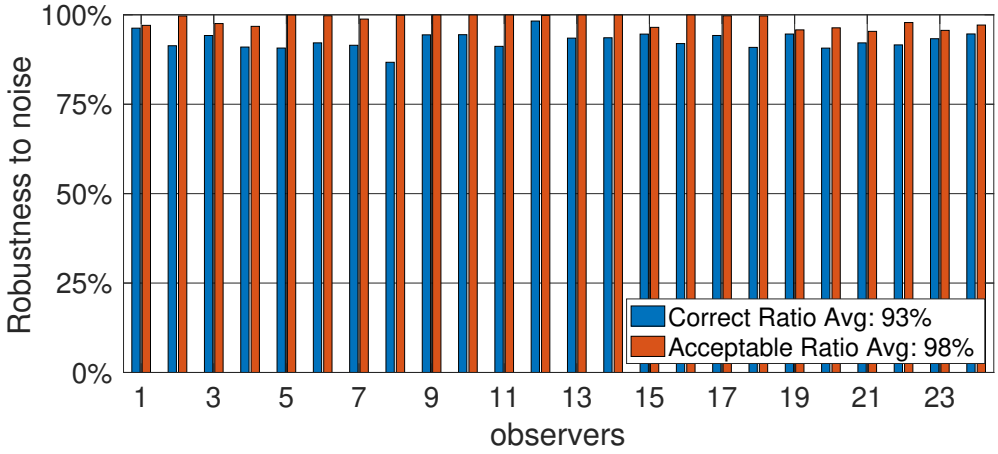



Figure 5.10: Probability, for each AIO, that its output will not change (the correct ratio) or will change by at most 1 quality level on the ACR scale (the acceptable ratio) after adding, to each input feature, a noise term which is uniformly distributed between -1% and 1% of the range of values assumed by such feature in the dataset.

Three main aspects were considered in evaluating the AIOs robustness: 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 on input data; iii) Assessing the ability of the AIOs to distinguish between two input PVSs with significantly different visual quality. For the second and the third aspects, the analysis was done with the 24 AIOs trained on the VQEG-HDTV dataset.

For both the VQEG-HDTV and the ITS4S datasets, the performance of the AIOs on the training set was not significantly different from that observed on the test set. In both cases the average performances observed on the training sets

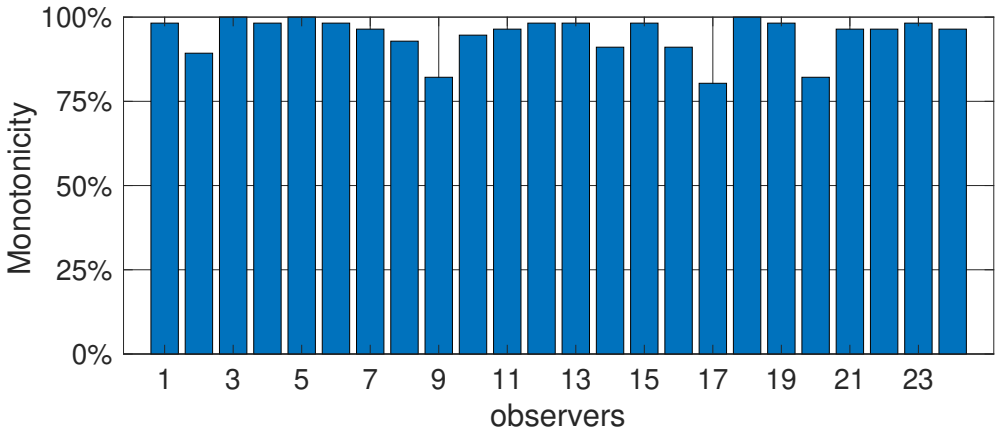



Figure 5.11: Probability that each AIO would predict a higher score for the PVS encoded with a higher bitrate when assessing the visual quality of a pair of PVSs generated from the same SRC and affected by the coding artifacts only. The closer it is to 100 %, the better.

differed than those obtained on the test sets by no more than 5% as it can bee seen from Figure 5.9. For instance, when it comes to the VQEG-HDTV 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 only memorize the training set and can therefore generalize what was learned on it to a set of data never seen before.

The robustness of the AIOs to the noise on input data was also studied. For each AIO, Figure 5.10 reports 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 5.10 were obtained by simulating 10,000 realizations of the noise and counting the number of times in which the AIO did not change its prediction with respect to the noiseless case (the correct ratio) or changed it at most by one quality level on the ACR scale (the acceptable ratio). Figure 5.10 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 quality level on the ACR scale. This result shows that the trained AIOs are rather robust to noise.

Finally, it was studied the ability of the AIOs to distinguish between two stimuli, involving the same source (SRC), with different visual quality. This was done to assess the ability of the AIOs to coherently rank two PVSs coming from the same SRC based on their visual quality. The analysis was performed on the PVSs of

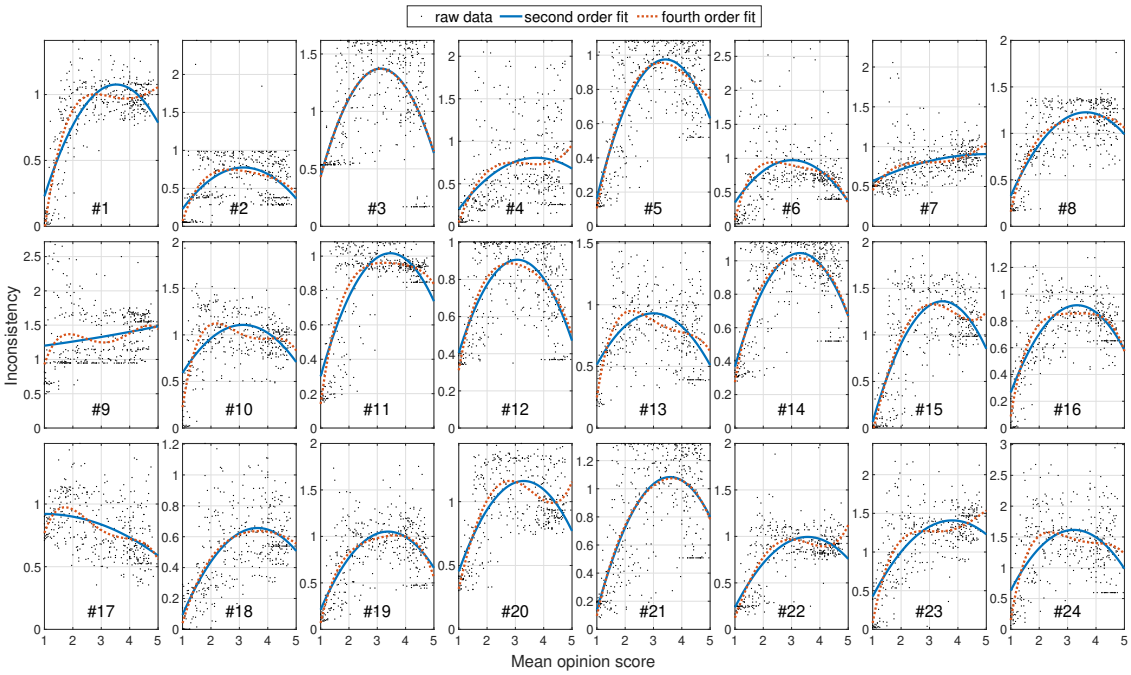



Figure 5.12: Fitting of the inconsistency value with second (red) and fourth (yellow) order polynomials. Fitting functions tend to present an absolute maximum in the central part of the quality scale for almost all the observers, as expected by an inconsistency measure.

the VQEG-HDTV dataset affected only by coding artifacts. Pairs of PVSs derived from the same SRC but encoded at different bit rates were considered. For each AIO, it was computed the fraction of times when a lower score for the PVS encoded with the lower bit rate, as it is typically expected for PVSs derived from the same SRC, was predicted. The results are summarized in Figure 5.11. On average, in 95% of the cases, the AIOs were, able to effectively classify the input stimuli as expected, even though #2, #9, #17 and #20 were a bit less accurate than the other AIOs in this regard.

### 5.5.4 Subjects' Inconsistency

Let's focus now on the results related to the inconsistency measure introduced in (5.3). It is worth reminding 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 (5.3). In the experiments, the 24 AIOs on all the PVSs trained on the VQEG-HDTV datasets were used. Hence for each observer, the value of its inconsistency for each PVS was also estimated. To analyze the properties of the introduced inconsistency measure as a function of the MOS, the widely used mapping that assumes that the five alternatives on the ACR scale are

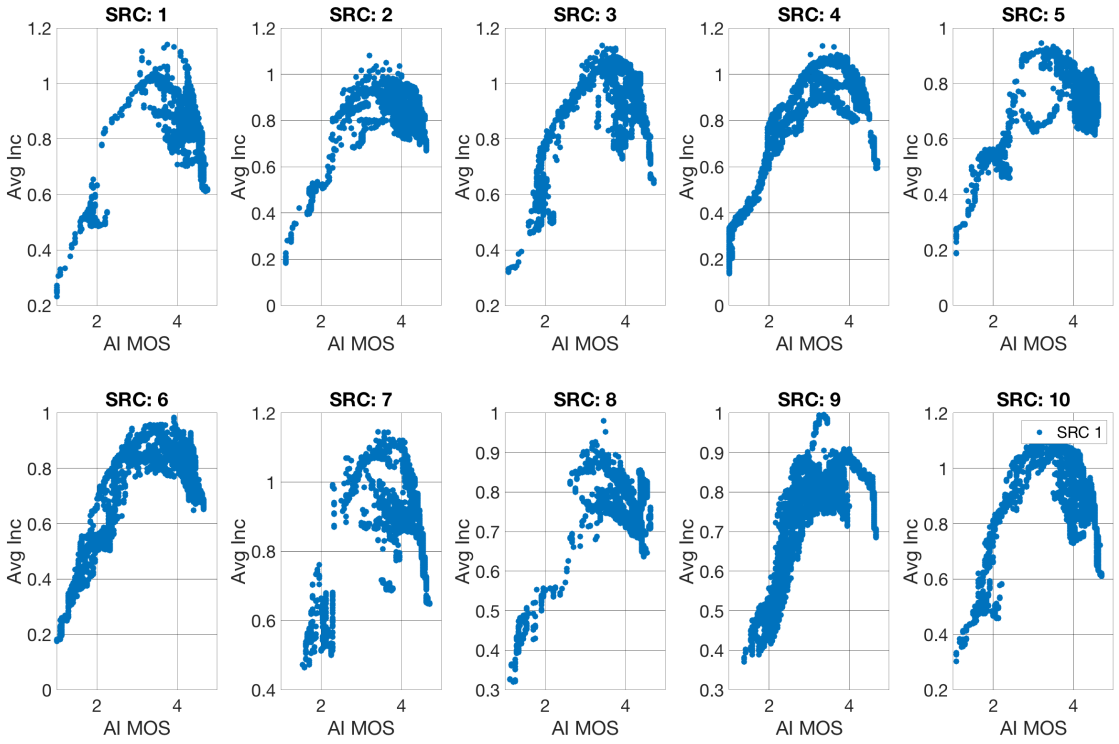



Figure 5.13: The effectiveness of the proposed inconsistency measure on the large scale JEG-Hybrid dataset. The results show that low quality PVSs create less ambiguity (average inconsistency) for the AIOs independently from the SRC as it would have happened with real observers.

equidistant was employed. Therefore, in the analysis, with reference to the Eq 5.3, $v_i = i, \quad i = 1,2,\ldots,5$.

Figure 5.12 reports the inconsistency of each observer on each PVS as a function of the MOS of that PVS. To better visualize the average trend from the points, it was 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 AIOs 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. There were however few AIOs, in particular #7 and #17, that tended to show higher inconsistency as the perceived quality increases.

To investigate the properties of the inconsistency measure at a larger scale, the 24 AIOs were deployed on the JEG-Hybrid dataset that contains almost 60,000 (not subjectively annotated) PVSs, whose characteristics are explained in Table 5.1. For each PVS, the AIOs opinion scores were computed and then averaged to obtain

the AI MOS. The inconsistency of each AIO on each PVS was also computed. Figure 5.13 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.

Finally, it was shown that there is a relationship between the introduced measure of inconsistency and the prominence of artifacts caused by the loss of blocks during transmission. In fact, as shown in Figure 5.14, the inconsistency of almost all the AIOs decreases as the visual disturbance due to the amount of macroblocks lost during transmission becomes more and more perceptible. The Figure 5.14 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 AIOs indicates that the introduced measure of inconsistency captures the typical reliably of human viewers in recognizing the distortion caused by the lost of blocks.

## 5.6 Conclusion

In this chapter a different approach to objectively evaluate media quality as perceived by the end users has been introduced. In particular, every single subject was modeled through a neural network rather than predicting the MOS, differently

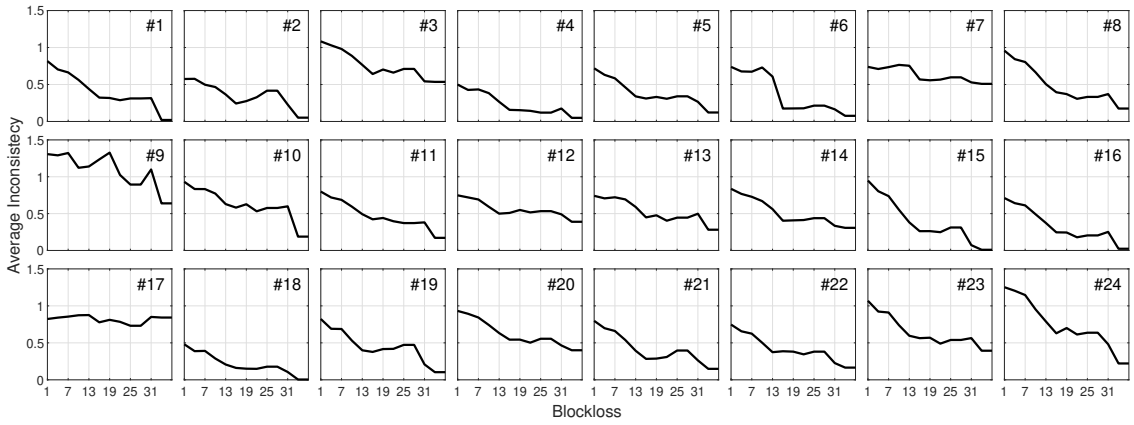



Figure 5.14: The average observers' inconsistency tends to decrease as the "Blockloss" feature value increases for almost all the AIOs. For each value of the Blockloss feature on the x axis, the graph shows the average inconsistency of the observer evaluated on PVSs for which the Blockloss feature value is greater than or equal to the one on the x axis.

from what has traditionally been done in the literature. The approach was called the AIOs-based approach, since the NN of each subject can be seen as an observer with artificial intelligence. Using the ratings of an actual observer gathered during a subjective experiment, a neural network that can then be used as a substitute for that observer was trained. In this way, each NN accounts for the individual characteristics of the related observer, such as personal expectations etc., which have a significant influence on the perception of quality. A deep qualitative analysis illustrated the advantages of the AIOs-based approach and also the flexibility it offers in evaluating the perceived media quality in different contexts.

The computational results demonstrated both the feasibility and the effectiveness of the AIOs-based approach. In particular, neural networks were seen to be a suitable tool in mimicking the choices of an individual subject, and also in allowing to estimate how much confident the subject is in expressing his/her opinion on the quality of a given PVS.

The implementation of the approach was seen to be however data demanding since raw opinion scores of individual viewers are typically noisy. Therefore, specially designed extensive subjective experiments where each subject is willing to evaluate thousands of stimuli, could be required to enhance the effectiveness of the AIOs-based approach using more complex models such as deep convolutional neural networks (CNNs). This could allow for instance to create accurate AIOs of golden eyes viewers. Such AIOs would potentially be very valuable for simulating subjective tests.

As an alternative to designing large scale subjective experiments that are time consuming and require effective tools to manage subjects fatigue, one might relies on approaches such as transfer learning to enable the use of deep CNNs in the design process of the AIOs. The next chapter will investigate such a direction. In particular, it will be presented a deep CNNs based approach to automatically construct, for each observer, the relevant perceptual features to be extracted from the PVSs instead of choosing among a predefined set of hand-crafted features as done in this chapter.

# Chapter 6

# CNNs-based AIOs for No Reference Images Quality Assessment

## 6.1 Introduction

In Chapter 5, the concept of Artificial Intelligence-based Observer (AIOs) has been introduced. An AIO has been defined as a neural network (NN) trained to mimic a human viewer in terms of quality perception. To obtain an AIO, a set of 13 hand-crafted features were first extracted. The AIO of each subject was then obtained by finding those of these features that best characterized that subject and mapping them to his/her opinion scores with a shallow NN.

While the use of hand-crafted features and shallow NNs allows to overcome, up to certain extend, the issues posed by the lack of large scale subjectively annotated datasets, it introduces two main sources of noise in the AIOs' training process: i) hand-crafted features derives from algorithms that attempt to estimate the contribution of a specific artifact or characteristic of the image/video to the determination of its perceptual quality; therefore, due to a potential inaccuracy of these algorithms, the used features might not correctly and/or exhaustively represent the raw stimulus that the observer, to be modeled, has seen and evaluated; ii) the computation of hand-crafted features is not based on the ratings of the observer to be modeled. Different human viewers typically rely on different characteristics of a signal when expressing their opinion on its perceptual quality. In fact, viewers have different sensitivity to the same artifact. Hand-crafted features that are relevant for an observer might not be important for another one. Therefore, when a pre-selected set of hand-crafted features is used, it might be ineffective for modeling certain observers.

To address the first issue, it is important to design AIOs such that they directly process the raw stimulus just like the subject they are modeling do when rating

the perceptual quality. In order to address the second issue, the whole training process should enable the extraction of all and only the features that are really useful to model the opinion scores of each observer. Relying on deep convolutional NNs (CNNs) when training the AIOs would solve both issues.

In fact, deep CNNs-based AIOs would directly receive as input the raw signal and extract from it, through a sequence of convolutional layers, the features that really matter based on the opinion scores gathered from the observer to be modeled. Unfortunately training deep CNNs-based AIOs would be much more demanding in terms of training samples than the approach presented in Chapter 5. This chapter aims at showing, in the still image case (for simplicity's sake), how the transfer learning concept [145] can be leveraged in this context.

The research focusing on the design of Deep CNNs tailored to image classification has attracted and continues to attract the attention of several researchers. As such a large number of deep CNNs architectures with the related weights have been proposed in the computer vision literature [8]. Some attempts to leverage these pre-trained models in media quality assessment for mean opinion score (MOS) prediction, after a transfer learning step, have already been explored [155, 45]. A similar approach is adopted in this chapter. More precisely, instead of designing from scratch a new deep CNN architecture, we make use of the ResNet architecture that has proven to be suitable for the design of media quality assessment models [152].

In particular, two learning steps were seen to be necessary in order to reach accurate deep CNNs-based AIOs. During the first learning step, the architecture of the ResNet50 [40], pre-trained for image classification, was modified and the weights were progressively updated to reach a new deep CNN called JEPGResNet50. The JEPGResNet50 is able to classify images based on their level of JPEG compression. The second learning step refined the generic perceptual quality features already learned by the JEPGResNet50 to obtain new ones that really characterize each subject's quality perception based on his opinion scores. The main stages behind this approach were summarized in the following journal paper [31].

Applying this approach on the data collected in the phase 1 of the "LIVE Multiply Distorted Image Quality" (LIVE-MD-ph1) experiment [56], each of the 19 observers that participated in that experiment was modeled. Doing so, 19 deep CNNs were obtained, one for each observer. These CNNs take an image as input and predict the opinion that the corresponding observer would have expressed after evaluating the quality of that image. The JPEGResNet50 as well as the 19 deep CNNs-based AIOs mimicking actual observers are freely available for research purposes at http://media.polito.it/AIobservers.

Extensive computational experiments were conducted in order to assess the accuracy of the JPEGResNet50 as well as that of the 19 AIOs. When compared to the PSNR [147], SSIM [157] and the BRISQUE [80], it was observed that the JPEGResNet50 is particularly suitable to assess the quality of JPEG compressed

images. Each AIO can mimic, with a rather good accuracy, the corresponding observer yielding the state-of-the-art performance in terms of MOS prediction while also providing an estimation of the distribution of opinion scores.

The remainder of the chapter is organized as follows. Section 6.2 presents related work, followed by Section 6.3 that analyzes in detail the importance of modeling individual observers with deep CNNs. In Section 6.4 the training process of the JPEGResNet50 and the derivation of the AIOs are explained. Computational experiments and the related results are presented in Section 6.5, while conclusions are drawn in Section 6.6.

## 6.2 Related Work

The training of effective deep CNNs-based models is a task that is demanding in terms of the number of training samples. [60]. Because of the lack of large scale subjectively annotated datasets, the question of how to effectively train deep NNs in the context of image quality assessment (IQA) is still an open issue [72]. Several authors have however obtained promising results by relying on deep CNNs. When making use of deep NNs, the following three approaches have mainly been adopted:

1. *Features extraction followed by the use of a deep NN with very few hidden layers.* A Some hand-crafted features are first extracted from the image. A deep NN is then trained feeding it just with the extracted features. In this case, the deep NN builds more detailed features starting from the high-level features provided initially in input. Then it maps these new features to the quality scale. This approach has been used for example in [20, 37, 38].

2. *Direct use of a deep CNN with few hidden layers.* In this case, feature learning and regression are jointly considered in a single optimization process. More precisely, *few* convolutional and pooling layers are subsequently used to extract the perceptual features that model in the best way the average quality perceived by final users. This approach has been leveraged in many papers in the literature, e.g. [58, 26, 142].

3. *Relying on transfer learning.* This last alternative is more recent than the previous ones. To overcome the issue related to the limited size of the training set that precludes the use of a deep CNN with a large number of convolutional layers, transfer learning techniques can be employed. Typically, a large scale pre-trained deep CNN for another task, e.g. image classification, is adjusted by modifying the architecture and the weights of the network. Typically, the weights on the last layers are updated, through an additional training step with a large learning rate on subjectively annotated datasets. In practice, the parts of the network that produce good results tend to remain unaltered,

while the rest is adapted to best fit the new task. Transfer learning has been considered in [77] yielding an objective metric with state-of-the-art performance. In [154] the authors relied on a pre-trained deep NN to determine the type of distortion by which the image quality is impaired and then assessed its visual quality accordingly.

To further enhance the effectiveness of deep learning-based approaches for perceptual quality prediction, before using one of the three aforementioned approaches, many authors considered the possibility of increasing the size of the training set by relying on data augmentation methods (see Section 1.2.3).

The approach in this chapter aims at being a first step toward modeling the quality perception of individual human observers with deep CNNs. The related learning task is more challenging since the noisy nature of the raw individual opinion scores further emphasizes the need of a large set of training samples. To overcome this issue, a deep CNNs, i.e., the JPEGResNet50, was first trained with many synthetically generated training samples. The weights of the JPEGResNet50 were then readjusted on a small scale subjectively annotated dataset in order to obtain accurate deep CNNs-based AIOs.

## 6.3 From Shallow NN to Deep CNN-based AIOs

### 6.3.1 Motivation

Several complex features concur to determine the choices of a subject in terms of quality perception [107]. Furthermore, the set of relevant features might vary from one subject to another. This is because different subjects have different sensitivity to a given type of artifact. Therefore when modeling individual subjects, it is fundamental to make use of approaches that account for such a diversity in the characterization of the same image by different observers. Deep CNN-based models suitably serve this purpose.

Figure 6.1 presents a comparison between the deep CNN-based AIOs presented in this chapter and the shallow NN-based ones trained in the previous chapter. In particular, the use of deep CNNs instead of shallow NNs for modeling the quality perception of an observer allows to address two major issues:

1. **Noise due to the approximation of the input image with hand-crafted features**. The hand crafted features, that are required by shallow NNs as an input, provide a representation of the raw image that might not be perfect. On the other hand, deep CNNs instead directly receive, as an input, the raw image. This allows, first of all, to eliminate a potential noise due to feature inaccuracy that would affect the quality of the final model.

Furthermore, by directly receiving the raw image as an input, deep CNN-based AIOs more precisely resemble actual ones. In fact, observers watch and assess the raw image, not a set of features.

2. **Shallow NNs do not extract features based on the observers' ratings**. Shallow NNs take into account the peculiarities and expectation of an individual observer only during the regression/classification phase of the subject modeling process. This is because the observer's opinion scores are simply used to determine the best way to map the *already chosen* set of hand-crafted features to the quality scale. Therefore, if the selected hand-crafted features are not able to accurately model the quality perception of a given observer, the resulting model will definitely be inaccurate. On the contrary, by relying on deep CNNs, specific features that are of interest for each observer are simultaneously computed together with the mapping from the feature space to the quality scale. In fact, an optimal set of deep CNN-based features for each observer is determined on the basis of his/her opinion scores directly during the training process of his/her AIO.

### 6.3.2 Challenges and Solution Approach

Based on the discussion in Section 6.3.1, it should be clear that the transition from the shallow NN-based approach to the Deep CNN-based one brings considerable advantages when modeling the quality perception of individual observers to create AIOs that can then be used as needed. In practice, however, this transition is hindered by the absence of large-scale annotated datasets, which are fundamental for training deep CNNs.

To overcome this difficulty, a synthetically annotated large-scale dataset was first created. Relying on this dataset, the JPEGResNet50 was trained. Finally, the deep CNN-based AIOs were then derived from the JPEGResNet50 by jointly exploiting a small scale annotated dataset with data collected during a subjective experiment and the transfer learning concept.

In order to make the transfer learning process efficient, the ResNet50 architecture was chosen as the underlying one for training the JPEGResNet50. In fact, the ResNet (Residual Network) architecture has the following advantage compared to the other deep CNNs that have been used in the literature so far: through the introduction of the so called "addition layers" that are computationally cost-less, they learn a residual mapping instead of a direct mapping from the inputs to the label space. It has been empirically observed that this simple trick speeds up the weight optimization process considerably [40], thus allowing to train networks with a very high number of hidden layers in less time than otherwise required.

## 6.4 Training Deep CNNs-based AIOs

In this section, the procedure followed for training the JPEGResNet50 and the AIOs is described in detail. First, the approach adopted for synthetically generating a large scale annotated training set is explained; then, details on the training process of the JPEGResNet50 are provided, followed by a description of the transfer learning steps that yielded the deep CNN-based AIOs.

### 6.4.1 Large Scale Synthetically Created Annotated Dataset

The main idea adopted here to generate training samples is that of finding a way to map different levels of JPEG compression to the ACR scale (see the final result in Table 6.1).

To figure out such a mapping, the data gathered during the phase 1 of the first release of the LIVE image quality assessment (LIVE-IQA-r1-ph1) experiment [118]

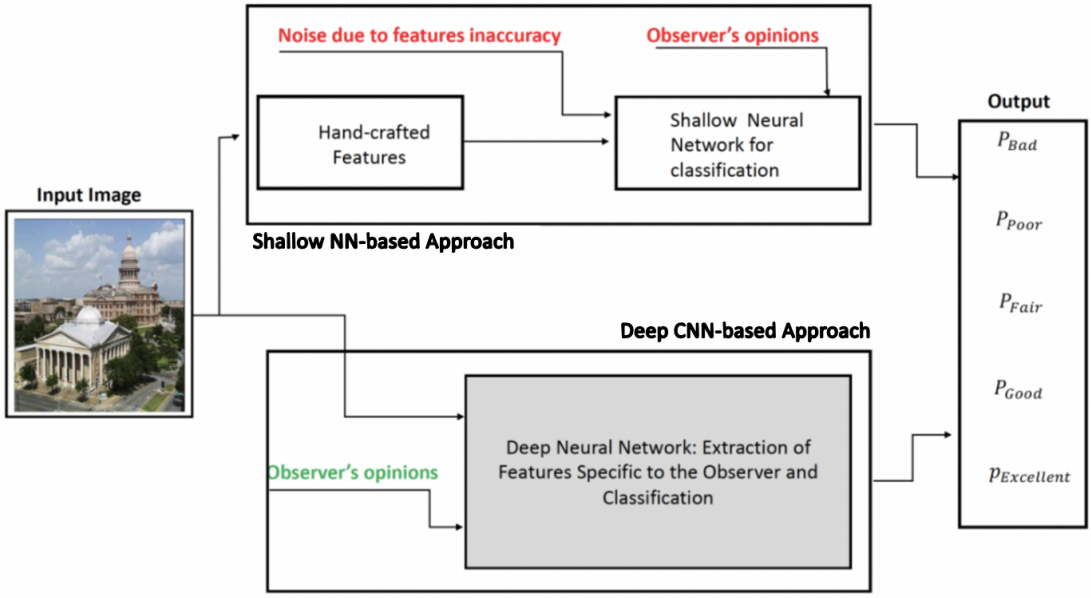



Figure 6.1: Comparison of the deep CNN-based approach and the shallow NN-based one. In both cases the system, after receiving an image or a set of features returns the probability of choosing any of the five options offered by the ACR scale. Note however that, unlike the deep CNNs that receive as input the raw image, the shallow NNs receive hand-crafted features that may not correctly and/or exhaustively characterize the input image. Furthermore, the hand-crafted features are not computed based on the opinion scores of the observers to be modeled when relying on shallow NNs. As such, they might not be the most suitable ones for the observer to be mimicked.

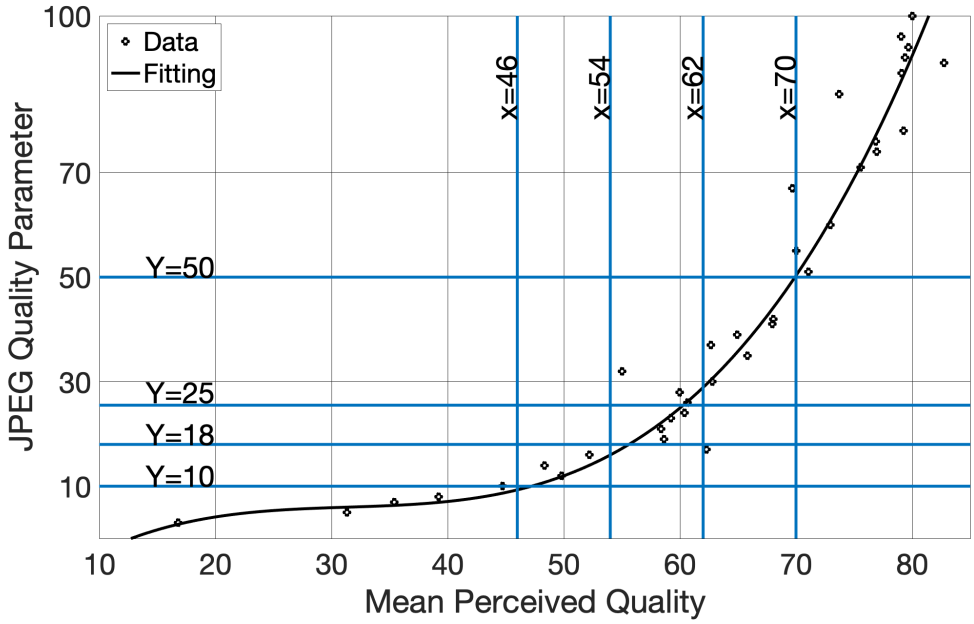



Figure 6.2: Least square fitting of the JPEG quality parameter to the average perceptual quality, on the phase 1 of the first release of the LIVE image quality assessment dataset, using a third order polynomial function.

were considered. Since it was not possible to access the JPEG quality parameter value $Q$ used to create the images in the original dataset, the following procedure was used to estimate it. For each distorted image used during that experiment, its PSNR score $s$ was computed, then the related source image was compressed using many different JPEG quality parameters $Q$, each time computing the PSNR value. Finally, the $Q$ value for which the obtained PSNR was the closest to $s$ was chosen. In this way, for each subjectively evaluated image, the JPEG quality parameter $Q$ that corresponds to its MOS was obtained.

Figure 6.2 reports the average perceived quality for each value of the JPEG quality parameter. The average perceived quality represents the mean of the MOS values of all stimuli sharing the same JPEG quality parameter. The black curve in the figure was obtained by performing a least square fitting of the $Q$ values to the quality scale using a third order polynomial function. This curve provides indications on how different levels of the JPEG compression can be mapped to the subjective quality scale.

Looking at the Figure 6.2, it can be noticed that the viewers did not use the whole 0 to 100 quality scale, as it typically happens [98]. For instance, an average quality of about 45 is observed for images compressed in the very low JPEG quality parameter range of 0 to 10. For this reason, in order to obtain a mapping of the 0 to 100 scale to the ACR scale, a clipping is often used for the boundaries and the

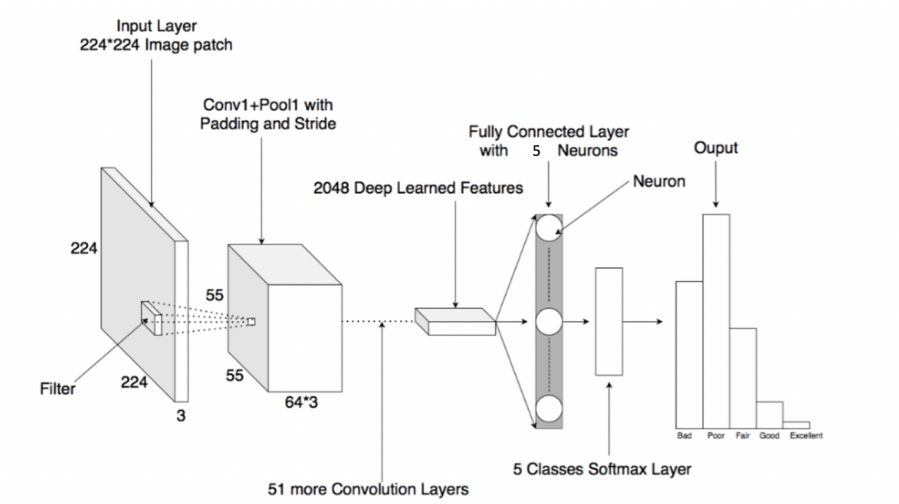



Figure 6.3: Architecture of the JPEGResNet50 as well as of the AIOs. This network receives as an input a $224 \times 224$ color image and provides as an output an estimation of the probability that an average viewer would choose any of the five options of the ACR scale.

remaining part is linearly mapped. In particular, the original 0 to 100 scale was divided into five quality ranges, and converted to the ACR scale as follows: any quality score lying in [0, 46] was mapped to "Bad" (1); the interval [46, 70] was divided into three equally large intervals corresponding respectively to "Poor" (2), "Fair" (3) and "Good" (4); finally, any quality score in [70, 100] was considered as "Excellent" (5). Then, exploiting the fitting curve in Figure 6.2, each of these five quality ranges was mapped to a range of JPEG quality values and the mapping rule in Table 6.1 was obtained.

It is important to note that this mapping rule is not intended to be directly used for predicting the quality of individual images but to generate data for training a neural network to detect quality degradation in a generic way before fine-tuning it and thus optimizing its accuracy for the targeted task. For instance, as it will

Table 6.1: Mapping JPEG Quality parameter intervals to ACR scale.

| **JPEG Quality parameter interval** | **Opinion score** | **Image label** |
|---|---|---|
| [2, 10] | 1 | Bad |
| [11, 18] | 2 | Poor |
| [19, 25] | 3 | Fair |
| [26, 50] | 4 | Good |
| [51, 100] | 5 | Excellent |

be observed later in Section 6.5, if the output of the JPEGResNet50, trained from the data generated based on this rule, is interpreted under probabilistic terms to account for the potential imprecision of this mapping rule, it could provide a good estimator of the MOS of JEPG compressed images.

On the basis of the mapping in Table 6.1, it was created a large-scale synthetically annotated dataset starting from the images available in the ImageNet competition dataset [63] that contains over a million images dedicated to the training and evaluation of deep NN-based models for image classification.

More precisely, 100,000 pristine quality images were selected from the ImageNet dataset. For each of these images, five distorted images were generated by compressing the original pristine quality image using five different values of the JPEG quality parameter. The five values of the JPEG quality parameter were selected at random by choosing one in each interval in the first column of Table 6.1. The quality of each generated image was then annotated by the label associated with the interval to which the related JPEG quality parameter belongs. Therefore, in the end of the procedure, a dataset containing 500,000 annotated JPEG compressed images was obtained.

### 6.4.2 The JPEGResNet50: Architecture and Training Process

Relying on the large-scale synthetically annotated dataset described in the previous section, the JPEGResNet50, i.e. a DNN having the same architecture as that of the ResNet50 except for the last three layers was trained. More precisely, the fully connected, and softmax layers were redesigned to output five probability values. In fact, the JPEGResNet50, after receiving an image as an input, attempts to figure out the probability that an average observer would choose any option among the five available on the ACR scale, after watching and rating the same image. The prediction of the JPEGResNet50 was assimilated to that of an average viewer since the annotation of the training set, as discussed before, is based on an objective rule that maps JPEG compression levels to the average perceived visual quality.

Figure 6.3 presents the architecture of the JPEGResNet50. A $224 \times 224 \times 3$ patch is taken as input image, it then goes through 52 convolutional layers that are meant to progressively extract more and more detailed features characterizing the visual quality of it. Once such features are obtained, they are mapped to the quality scale through the fully connected and softmax layers. The output of such a layer estimates, as shown in Figure 6.3, the probability that the quality of the input image will be assessed as "Bad" (1), "Poor" (2), "Fair" (3), "Good" (4) or "Excellent" (5) by an average observer.

More formally, to train the JPEGResNet50, the label of each image $i$ in the synthetically created large-scale dataset, was encoded as a binary vector $V_i$ whose

entries were defined as follows:

$$V_i(t) = \begin{cases} 1 & \text{if t is the opinion score of image } i \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

where $t = 1,2,\ldots,5$.

Denoting by

- $I$ the total number of images in the training set;
- $\beta$ a vector containing all the weights of the JPEGResNet50 to be computed;
- $p_i^t(\beta) \quad i = 1,2,\ldots,I, \quad t = 1,2,\ldots,5$ the predicted probability that the visual quality of the image $i$ will be rated as $t$, given the weights defined in $\beta$;

the optimization problem that guided the training process of the JPEGResNet50 was formulated as follows:

$$\min_{\beta} \sum_{i=1,2,\ldots,I} \sum_{t=1,2,\ldots,5} -V_i(t) \log(p_i^t(\beta)) \tag{6.2}$$

$$\sum_{t=1,2,\ldots,5} p_i^t(\beta) = 1 \quad i = 1,2,\ldots,I \tag{6.3}$$

$$p_i^t(\beta) \in [0,\ 1]; \quad i = 1,2,\ldots,I; \quad t = 1,2,\ldots,5. \tag{6.4}$$

Eq. (6.2) expressed the minimization of the cross entropy, chosen as the cost function, whereas Eq. (6.3) and (6.4) established the fact that the JPEGResNet50 outputs a probability distribution.

To solve the problem in Eq. (6.2)-(6.4) and thus to train the JPEGResNet50, the well known and widely used stochastic gradient descent with momentum (SGDM) algorithm was used. The SGDM was deployed on a batch containing 90 images at each iteration, this was repeated for 60 periods, i.e. a total of $60 \cdot I/90$ iterations. A small learning rate (0.0001) was adopted to enable the network to progressively transform the initial image classification features into new ones useful for quality assessment. The momentum parameter was set to 0.9 as typically recommended when using the SGDM algorithm.

At the end of the training process all the weights, i.e. the entries of the vector $\beta$, were known. Therefore, receiving an image $i$ as an input, the JPEGResNet50 was able to provide, as output, the following five probability values: $p_i^t(\beta) \quad t = 1,2,\ldots,5$, that represent an estimate of the probability of each of the five options on the ACR scale.

An estimation of the MOS of each image $i$ using the JPEGResNet50 was expressed as:

$$MOS_{res}^i = \sum_{t=1}^{5} t p_i^t(\beta). \tag{6.5}$$

### 6.4.3 Deriving the Deep CNNs-based AIOs

The weights of the JPEGResNet50 can be considered as a suitable starting point for the training of the deep CNN-based AIOs, since it is a deep CNN able to process JPEG compressed images and figure out complex features that characterize their perceptual quality.

To train the AIOs, the data collected during the LIVE-MD-ph1 experiment [56] were considered. That experiment was done with 19 participants. Starting from the JPEGResNet50, exploiting the ratings of the individual subjects and a transfer learning approach, 19 additional deep CNNs were trained, thus obtaining, for each participant, a model capable of predicting his/her choices in terms of perceptual quality.

In more detail, in order to obtain a deep CNN mimicking the quality perception of each of the 19 observers, the training process of the JPEGResNet50 was continued using, as ground truth data, the ratings provided by the observer during the subjective experiment. In this way, the deep CNN modeling each observer takes advantage of the perceptual features previously learned during the training of the JPEGResNet50 on the synthetically generated large-scale dataset. During this additional training phase, the pre-computed features, i.e. those extracted by the JPEGResNet50, were further refined on the basis of the ratings actually provided by each observer. This yielded a new set of deep features for each observer that are expected to better model his/her quality perception. In order not to overfit the small scale subjectively annotated training set, the deep CNN modeling each of the 19 observers was trained for 10 epochs with a learning rate 100 times larger than the one used for training the JPEGResNet50.

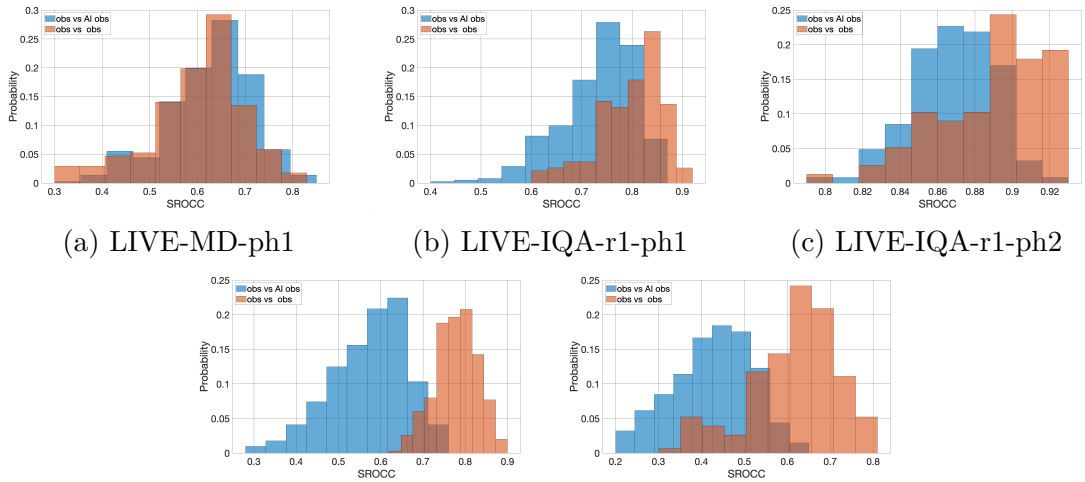
All the 19 deep CNNs obtained at the end of this process share the architecture with the JPEGResNet50 (shown in Figure 6.3) but have different weights. As such, they are deep CNN-based AIOs.

Let consider the deep CNN mimicking the quality perception of an observer $o$: upon receiving in input an image $i$, it provides as output the following five values $p^o_{it}$, $t = 1,2,\ldots,5$, that indicate with which probability the observer $o$ would choose one of the five possible option of the ACR scale, when he/she would be asked to assess the quality of the image $i$. The predicted opinion score $OS^o_i$ of the observer $o$ for the image $i$ is then the one with the highest probability, i.e.

$$OS^o_i = \arg\max_t \left(p^o_{it}\right). \tag{6.6}$$

The MOS of each image $i$ can therefore be estimated by the mean of the opinion scores predicted by the AIOs. It will be referred to as the $MOS_{AI}$.

Modeling individual observers has the advantage of allowing to estimate not only the MOS but also and above all the expected distribution of observer's opinion scores regarding the quality of a given image. Given any image $i$, one might be interested in determining the five probabilities $\alpha^t_i$, $t = 1,2,\ldots,5$, representing the




(a) LIVE-MD-ph1 (b) LIVE-IQA-r1-ph1 (c) LIVE-IQA-r1-ph2

(d) MICT (e) LIVE-MD-ph2

Figure 6.4: Comparing the correlation values observed between the actual observers and the ones of the actual observers and AIOs. The higher the overlap, the better. MD stands for Multi distortion.

expected percentage of the end users that will rate the quality of $i$ choosing $t$ as the corresponding opinion score.

Exploiting the output of the trained deep CNN-based AIOs, such percentages can be estimated as it follows:

$$\alpha_i^t = \frac{1}{19}\sum_{o=1}^{19} p_{it}^o \quad t = 1,2,\ldots,5, \quad i = 1,2,\ldots,I. \tag{6.7}$$

Note that the proposed estimate of the distribution of the observer's opinion scores is not just an empirical distribution derived from the 19 score predicted by the AIOs. Instead, it is derived from the probability values $p_{it}^o$ that takes into account the inability of each observer $o$ to repeat his/her assessment upon many ratings of the image $i$, i.e. his/her inconsistency. By considering the subject inconsistency, one expects that the formula in Eq. (6.7) provides a more robust estimate of the desired distribution than the one that is based on a really limited number of opinion scores.

## 6.5 Numerical Experiments

A number of numerical experiments were conducted to assess the effectiveness of the approach described in this chapter. First, the designed deep CNN-based AIOs were compared to actual observers when used to simulate a subjective test. Then, the accuracy of the deep CNN-based AIOs, in terms of MOS prediction and estimation of the distribution of final user opinion scores, was evaluated.

Table 6.2: PLCC value between the scores of each measures and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive PLCC values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

| DATASET | DISTORTION | BRISQUE | PSNR | SSIM | $\mathbf{MOS_{res}}$ | $\mathbf{MOS_{AI}}$ |
|---|---|---|---|---|---|---|
| CSIQ [64] | JPEG | 0.86 | 0.89 | 0.94 | 0.95 | 0.91 |
| MICT [89] | JPEG | 0.90 | 0.64 | 0.64 | 0.88 | 0.75 |
| SDIVL [23] | JPEG | 0.56 | 0.73 | 0.77 | 0.82 | 0.43 |
| TID2013 [102] | JPEG | 0.81 | 0.91 | 0.92 | 0.94 | 0.84 |
| VCL-FER[151] | JPEG | 0.76 | 0.57 | 0.82 | 0.93 | 0.76 |
| LIVE-IQA-r1 [118] | JPEG | 0.94 | 0.85 | 0.96 | 0.96 | 0.92 |
| LIVE-IQA-r2 [117] | JPEG | 0.96 (T) | 0.95 | 0.92 | 0.91 | 0.86 |
| MICT [89] | JP2K | 0.87 | 0.84 | 0.84 | 0.46 | 0.69 |
| LIVE-IQA-r1 [118] | JP2K | 0.91 | 0.85 | 0.88 | 0.59 | 0.83 |
| LIVE-MD-ph1 [56] | BLUR + JPEG | 0.12 | 0.37 | 0.36 | 0.25 | 0.83 (T) |
| LIVE-MD-ph2 [56] | BLUR + NOISE | 0.01 | 0.53 | 0.42 | 0.02 | 0.52 |

Table 6.3: SROCC value between the scores of each measures and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive SROCC values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

| DATASET | DISTORTION | BRISQUE | PSNR | SSIM | $\mathbf{MOS_{res}}$ | $\mathbf{MOS_{AI}}$ |
|---|---|---|---|---|---|---|
| CSIQ | JPEG | 0.85 | 0.90 | 0.93 | 0.93 | 0.87 |
| MICT | JPEG | 0.92 | 0.60 | 0.66 | 0.87 | 0.75 |
| SDIVL | JPEG | 0.54 | 0.76 | 0.82 | 0.71 | 0.29 |
| TID2013 | JPEG | 0.83 | 0.93 | 0.90 | 0.92 | 0.83 |
| VCL-FER | JPEG | 0.79 | 0.58 | 0.82 | 0.94 | 0.74 |
| LIVE-IQA-r1 | JPEG | 0.92 | 0.93 | 0.94 | 0.92 | 0.85 |
| LIVE-IQA-r2 | JPEG | 0.97 (T) | 0.94 | 0.95 | 0.90 | 0.86 |
| MICT | JP2K | 0.90 | 0.88 | 0.88 | 0.52 | 0.67 |
| LIVE-IQA-r1 | JP2K | 0.92 | 0.92 | 0.91 | 0.69 | 0.78 |
| LIVE-MD-ph1 | BLUR+JPEG | 0.12 | 0.37 | 0.36 | 0.27 | 0.83 (T) |
| LIVE-MD-ph2 | BLUR+NOISE | 0.16 | 0.52 | 0.37 | 0.01 | 0.53 |

### 6.5.1 Deep CNN-based AIOs vs Human Observers

The aim of this first experiment was to verify whether, in case the AIOs are used to simulate a subjective test already done with actual observers, the observed correlation between the ratings expressed by two actual observers will be similar to that between the ratings of an AIO and an actual one. If so, the AIOs may be considered as valid substitutes for real observers. To run the experiment, the datasets coming from five different subjective tests, i.e. the LIVE-MD-ph1 [56], the phase 1 and 2 of the of the first release of LIVE image quality assessment dataset, here abbreviated respectively as (LIVE-IQA-r1-ph1, LIVE-IQA-r1-ph2) [118], the

Table 6.4: RMSE value between the scores of each measure and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive RMSE values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

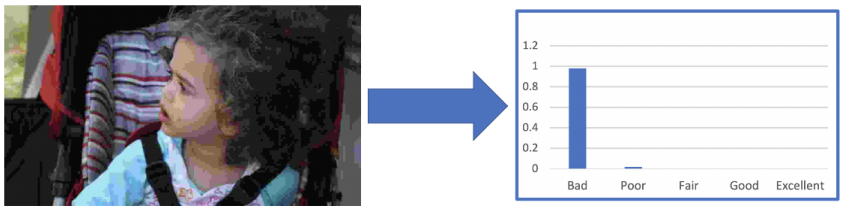
| DATASET | DISTORTION | BRISQUE | PSNR | SSIM | $\mathbf{MOS_{res}}$ | $\mathbf{MOS_{AI}}$ |
|---|---|---|---|---|---|---|
| CSIQ | JPEG | 0.63 | 0.56 | 0.43 | 0.37 | 0.51 |
| MICT | JPEG | 0.51 | 0.89 | 0.90 | 0.55 | 0.76 |
| SDIVL | JPEG | 0.77 | 0.64 | 0.60 | 0.54 | 0.85 |
| TID2013 | JPEG | 0.40 | 0.28 | 0.26 | 0.24 | 0.38 |
| VCL-FER | JPEG | 0.56 | 0.70 | 0.49 | 0.31 | 0.56 |
| LIVE-IQA-r1 | JPEG | 0.33 | 0.49 | 0.25 | 0.26 | 0.35 |
| LIVE-IQA-r2 | JPEG | 0.26 (T) | 0.31 | 0.38 | 0.42 | 0.50 |
| MICT | JP2K | 0.60 | 0.64 | 0.65 | 1.06 | 0.87 |
| LIVE-IQA-r1 | JP2K | 0.35 | 0.45 | 0.41 | 0.69 | 0.47 |
| LIVE-MD-ph1 | BLUR+JPEG | 0.49 | 0.45 | 0.46 | 0.47 | 0.27 (T) |
| LIVE-MD-ph2 | BLUR+NOISE | 0.54 | 0.46 | 0.49 | 0.54 | 0.46 |

MICT dataset [89] and the phase 2 of the LIVE Multiply Distorted Image Quality dataset (LIVE-MD-ph2) [56], were considered. For each of these datasets, the opinion scores expressed by actual observers were available. The AIOs were used as substitutes of the actual observers to simulate the considered subjective tests. More precisely, each image used in each of these five subjective experiments was used as an input to the 19 deep CNN-based AIOs. The opinion scores of each AIO were computed as indicated by Eq. (6.6).

Figure 6.4 shows the histograms of the Spearman Rank Order Correlation Coefficient (SROCC) values between the ratings of a pair made by two real observers (orange histogram) and a pair made by an AIO and an actual observer (blue histogram). The SROCC values between the AIOs and the actual observers are quite similar to those obtained for any pair of the actual observers in the case of the LIVE-MD-ph1, LIVE-IQA-r1-ph1 and the LIVE-IQA-r1-ph2, since the histograms overlap well. This basically indicates that the choices of the AIOs are coherent with those of the actual observers, as expected. For the MICT and LIVE-MD-ph2 datasets, less overlap is observed between the histograms, lower SROCC values are observed among the AIOs and actual observers (from 0.3 to 0.75 for the MICT dataset, from 0.2 to 0.65 for the LIVE-MD-ph2) than those obtained for the actual observers (from 0.6 to 0.9 for the MICT dataset, from 0.3 to 0.8 for the LIVE-MD-ph2). The difference between the SROCC values raises a certain number of questions that deserve more attention and will form the basis of the next research steps. In particular, it should be stressed that the AIOs were trained on the LIVE-MD-ph1 dataset, then tested on other datasets, thus performance might not be optimal. On the other hand, it is important to investigate how context influence factors of any of the involved experiments, e.g. the lighting conditions in the laboratory, the monitor size, the subject's expertise, impacted the reported SROCC

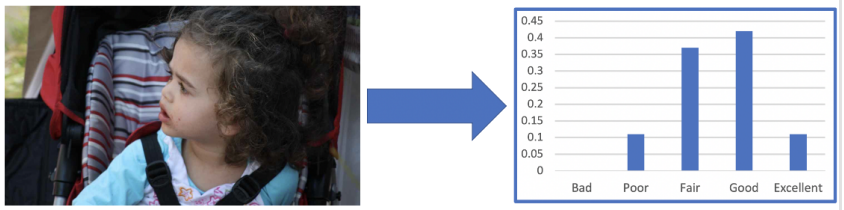
Table 6.5: Results of the statistical test performed for comparing the PLCC values provided by the different metrics on all the datasets. Considering the datasets ordered as they appear in Table 6.2, the k-th digit of the binary sequence in the i-th row and j-th column is 1 if and only if on the k-th dataset, the i-th metric performed significantly better than the j-th one with 95% of confidence. For instance, on the TID2013 dataset (k=4) the $\mathbf{MOS_{res}}$ performed significantly better than the **BRISQUE**.

| | **BRISQUE** | **PSNR** | **SSIM** | $\mathbf{MOS_{res}}$ | $\mathbf{MOS_{AI}}$ | **Total** |
|---|---|---|---|---|---|---|
| **BRISQUE** | — | 01001100100 | 01000010000 | 00000011100 | 01000011100 | 13 |
| **PSNR** | 00110000011 | — | 00000010000 | 00000011101 | 00110011000 | 13 |
| **SSIM** | 1 0110100011 | 10001100000 | — | 0000 0001101 | 10110111100 | 19 |
| $\mathbf{MOS_{res}}$ | 10111100000 | 11101100000 | 01001000000 | — | 11111110000 | 18 |
| $\mathbf{MOS_{AI}}$ | 100000 00011 | 00001100010 | 00000000010 | 00000001111 | — | 10 |

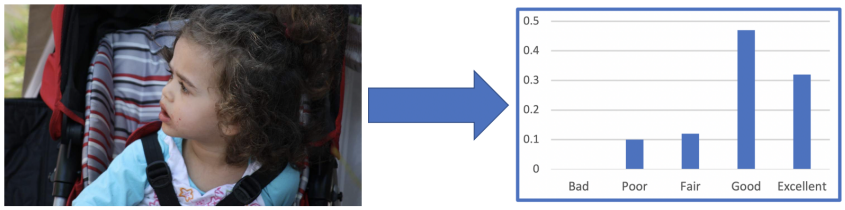



(a) JPEG Quality Parameter equal to 5

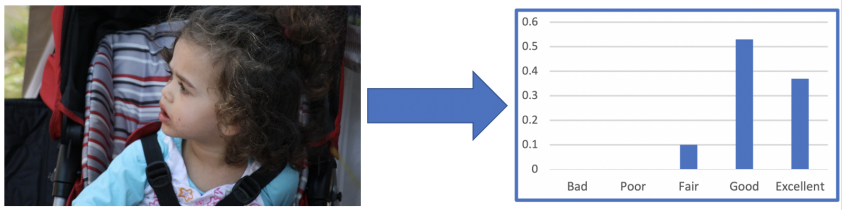

(b) JPEG Quality Parameter equal to 15




(c) JPEG Quality Parameter equal to 35




(d) JPEG Quality Parameter equal to 65

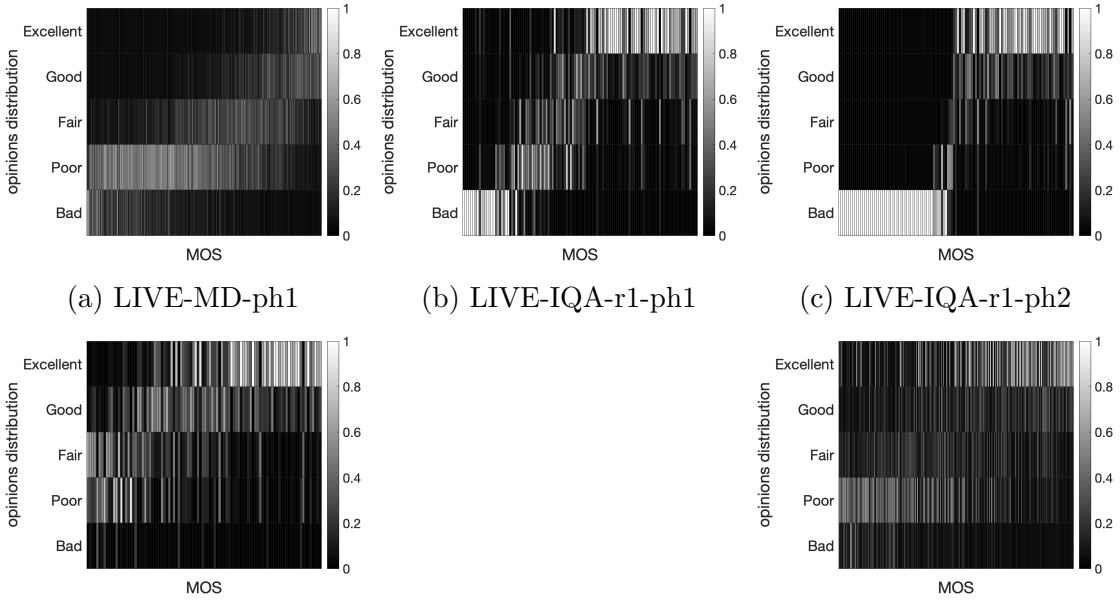



(e) JPEG Quality Parameter equal to 95

Figure 6.5: Showcasing the use of the AIOs in practice. The figure shows the distribution of the user opinions as predicted by the AIOs. The quality of the image given as an input is progressively degraded by applying JPEG compression.

values. In fact, while the ratings of the actual observer are influenced by the context in which they were obtained, those of the AIOs are always determined by the

context influence factors of the subjective experiment whose data are used as the training set. Furthermore, the quality of part of the images in the LIVE-MD-ph2 is impaired by noise artifacts that have never been seen by the AIOs during the training process. The test was done to verify to which extend the accuracy of AIOs depends on the type of distortion considered in the training set.




(a) LIVE-MD-ph1 (b) LIVE-IQA-r1-ph1 (c) LIVE-IQA-r1-ph2

(d) MICT (e) LIVE-MD-ph2

Figure 6.6: The predicted distribution of the user opinions for each image as a function of its MOS. Note that the mode of the distribution tends to increase as the MOS increases. Furthermore, as expected, the distribution is concentrated around the value of the mode in most of the cases.

Table 6.6: Percentage of the images for which the predicted users' distribution of opinions is not statistically different from the empirically observed one.

| **Dataset** | **Percentage of images** |
|---|---|
| LIVE-MD-ph1 | 100% |
| LIVE-IQA-ph1 | 66% |
| LIVE-IQA-ph2 | 90% |
| MICT | 50% |
| LIVE-MD-ph2 | 68% |
| Average | 75% |

### 6.5.2 Predicting the MOS

The accuracy of the JPEGResNet50 as well as that of the AIOs in predicting the MOS of an image was evaluated. The results are summarized in Table 6.2, 6.3 and 6.4. For each image, the PSNR [147], SSIM [157] and BRISQUE [80] scores were computed. The latter is a no reference quality measure, similarly as the models trained in this chapter, while the PSNR and SSIM are full reference measures, which are therefore expected to provide a higher accuracy in terms of MOS prediction. The $MOS_{res}$, i.e. the estimation of the MOS by the JPEGResNet50 as indicated by Eq. (6.5), was also computed. Finally, the $MOS_{AI}$, i.e. the mean of the predicted opinions provided by the 19 AIOs, upon receiving as an input the corresponding image was determined.

Before calculating the Pearson linear correlation coefficient (PLCC) and the root mean square error (RMSE) shown in the Table 6.2 and 6.4, the quality scores of all the considered measures were normalized from their original scale to the MOS scale performing a least square fitting using the following logistic function:

$$\widehat{MOS} = \beta_1 \left( 0.5 + \frac{1}{1 + \exp \beta_2 (VQM - \beta_3)} \right) + \beta_4 \cdot VQM + \beta_5 \tag{6.8}$$

The PLCC, SROCC and RMSE values presented respectively in Table 6.2, Table 6.3 and Table 6.4 show that the trained models are very competitive with respect to all the other measures considered in the experiments in terms of MOS prediction. The JPEGResNet50 is particularly accurate when estimating the quality of the JPEG compressed images. For instance, on the VCL-FER dataset, the $MOS_{res}$ provided by the JPEGResNet50 yielded a PLCC of 0.93 and a SROCC of 0.94, while the BRISQUE only achieved 0.76 and 0.79, respectively. In this case even the PSNR and SSIM led to a lower accuracy in comparison to the output provided by the JPEGResNet50. This is really interesting, if one takes into account the fact that the JPEGResNet50 has been trained using only synthetically generated data. One might hypothesize that such accuracy is due to the fact that the weights of the JPEGResNet50 are learned in such a way that the probability values $p_i^t(\beta)$ in Eq. (6.5) take into account the potential imprecision that affects the labels in the synthetically generated dataset. Specific experiments are however needed to verify the validity of this hypothesis.

The accuracy of the JPEGResNet50 is however strongly dependent on the type of distortion that affects the perceptual quality of the processed image. In fact, the JPEGResNet50 is not able to accurately process images whose quality is impaired by artifacts jointly caused by the blur and the JPEG compression as well as the blur and the noise. This was somehow expected, since the compression process used for generating the synthetic data used for the training of the JPEGResNet50 was strongly related to the JPEG quality parameter. This latter observation highlights the necessity to develop, in future, approaches for artificially generating large-scale

datasets suitable for training deep neural network models that can be deployed for a wider range of applications.

When looking at the prediction of the MOS through the mean of the opinions of the AIOs, i.e. the $MOS_{AI}$, one can observe that the $MOS_{AI}$ predicts the quality of the JPEG compressed images with a lower accuracy than the JPEGResNet50. However, it does perform better when it comes to the assessment of the visual quality of the images distorted by blur and noise artifacts.

The competitiveness of the trained models was further investigated by conducting statistical tests. More precisely, the measures were compared in terms of PLCC on all the datasets while taking into account the statistical significance of the performance gaps. The results are summarized in Table 6.5. As one can notice, none of the measures was significantly more accurate than all the others on all the datasets. Indeed, the results show that the JPEGResNet50 is capable of predicting the quality of JPEG compressed images with an accuracy comparable to that of full reference measures. In fact, while SSIM was significantly more accurate in terms of the PLCC in 19 pair comparisons, the same happened for the $MOS_{res}$ in 18 cases. On the other hand, the $MOS_{AI}$ demonstrated a lower performance. However, in comparison to the other measures, it showed a greater robustness in predicting the quality of the images affected by multiple distortions.

It is fundamental to notice that beyond the high competitiveness of the trained metrics, i.e. the $MOS_{res}$ and the $MOS_{AI}$, in terms of the MOS prediction, they offer in parallel a considerable advantage over the other measures. In fact, both the JPEGResNet50 as well as the model of each single AIO return a discrete probability distribution that can be used to estimate not only the MOS but also the distribution of the opinion scores of the end users on the quality scale. The results related to this particular advantage are presented in the following section.

### 6.5.3 Predicting the Distribution of Users' Opinion Scores

Some applications might demand further measures such as the percentage of end-users that will assess the quality of a given processed image at least as "Fair". Knowing such a percentage would definitely help in enhancing the user QoE. The computation of the probability of each option of the ACR scale, for each image, was performed according to the Eq. (6.7) after passing the image as an input to the 19 AIOs.

Let's start by showcasing the effectiveness of the trained AIOs on an image whose quality is progressively degraded by JPEG compression. Starting from the original pristine image, five distorted images were created by employing different level of JPEG compression. These images were then given as input to the 19 AIOs. The distribution of opinion scores was derived on the basis of the output of the various AIOs. Figure 6.5 illustrates the results. One can notice that the support of the predicted distribution moves progressively to the right as the level of JPEG

compression decreases. Furthermore, the predicted distribution shows a greater variance when the JPEG quality parameter is 35, while for small values of this parameter (5 and 15) the obtained distribution is almost totally concentrated on a single opinion score. This is a very interesting observation because it suggests that deep CNN-based AOIs can replicate the well-known ability of human subjects to consistently evaluate low-quality content.

This preliminary experiment was then generalized by predicting the distribution of user opinions for all the images included in five annotated datasets for which the individual opinion scores are publicly available. Figure 6.6 shows the estimated distribution for each image as a function of its MOS. It can be noticed, as expected, that as the MOS increases, higher probability values are progressively concentrated on the larger opinion scores. In fact, a positive correlation between the MOS and the mode of the distribution of the user opinions can be observed. It is also important to notice that, as it often happens in practice during subjective experiments, the support of the predicted distribution is in almost all the cases concentrated on consecutive opinion scores. This highlights the fact that the deep CNN-based AIOs, during the training process, have been able to capture the ordinal nature of the quality scale. This, however, was not trivial, since none of the constraints of the optimization problem guiding the training process of the AIOs explicitly imposes that.

Statistical tests aiming at determining, for each image, whether the predicted distribution of the user opinion scores is different from the empirical fractions observed during the subjective experiment, with a statistical significance were computed. The Kolmogorov–Smirnov test was used. The tests are performed with 95% of confidence. Table 6.6 reports, for each considered dataset, the percentage of the images for which the predicted distribution can be considered not statistically different than the one observed during the subjective test. In all the cases such a percentage is greater than 50% and on average, in 75% of the cases, the predicted distribution is to be considered not statistically different from the distribution of opinion scores observed during the subjective experiment. The results show therefore a high potential of the AIOs-based approach for going beyond the MOS in QoE measurement scenarios.

## 6.6 Conclusion

This chapter focused on the issue of modeling the quality perception of individual observers using deep CNNs. The purpose of the study was to create deep CNN-based models able to replicate the choices of a real observer in terms of perceptual quality with a high accuracy. To cope with the difficulties related to the training of deep CNNs on small-scale annotated datasets, a synthetically generated large-scale dataset was first created. This was done by mapping progressive levels

of JPEG compression to the ACR scale. Using this dataset, the JPEGResNet50, i.e., a very deep neural network with 52 hidden convolutional layers was trained. The results demonstrate that the JPEGResNet50 can be readily used to accurately evaluate the quality of JPEG compressed images.

To obtain the desired deep CNN-based model of each single observer, a transfer learning step was conducted exploiting the JPEGResNet50 and a small scale subjectively annotated dataset. More precisely, the model that mimics the quality perception of each of the 19 observers considered in this chapter was obtained by continuing the training of the JPEGResNet50 on the considered small scale annotated dataset. During this second learning phase, the perceptual features learned by the pre-trained JPEGResNet50 were further updated/fine-tuned on the basis of the opinions expressed by the observer during the subjective test. This allowed to obtain, for each observer, the optimal set of features that can accurately model his/her quality perception. A total of 19 deep CNNs were therefore trained, one for each observer.

The experiments performed on several datasets highlighted the accuracy of the trained deep CNNs in terms of MOS prediction, while promising results were obtained when comparing the proposed models to the actual observers and estimating the distribution of the user opinion scores on the quality of a given image.

# Chapter 7

# Conclusions

This thesis focused on a number of open problems in the context of media quality assessment that were briefly presented in Chapter 1. In Chapter 2, it was highlighted the importance of reporting the perceptual quality score of any processed video sequence (PVS) under probabilistic terms rather than using the traditional deterministic mean opinion score (MOS) value. In fact, the large number of stochastic influence factors (IFs) that affects the judgment of a human viewer when rating the perceptual quality of a stimuli makes any MOS value, obtained in a subjective test, a possible realization of a random variable, since by repeating the same experience a different MOS value would be observed. Therefore the MOS was treated as a random variable and an approach to compute an interval to which it belongs with a user specified probability was presented. Computational results showcased the effectiveness of such an approach. The approach was published in [29].

The probabilistic representation of the perceptual quality presented in Chapter 2 yields an estimation of a range of quality scores. However, discussions with practitioners in the media industry revealed that having a single numerical quality score coupled with an index that measures the uncertainty that affects such a quality score because of stochastic IFs could be a better option in some cases. For this reason, in Chapter 3 and Chapter 4, approaches to derive indexes that can be used to measure the reliability of any MOS prediction were introduced.

In Chapter 3, the standard deviation of opinion scores (SOS) observed during a subjective text was modeled as the sum of two components: i) a deterministic component, called ground truth SOS (gtSOS), predicted from the scores of several different video quality measures (VQMs); ii) a normally distributed error term caused by the quantization of the quality scale and the use of limited number of participants in subjective tests. In that context, the gtSOS of a PVS can be seen as an index that quantifies its actual ability to confuse viewers and hence the VQMs that are trained using subjective scores. In Chapter 4 instead, a different approach was adopted. An index measuring the level of disagreement between the

quality scores predicted by different VQMs was introduced. A subjective test was conducted to show the effectiveness of such an index in estimating the reliability of a MOS prediction computed using a VQM. The results showed that this index allows distinguishing cases in which the VQMs prediction is likely to be accurate and those where they might fail in predicting the MOS. Part of the analysis presented in Chapter 3 and Chapter 4 was summarized and published in the following journal papers [30, 130].

The last two chapters of this thesis, i.e., Chapter 5 and Chapter 6, focused on the concept of artificial intelligence-based observers (AIOs). An AIO is a neural network (NN) trained to mimic an individual observer in terms of quality perception. Unlike the approaches from Chapter 2 to Chapter 4, the AIOs-based approach to quality assessment accounts to single viewers' characteristics and expectations. A comparative analysis of the AIOs-based approach with respect to the traditional approach based on MOS prediction yielded the conclusion that AIOs allow for a more complete objective quality assessment. The research conducted on the AIOs-based approach has yielded two journal papers [129, 31] and a conference paper [32].

The derivation of AIOs requires performing a learning task in a more noisy context than what typically happens when training models for MOS prediction. This is because the MOS is less noisy than raw opinion scores. The learning task associated with the design of AIOs is typically more demanding in terms of training samples. This makes the learning task challenging, especially in the context of media quality assessment field that is strongly missing large scale annotated dataset with reliable subjective raw opinion scores.

In the context of this thesis, a data augmentation approach was introduced in Chapter 5 in order to train AIOs suitable for a wide range of applications. More precisely, an optimization problem aiming at identifying viewers with similar perception of quality was formulated. Based on its solution, the opinion scores of each observer to model were augmented by those of two observers that were seen to have similar quality perception. The AIOs were then trained feeding shallow NNs with a number of hand-crafted features.

In Chapter 6, two main sources of noise related to the use of hand-crafted features and shallow NNs for the training of AIOs were highlighted. In particular, hand-crafted features might not correctly represent the raw signal; also, the chosen hand-crafted features could not be the most suitable ones to model a given observer. It was then observed that deep convolutional NN (CNN)-based AIOs would not suffer from these two sources of noise.

To overcome the challenges posed by the lack of large scale datasets for the training of deep CNNs, the transfer learning concept was leveraged in Chapter 6, in a ''two-steps" learning approach to obtain deep CNN-based AIOs. The first learning step consisted in training, on a synthetically annotated large scale dataset, the so called JPEGResNet50, i.e., a NN able to classify images based on their level of

JPEG compression. As such, it is a deep CNN that learns relevant features characterizing the perceptual quality. To obtain the deep CNN-based AIO modeling a subject, a second learning step was conducted to adjust the weights of the JPEGResNet50 on the basis of that subject's opinion scores. This approach yielded the training of 19 deep CNN-based AIOs made publicly available for research purposes.

## 7.1 Future Developments

Future developments will mainly focus on the modeling of individual behaviors in terms of quality perception and thus refining the preliminary results of Chapter 5 and Chapter 6. Also, individual behavior will be modeled by relying on some well known discrete choice models.

To further illustrate how general the proposed AIOs-based approach is, additional computational experiments are required to properly assess the correlation between the metrics deriving from AIOs and those proposed in Chapter 2 to Chapter 4. This will constitute one of the directions to which future developments will be devoted.

In particular, at a first glance, the problems addressed from the second chapter of this thesis to the fourth may seem unrelated to the concept of AIO introduced and treated in Chapter 5 and Chapter 6. However, this is not the case.

Since the AIOs can simulate, with their probabilistic output, the inability of actual observers to repeat their first opinion score when rating again a PVS they already evaluated, one can use them to derive many different realizations of the MOS of the same PVS. From these realizations an empirical distribution of the MOS can be computed and the related quantiles would yield quality ranges to be compared to those introduced in Chapter 2.

The diversity observed among the opinion scores predicted by the AIOs for a given PVS (AI-SOS introduced in Chapter 5) could be seen as an indicator of how reliable a MOS prediction is. Hence, the AI-SOS is an alternative approach to be put in comparison with those proposed in Chapter 3 and Chapter 4.

It is worth noting that models for the behavior of individual subjects in media quality assessment are proving to be a very important tool. For instance, the subjects' model proposed in [55] has provided the theoretical foundation to support the fact that reliable subjective experiments can be performed by collecting repeated votes from a few subjects [95]. A future analysis will be dedicated to understanding how the AIOs can concretely help in a similar direction, i.e., making subjective experiments more efficient.

The deep CNN-based AIOs presented in Chapter 6 were designed for still image applications. A natural extension of the approach would be that of deriving deep CNN-based AIOs that can operate on video content. Another question of high interest for the design of the AIOs is how to collect enough reliable subjective raw

opinion scores. In fact new recommendations, tailored for the design of subjective tests aiming at the training of AIOs needs to be thought. Finally, it is important to investigate the aspects of the human perception of quality that a deep NN can really mimic. In other words, it would be interesting to understand whether a deep NN, trained to predict the opinion scores of a human subject, attempts to simulate the mental process that guides human choices or implements a totally different approach that however yields the same prediction. A starting point in this direction could be a comparative analysis of the sensitivity of a human subject and that of his/her AIO to specific modifications on the input signal.

As an alternative to NNs, future work will investigate the modeling of individual choices in media quality assessment using a discrete choice model called "Logit model". The Logit model has already been used in different research fields to model human subjects' choices when they have to select one option from a finite number of alternatives. This is exactly what happens in subjective experiments that use a discrete scale. Being aware that the ability to model discrete choices could have turned out to be a suitable asset in media quality assessment, in the context of my PhD, I dedicated part of my research activity to this aspect. I have published two journal papers involving discrete choice modeling [28, 111]. An application of my findings in the context of discrete choice modeling to media quality assessment has yielded really promising results that I am planning to publish in the next future.

# Appendix A

# List of my Publications

My PhD research activities led to seven journal papers, of which six have already been published and one is ready to be submitted for publication; one book chapter and seven proceedings.

## A.1 Journal Papers

1 2020. L. Fotio Tiotsop, A. Servetti, E. Masala. "An Integer Linear Programming Model for Efficient Scheduling of UGV Tasks in Precision Agriculture under Human Supervision". In: Computer and Operation Research journal.

2 2020. E. Fadda, L. Fotio Tiotsop, D. Manerba, R. Tadei. "The stochastic multi-path Traveling Salesman Problem with dependent random travel costs". In: Transportation Science.

3 2020. L. Fotio Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, E. Masala (2020). "Modeling and Estimating the Subjects' Diversity of Opinions in Media Quality Assessment: A Neural Network Based Approach". In: Multimedia Tools and Applications.

4 2021. L. Fotio Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, E. Masala. "Mimicking Individual Media Quality Perception with Neural Networks based Artificial Observers". In: ACM Transactions on Multimedia Computing, Communications and Applications .

5 2021. M. Roohnavazfar, D. Manerba, L. Fotio Tiotsop, S. H. Reza Pasandideh R. Tadei. "Stochastic single machine scheduling problem as a multi-stage dynamic random decision process". In: Computational Management Science journal.

6 2021. L Fotio Tiotsop, F Agboma, G Van Wallendael, A Aldahdooh, S Bosse, L Janowski, M Barkowsky, E Masala. "On the Link between Subjective Score Prediction and Disagreement of Video Quality Metrics". In: IEEE Access

## A.2 To Be Submitted to Journal

7 2021. L. Fotio Tiotsop, A. Servetti, T. Mizdos, M. Uhrina, P. Pocta, G.Van Wallendael, M. Barkowsky, E. Masala. "Predicting Individual Quality Ratings of Compressed Images through Deep Neural Networks-based Artificial Observers". To be submitted to Signal Processing: Image Communication.

## A.3 Book Chapters

8 2020. E. Fadda, L. Fotio Tiotsop, D. Manerba, R. Tadei. "Optimization Problems under Uncertainty in Smart Cities" In the Handbook of Smart city

## A.4 Proceedings

9 2018. Fadda, E., Fotio Tiotsop, L., Perboli, G., Tadei, R. "The Multi-Path Traveling Salesman Problem with Dependent Random Cost Oscillations". In: Proceedings of Odysseus 2018 - 7th International Workshop on Freight Transportation and Logistics

10 2019. Fotio Tiotsop, L., Masala, E., Aldahdooh, A., Van Wallendael, G., Barkowsky, M. "Computing Quality-of-Experience Ranges for Video Quality Estimation". In: Proceedings of QoMEX 2019 - 11th International Conference on Quality of Multimedia Experience.

11 2019. Fotio Tiotsop, L., Servetti, A., Masala, E. "Optimally Scheduling Complex Logistics Operations Involving Acquisition, Elaboration and Action Tasks". In: Proceedings of RTSI 2019 IEEE 5th International Forum on Research and Technologies for Society and Industry.

12 2020. L. Fotio Tiotsop, A. Servetti, E. Masala. "Full Reference Video Quality Measures Improvement using Neural Networks". In Proceedings of IEEE ICASSP Conference.

13 2020. L. Fotio Tiotsop, A. Servetti, E. Masala. "Investigating Prediction Accuracy of Full Reference Objective Video Quality Measures through the ITS4S Dataset". In Proceedings of Human Vision and Electronic Imaging (HVEI) Conference.

14 2020. L. Fotio Tiotsop, T. Mizdos, M. Uhrina, P. Pocta, M. Barkowsky, E. Masala. "Predicting Single Observer's Votes from Objective Measures using Neural Networks". In Proceedings of Human Vision and Electronic Imaging (HVEI) Conference.

15 2021. L. Fotio Tiotsop, T. Mizdos, E. Masala, M. Barkowsky, P. Pocta. "How to Train No Reference Video Quality Measures for New Coding Standards using Existing Annotated Datasets?". In Proceedings of the IEEE 23th international Workshop on Multimedia Signal Processing (MMSP 2021).

# Appendix B

# Datasets Description and Usage

A brief description of the main datasets used in this thesis to train or motivate the proposed models is provided here together with the main reason why they were chosen. A list of the other datasets that have been used only as test set is also provided with a reference the interested reader may refer to for further details.

## B.1 VQEG HD Phase 1 Experiment Datasets

The VQEG HD Phase 1 Experiment was run in six different laboratories yielding six different datasets traditionally called VQEG-HD1, 2, 3, 4, 5 and 6 within the media quality assessment community. Each of these datasets contains around 160 processed video sequences (PVSs) that were evaluated by 24 subjects. The absolute category rating with hidden reference approach was used to collect participants' opinion scores.

The participants were carefully screened for normal visual acuity (with or without corrective glasses) by means of the Snellen test [123] and for normal color vision by means of the Ishihara test [18]. Furthermore, after the test, a statistical approach was used to check whether the opinion scores provided by each participants were consistent enough with those of the other participants. In case of large inconsistency, the ratings of the related participant were discarded and a new participant was invited. Futher details can be found in the test final report [139]

In each laboratory, the source video sequences were chosen in such a way to cover a large range of content type, i.e., movies, sports, general TV material with much variability as possible. The PVSs were then obtained from the selected source content by applying both compression (e.g., AVC and MPEG-2 encoding) and transmission artifacts (e.g., bursty packet loss).

Based on the efforts invested to ensure the consistency of the gathered opinion scores and considering that a very large set of sources and distortions were integrated, the VQEG Phase 1 experiment datasets can be considered a suitable

assets for research. For this reason, this thesis has made extensive use of them in Chapter 2, Chapter 3 and Chapter 5 where they have been used both as training and testing set. This allowed to infer on the accuracy of the proposed models and tools in a more general application context.

## B.2 The ITS4S Dataset

The ITS4S dataset [96] includes 813 unique source sequences at 1280x720 resolution, each about 4 second long. A subset comprising 514 of these sequences has been AVC compressed by the authors at one of these 5 different bit rate values: 512, 951, 1256, 1732, 2340 kbps. For lower bit rate encoding, lower resolutions have been used [96], however all content has been decoded and upscaled again at 1280x720 before performing any subjective evaluation.

It is worth noting that the ITS4S dataset was originally designed for the training of a no reference video quality metric. As such, a single stimuli approach was adopted during the test. This dataset is somewhat particular as the quality of each PVS was evaluated in absolute terms, i.e., without referring neither to the related source content nor to the quality of another PVS deriving from the same source.

The ITS4S dataset was used in Chapter 3 in order to evaluate whether the proposed SOS model is valid also in the case where the quality is evaluated under absolute terms, since one might expect such a situation to generate more diversity among subjects' opinion scores. It was then also used in Chapter 5 to train and assess the effectiveness of the AIOs when dealing with coding only artifacts.

## B.3 VQMs Disagreement-based Dataset

This dataset was specially designed to assess the effectiveness of the VQMs disagreement index proposed in Chapter 4. The experiment was run from scratch since there was no dataset in the literature including PVSs selected based on a potential disagreement among the VQMs' scores.

The PVSs used in the subjective test were carefully selected to include encoded videos sequence that had high disagreements between the VQMs. A number of encoded videos sequence where all the metrics were consistent with each other were also included in the subjective test. A total of 83 PVSs were put forward to a panel of viewers for the subjective tests. A total of 16 subjects participated in this evaluation across two laboratories in Italy and Germany. The PVSs were selected to cover the full range of impairments from low to high quality, and the double stimuli impairment scale was used, i.e., the reference/source video was shown first, followed by the PVS, and the subject was asked to score, on a five points scale, the artifacts' annoyance in the PVS with respect to the related source.

## B.4 The LIVE Multiply Distorted Phase 1 Experiment Dataset

This dataset derives from a subjective experiment run at the Laboratory for Image and Video Engineering (LIVE). A total of 15 reference/source images was considered. To each of these source images, nine hypothetical reference circuits (HRCs) were applied. These HRCs included: three different levels of JPEG compression, three different levels of blur, and finally 9 HRCs obtained by combining blurring artifacts with JPEG compression. The 15 reference images were put together with the generated distorted images (15*15) yielding a total of 240 images that were evaluated by 19 subjects on a 0 to 100 quality scale.

This dataset was used in Chapter 6 of this thesis to implement the transfer learning step yielding the 19 trained AIOs. The choice to use this specific dataset was mainly motivated by the fact that it offers up to 240 opinion scores for each individual subject.

## B.5 Other Datasets

- The Netflix Public Dataset [68] contains 70 Full HD PVSs subjectively annotated by 26 viewers. While its small number of PVSs precludes the possibility to use it as an effective training set, the fact that the qualities of the 70 PVSs cover the whole quality scale makes it a valid test set. This dataset was used in Chapter 2 and Chapter 3 to validate the proposed approaches.

- The JEG-Hybrid large scale dataset [14] contains around 60,000 HEVC encoded PVSs not subjectively annotated. It was used in Chapter 2 and Chapter 5 to assess the accuracy of the proposed models on a large set of PVSs.

- The TID2013 [102], the MICT [89], the CSIQ [64], the SDIVL [23], the VCL-FER [151] The LIVE image quality assessment dataset [117, 118] and the LIVE multiply distortion phase 2 dataset [56] are well known datasets within the media quality assessment community. They are commonly used to benchmark newly proposed quality metrics. These datasets were used in Chapter 6 to compare the proposed video quality measures with some state-of-the-art ones.

- The ImageNet competition dataset [63] is a well known dataset within the computer vision community. It contains over a million of images aimed at the training of deep neural networks for image classification. 100,000 images were selected from this dataset and used to generate a synthetically annotated training set in Chapter 6

# Bibliography


[1] David W Aha and Richard L Bankert. "A comparative evaluation of sequential feature selection algorithms". In: *Learning from data*. Springer, 1996, pp. 199–206.

[2] Zahid Akhtar and Tiago H. Falk. "Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey". In: *IEEE Access* 5 (2017), pp. 21090–21117.

[3] Shahriar Akramullah. *Digital video concepts, methods, and metrics: quality, compression, performance, and power trade-off analysis*. Springer Nature, 2014.

[4] Ahmed Aldahdooh et al. "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database". In: *Eighth international conference on quality of multimedia experience (QoMEX)*. 2016, pp. 1–6.

[5] Ahmed Aldahdooh et al. "Comparing temporal behavior of fast objective video quality measures on a large-scale database". In: *Picture Coding Symposium (PCS)*. 2016, pp. 1–5.

[6] Ahmed Aldahdooh et al. "Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures". In: *Signal Processing: Image Communication* 74 (2019), pp. 32–41. ISSN: 0923-5965.

[7] Ahmed Aldahdooh et al. "Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in VQA datasets using objective measures". In: *Signal Processing: Image Communication* 74 (2019), pp. 32–41.

[8] Md Zahangir Alom et al. "The history began from AlexNet: A comprehensive survey on deep learning approaches". In: *arXiv preprint arXiv:1803.01164* (2018).

[9] Balasubramanyam Appina and Sumohana S. Channappayya. "Full-Reference 3-D Video Quality Assessment Using Scene Component Statistical Dependencies". In: *IEEE Signal Processing Letters* 25.6 (2018), pp. 823–827.

[10] Apple. *HLS Authoring Specification for Apple Devices*. http://apple.co/39VrP6t. 2020.

[11] Vandana Azad and Pooja Sharma. "A review on objective image quality assessment techniques". In: *International Journal of Emerging Engineering Research and Technology* 2.5 (2014), pp. 188–192.

[12] C. G. Bampis, Z. Li, and A. C. Bovik. "Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.8 (Aug. 2019), pp. 2256–2270.

[13] Christos G Bampis, Zhi Li, and Alan C Bovik. "Spatiotemporal feature integration and model fusion for full reference video quality assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.8 (2018), pp. 2256–2270.

[14] M. Barkowsky et al. "Objective video quality assessment-towards large scale video database enhanced model development". In: *IEICE Transactions on Communications* E98B.1 (2015), pp. 2–11.

[15] N. Barman and M. G. Martini. "QoE Modeling for HTTP Adaptive Video Streaming – A Survey and Open Challenges". In: *IEEE Access* 7 (2019), pp. 30831–30859.

[16] Nabajeet Barman et al. "No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications". In: *IEEE Access* 7 (2019), pp. 74511–74527.

[17] Ankan Bhattacharya and Sarbani Palit. "Measurement of image degradation: a no-reference approach". In: *Multimedia Tools and Applications* 79.9 (2020), pp. 5545–5572.

[18] Jennifer Birch. "Efficiency of the Ishihara test for identifying red-green colour deficiency". In: *Ophthalmic and Physiological Optics* 17.5 (1997), pp. 403–408.

[19] S. Bosse et al. "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment". In: *IEEE Transactions on Image Processing* 27.1 (Jan. 2018), pp. 206–219. ISSN: 1057-7149.

[20] A. Bouzerdoum, A. Havstad, and A. Beghdadi. "Image quality assessment using a neural network approach". In: *Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004.* 2004, pp. 330–333.

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

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

[23] S. Corchs, F. Gasparini, and R. Schettini. "No reference image quality classification for JPEG-distorted images". In: *Digital Signal Processing* 30 (2014), pp. 86–100. ISSN: 1051-2004.

[24] Yashar Deldjoo et al. "Recommender systems leveraging multimedia content". In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–38.

[25] Sai Deng, Jingning Han, and Yaowu Xu. "VMAF Based Rate-Distortion Optimization for Video Coding". In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. 2020, pp. 1–6.

[26] Y. Ding et al. "No-Reference Stereoscopic Image Quality Assessment Using Convolutional Neural Network for Adaptive Feature Extraction". In: *IEEE Access* 6 (2018), pp. 37595–37603.

[27] Yong Ding, Yang Zhao, and Xinyu Zhao. "Image quality assessment based on multi-feature extraction and synthesis with support vector regression". In: *Signal Processing: Image Communication* 54 (2017), pp. 81–92.

[28] Edoardo Fadda et al. "The Stochastic Multipath Traveling Salesman Problem with Dependent Random Travel Costs". In: *Transportation Science* 54.5 (2020), pp. 1372–1387.

[29] Lohic Fotio Tiotsop et al. "Computing Quality-of-Experience Ranges for Video Quality Estimation". In: *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany: IEEE, June 2019, pp. 1–3.

[30] Lohic Fotio Tiotsop et al. "Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach". In: *Multimedia Tools and Applications* (2020), pp. 1–19.

[31] Lohic Fotio Tiotsop et al. "Predicting Individual Quality Ratings of Compressed Images through Deep Neural Networks-based Artificial Observers". In: *To be submitted to Signal Processing: Image Communication* (2021).

[32] Lohic Fotio Tiotsop et al. "Predicting Single Observer's Votes from Objective Measures using Neural Networks". In: *Electronic Imaging* 2020.11 (2020), pp. 130–1.

[33] Pedro Garcia Freitas, Welington YL Akamine, and Mylène CQ Farias. "Using multiple spatio-temporal features to estimate video quality". In: *Signal Processing: Image Communication* 64 (2018), pp. 1–10.

[34] Benoît Frénay and Michel Verleysen. "Classification in the presence of label noise: a survey". In: *IEEE transactions on neural networks and learning systems* 25.5 (2013), pp. 845–869.

[35] Chathura Galkandage et al. "Full-Reference Stereoscopic Video Quality Assessment Using a Motion Sensitive HVS Model". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.2 (2021), pp. 452–466.

[36] Iris Galloso et al. "On the influence of individual characteristics and personality traits on the user experience with multi-sensorial media: an experimental insight". In: *Multimedia Tools and Applications* 75 (Feb. 2016).

[37] Paolo Gastaldo et al. "No-reference quality assessment of JPEG images by using CBP neural networks". In: *International Conference on Artificial Neural Networks*. Springer. 2007, pp. 564–572.

[38] Paolo Gastaldo et al. "Objective quality assessment of displayed images by using neural networks". In: *Signal processing: Image communication* 20.7 (2005), pp. 643–661.

[39] P. Hanhart and R. Hahling. *Video Quality Measurement Tool (VQMT)*. http://mmspg.epfl.ch/vqmt. Sept. 2013.

[40] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[41] C. R. Helmrich et al. "XPSNR: A Low-Complexity Extension of The Perceptually Weighted Peak Signal-To-Noise Ratio For High-Resolution Video Quality Assessment". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 2727–2731.

[42] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. "SOS: The MOS is not enough!" In: *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. Mechelen, Belgium: IEEE, Sept. 2011, pp. 131–136.

[43] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger-Lampl. "SOS: The MOS is not enough!" In: *2011 Third International Workshop on Quality of Multimedia Experience*. Sept. 2011, pp. 131–136.

[44] Tobias Hoßfeld et al. "QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS". In: *Quality and User Experience* 1.1 (Sept. 2016). ISSN: 2366-0147.

[45] Rui Hou et al. "No-reference video quality evaluation by a deep transfer CNN architecture". In: *Signal Processing: Image Communication* 83 (2020), p. 115782.

[46] Qinghua Huang et al. “Personalized video recommendation through graph propagation”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10.4 (2014), pp. 1–17.

[47] Quan Huynh-Thu et al. “Study of Rating Scales for Subjective Quality Assessment of High-Definition Video”. In: *IEEE Transactions on Broadcasting* 57.1 (2011), pp. 1–14.

[48] Mansoor Hyder, Christian Hoene, and Noel Crespi. “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. Jamshoro, Pakistan: Springer, 2012.

[49] ITU-T Rec. BT.500. *Methodology for the subjective assessment of the quality of television pictures*. Jan. 2012.

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

[51] ITU-T Rec. J.149. *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)*. Mar. 2004.

[52] ITU-T Rec. P.1203.1. *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport - Video quality estimation module*. Jan. 2019.

[53] ITU-T Rec. P.1401. *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. July 2012.

[54] Lucjan Janowski and Zdzisław Papir. “Modeling subjective tests of quality of experience with a Generalized Linear Model”. In: *International Workshop on Quality of Multimedia Experience (QoMEX)*. San Diego, CA, USA: IEEE, July 2009, pp. 35–40.

[55] Lucjan Janowski and Margaret Helen Pinson. “The Accuracy of Subjects In A Quality Experiment: A Theoretical Subject Model”. In: *IEEE Transactions on Multimedia* 17 (Dec. 2015), pp. 2210–2224.

[56] Dinesh Jayaraman et al. “Objective quality assessment of multiply distorted images”. In: *2012 Conference record of the forty sixth Asilomar conference on signals, systems and computers*. IEEE. 2012, pp. 1693–1697.

[57] Michael I Jordan and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.

[58] Le Kang et al. “Convolutional neural networks for no-reference image quality assessment”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1733–1740.

[59] Jongyoo Kim and Sanghoon Lee. “Fully deep blind image quality predictor”. In: *IEEE Journal of selected topics in signal processing* 11.1 (2016), pp. 206–220.

[60] Jongyoo Kim et al. “Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment”. In: *IEEE Signal processing magazine* 34.6 (2017), pp. 130–141.

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

[62] L. Krasula, Y. Baveye, and P. Le Callet. “Training Objective Image and Video Quality Estimators Using Multiple Databases”. In: *IEEE Transactions on Multimedia* 22.4 (2020), pp. 961–969.

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[64] E. C. Larson and D. M. Chandler. “Most apparent distortion: full-reference image quality assessment and the role of strategy”. In: *Journal of Electronic Imagings* 19.1 (Mar. 2010), NA.

[65] Mikołaj Leszczuk et al. “Recent developments in visual quality monitoring by key performance indicators”. In: *Multimedia Tools and Applications* 75 (2016), pp. 10745–10767.

[66] Teng Li et al. “No-reference screen content video quality assessment”. In: *Displays* 69 (2021), p. 102030. ISSN: 0141-9382.

[67] Yang Li and Xuanqin Mou. “Joint Optimization for SSIM-Based CTU-Level Bit Allocation and Rate Distortion Optimization”. In: *IEEE Transactions on Broadcasting* (2021), pp. 1–12.

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

[69] Torrin M. Liddell and John K. Kruschke. “Analyzing ordinal data with metric models: What could possibly go wrong?” In: *Journal of Experimental Social Psychology* 79 (Nov. 2018), pp. 328–348. ISSN: 0022-1031.

[70] Joe Yuchieh Lin et al. “A fusion-based video quality assessment (FVQA) index”. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE. 2014, pp. 1–5.

[71] Haojie Liu et al. “Neural Video Coding using Multiscale Motion Compensation and Spatiotemporal Context Model”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[72] Maofu Liu et al. “Special Issue on Recent Advances on Deep Learning for Media Quality Modeling”. In: *Signal processing: Image communication* 78 (Oct. 2019).

[73] Qi Liu et al. “Reduced Reference Perceptual Quality Model With Application to Rate Control for Video-Based Point Cloud Compression”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 6623–6636.

[74] T. Liu, W. Lin, and C.C.-J. Kuo. “Image Quality Assessment Using Multi-Method Fusion”. In: *IEEE Transactions on Image Processing* 22.5 (2013), pp. 1793–1807.

[75] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. “Blind Quality Assessment of Compressed Videos Using Deep Neural Networks.” In: *ACM Multimedia*. 2018, pp. 546–554.

[76] Yongxu Liu et al. “Spatiotemporal Representation Learning for Blind Video Quality Assessment”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2021), pp. 1–1.

[77] T. Lu and A. Dooms. “A Deep Transfer Learning Approach to Document Image Quality Assessment”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1372–1377.

[78] Carlos Mello et al. “A comparative study of objective video quality assessment metrics”. In: *Journal of Universal Computer Science* 23 (Jan. 2017), pp. 505–527.

[79] Karan Mitra, Arkady Zaslavsky, and Christer Ahlund. “Context-Aware QoE Modelling, Measurement and Prediction in Mobile Computing Systems”. In: *IEEE Transactions on Mobile Computing* 14 (May 2015), pp. 920–936.

[80] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-Reference Image Quality Assessment in the Spatial Domain”. In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[81] Tomas Mizdos et al. “Linking Bitstream Information to QoE: A Study on Still Images Using HEVC Intra Coding”. In: *Advances in Electrical and Electronic Engineering* 17 (Dec. 2019).

[82] Decebal C. Mocanu et al. “No-reference video quality measurement: added value of machine learning”. In: *Journal of Electronic Imaging* 24.6 (2015), pp. 1–15.

[83] Arghir-Nicolae Moldovan and Cristina Hava Muntean. “QoE-aware video resolution thresholds computation for adaptive multimedia”. In: *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2017, pp. 1–6.

[84] T.K. Moon. “The expectation-maximization algorithm”. In: *IEEE Signal Processing Magazine* 13.6 (1996), pp. 47–60.

[85] Jim Mullin et al. “New techniques for assessing audio and video quality in real-time interactive communications”. In: *IHM-HCI Tutorial*. Lille, France, 2001.

[86] Anja B. Naumann, Ina Wechsung, and Jörn Hurtienne. “Multimodal interaction: A suitable strategy for including older users?” In: *Interacting with Computers* 22.6 (Nov. 2010), pp. 465–474. ISSN: 0953-5438.

[87] Netflix. *VMAF - Video Multi-Method Assessment Fusion (VMAF) v.0.6.2*. https://github.com/Netflix/vmaf. May 2018.

[88] Netflix developers. *Personal communication*. Jan. 2021.

[89] A. Ninassi et al. “Which semi-local visual masking model forwavelet based image quality metric?” In: *2008 15th IEEE International Conference on Image Processing*. 2008, pp. 1180–1183.

[90] Kento Nishi et al. “Augmentation strategies for learning with noisy labels”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8022–8031.

[91] Geoff Norman. “Likert scales, levels of measurement and the “laws” of statistics”. In: *Advances in Health Sciences Education* 15 (2010), pp. 625–632.

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

[93] Joana Palhais, Rui S. Cruz, and Mário S. Nunes. “Quality of Experience Assessment in Internet TV”. In: *Proc. Intl. Conf. on Mobile Networks and Management*. Aveiro, Portugal: Springer, 2012, pp. 261–274. ISBN: 978-3-642-30422-4.

[94] Andreas S. Panayides et al. “The Battle of the Video Codecs in the Healthcare Domain - A Comparative Performance Evaluation Study Leveraging VVC and AV1”. In: *IEEE Access* 8 (2020), pp. 11469–11481.

[95] Pablo Perez et al. “Subjective Assessment Experiments That Recruit Few Observers With Repetitions (FOWR)”. In: *IEEE Transactions on Multimedia* PP (July 2021), pp. 1–1. DOI: 10.1109/TMM.2021.3098450.

[96] Margaret Helen Pinson. *A Video Quality Dataset with Four-Second Unrepeated Scenes*. NTIA Technical Memo TM-18-532. 2018.

[97] Margaret Helen Pinson, Marcus Barkowsky, and Patrick Le Callet. "Selecting scenes for 2D and 3D subjective video quality tests". In: *EURASIP Journal on Image and Video Processing* 2013.1 (2013), pp. 1–12.

[98] Margaret Helen Pinson and Stephen Wolf. "An objective method for combining multiple subjective data sets". In: *Visual Communications and Image Processing 2003*. Vol. 5150. International Society for Optics and Photonics. 2003, pp. 583–592.

[99] Margaret Helen Pinson and Stephen Wolf. "Comparing subjective video quality testing methodologies". In: *Visual Communications and Image Processing (VCIP)*. Vol. 5150. 2003, pp. 573–582.

[100] Margaret Helen Pinson et al. "The influence of subjects and environment on audiovisual subjective tests: An international study". In: *IEEE Journal of Selected Topics in Signal Processing* 6.6 (2012), pp. 640–651.

[101] L. Po et al. "A Novel Patch Variance Biased Convolutional Neural Network for No-Reference Image Quality Assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.4 (2019), pp. 1223–1229.

[102] N. Ponomarenko et al. "Color image database TID2013: Peculiarities and preliminary results". In: *European Workshop on Visual Information Processing (EUVIP)*. 2013, pp. 106–111.

[103] ITU-T PSTR-CROWDS. *Subjective evaluation of media quality using a crowdsourcing approach*. May 2018.

[104] A. Raake et al. "Multi-Model Standard for Bitstream-, Pixel-Based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P.1204". In: *IEEE Access* 8 (2020), pp. 193020–193049.

[105] Tariq Rahim, Muhammad Arslan Usman, and Soo Young Shin. *Comparing H.265/HEVC and VP9: Impact of High Frame Rates on the Perceptual Quality of Compressed Videos*. 2020. URL: arXiv:2006.02671.

[106] Farah Diyana Abdul Rahman et al. "Reduced-reference Video Quality Metric Using Spatial Information in Salient Regions". In: *Telkomnika* 16.3 (2018), pp. 965–973.

[107] Ulrich Reiter et al. "Factors Influencing Quality of Experience". In: *Quality of Experience: Advanced Concepts, Applications and Methods*. Cham: Springer International Publishing, 2014, pp. 55–72.

[108] Mijke Rhemtulla, Patricia Brosseau-Liard, and Victoria Savalei. "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions". In: *Psychological methods* 17.3 (2012), pp. 354–373.

[109] Thomas Richter et al. "JPEG-XS - A High-Quality Mezzanine Image Codec for Video Over IP". In: *SMPTE Motion Imaging Journal* 127.9 (2018), pp. 39–49.

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

[111] Mina Roohnavazfar et al. "Stochastic single machine scheduling problem as a multi-stage dynamic random decision process". In: *Computational Management Science* (2021), pp. 1–31.

[112] M. J. Scott et al. "Do Personality and Culture Influence Perceived Video Quality and Enjoyment?" In: *IEEE Transactions on Multimedia* 18.9 (2016), pp. 1796–1807.

[113] K. Seshadrinathan et al. "Study of Subjective and Objective Quality Assessment of Video". In: *IEEE Transactions on Image Processing* 19.6 (2010), pp. 1427–1441.

[114] M. Seufert. "Fundamental Advantages of Considering Quality of Experience Distributions over Mean Opinion Scores". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany: IEEE, June 2019, pp. 1–6.

[115] H. R. Sheikh and A. C. Bovik. "Image information and visual quality". In: *IEEE Transactions on Image Processing* 15.2 (Feb. 2006), pp. 430–444.

[116] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms". In: *IEEE Transactions on Image Processing* 15.11 (2006), pp. 3440–3451.

[117] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. "A statistical evaluation of recent full reference image quality assessment algorithms". In: *IEEE Transactions on Image Processing* 15.11 (2006), pp. 3440–3451.

[118] H. R. Sheikh et al. "The LIVE image quality assessment database". In: *http://live.ece.utexas.edu/research/quality* (2005).

[119] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1 (2019), p. 60.

[120] Z. Sinno et al. "Quality Measurement of Images on Mobile Streaming Interfaces Deployed at Scale". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 2536–2551.

[121] Robert C Streijl, Stefan Winkler, and David S Hands. "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives". In: *Multimedia Systems* 22.2 (2016), pp. 213–227.

[122] Robert C. Streijl, Stefan Winkler, and David S. Hands. "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives". In: *Multimedia Systems* 22.2 (Mar. 2016), pp. 213–227. ISSN: 1432-1882.

[123] Stevens Sue. "Test distance vision using a Snellen chart." In: *Community Eye Health* 20.63 (2007), p. 52.

[124] Hossein Talebi and Peyman Milanfar. "NIMA: Neural image assessment". In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 3998–4011.

[125] Michela Testolina and Touradj Ebrahimi. "Review of subjective quality assessment methodologies and standards for compressed images evaluation". In: *Applications of Digital Image Processing XLIV*. Vol. 11842. International Society for Optics and Photonics. 2021, 118420Y.

[126] Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala. "An integer linear programming model for efficient scheduling of UGV tasks in precision agriculture under human supervision". In: *Computers & Operations Research* 114 (2020), p. 104826.

[127] Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala. "Full Reference Video Quality Measures Improvement Using Neural Networks". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 2737–2741.

[128] Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala. "Investigating prediction accuracy of full reference objective video quality measures through the ITS4S dataset". In: *Electronic Imaging* 2020.11 (2020), pp. 93–1.

[129] Lohic Fotio Tiotsop et al. "Mimicking Individual Media Quality Perception with Neural Network Based Artificial Observers". In: 18.1 (2022). ISSN: 1551-6857.

[130] Lohic Fotio Tiotsop et al. "On the Link between Subjective Score Prediction and Disagreement of Video Quality Metrics". In: *IEEE Access* (2021).

[131] Manish Madhava Tripathi, Mohammad Haroon, and Faiyaz Ahmad. "A Survey on Multimedia Technology and Internet of Things". In: *Multimedia Technologies in the Internet of Things Environment, Volume 2*. Ed. by Raghvendra Kumar, Rohit Sharma, and Prasant Kumar Pattnaik. Singapore: Springer Singapore, 2022, pp. 69–87.

[132] Zhengzhong Tu et al. "UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 4449–4464.

[133] Domonkos Varga. "No-Reference Video Quality Assessment Based on the Temporal Pooling of Deep Features". In: *Neural Processing Letters* 50.3 (Apr. 2019), pp. 2595–2608. ISSN: 1573-773X.

[134] Domonkos Varga. "No-reference video quality assessment based on the temporal pooling of deep features". In: *Neural Processing Letters* 50.3 (2019), pp. 2595–2608.

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

[136] *Video Quality Experts Group (VQEG)*. May 2021. URL: http://vqeg.org.

[137] Irene Viola and Pablo Cesar. "A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Contents". In: *IEEE Signal Processing Letters* 27 (2020), pp. 1660–1664.

[138] Heiko A. von der Gracht. "Consensus measurement in Delphi studies: Review and implications for future quality assurance". In: *Technological Forecasting and Social Change* 79.8 (2012), pp. 1525–1536. ISSN: 0040-1625.

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

[140] Mario Vranješ, Snježana Rimac-Drlje, and Krešimir Grgić. "Review of objective video quality metrics and performance comparison using different databases". In: *Signal Processing: Image Communication* 28.1 (2013), pp. 1–19. ISSN: 0923-5965.

[141] Meng Wang et al. "SSIM Motivated Quality Control for Versatile Video Coding". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 1122–1127.

[142] Xiaochuan Wang et al. "No-reference synthetic image quality assessment with convolutional neural network and local image saliency". In: *Computational Visual Media* 5.2 (2019), pp. 193–208.

[143] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[144] Ina Wechsung et al. "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*. New York, NY, USA: Springer, Aug. 2011, pp. 175–186.

[145] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big data* 3.1 (2016), pp. 1–40.

[146] M. J. Wierman and W. J. Tastle. "Consensus and dissention: theory and properties". In: *NAFIPS 2005 - 2005 Annual Meeting of the North American Fuzzy Information Processing Society*. 2005, pp. 75–79.

[147] S. Winkler and P. Mohandas. "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics". In: *IEEE Transactions on Broadcasting* 54.3 (Sept. 2008), pp. 660–668. ISSN: 0018-9316.

[148] L. Xu et al. "Free-Energy Principle Inspired Video Quality Metric and Its Use in Video Coding". In: *IEEE Transactions on Multimedia* 18.4 (Feb. 2016), pp. 590–602.

[149] Long Xu, Weisi Lin, and C-C Jay Kuo. *Visual quality assessment by machine learning*. Springer, 2015.

[150] Jiachen Yang et al. "Full-Reference Quality Assessment for Screen Content Images Based on the Concept of Global-Guidance and Local-Adjustment". In: *IEEE Transactions on Broadcasting* 67.3 (2021), pp. 696–709.

[151] A. Zarić et al. "VCL@FER image quality assessment database". In: *AUTOMATIKA* 53.4 (2012), pp. 344–354.

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

[153] F. Zhang et al. "Video Compression with CNN-based Post Processing". In: *IEEE MultiMedia* (2021).

[154] W. Zhang et al. "Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.1 (2020), pp. 36–47.

[155] Yu Zhang et al. "Objective Video Quality Assessment Combining Transfer Learning With CNN". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.8 (2020), pp. 2716–2730.

[156] Yu Zhang et al. "Video quality assessment with dense features and ranking pooling". In: *Neurocomputing* 457 (2021), pp. 242–253. ISSN: 0925-2312.

[157] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1057-7149.

[158] Yi Zhu et al. “Measuring individual video QoE: A survey, and proposal for future directions using social media”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.2s (2018), pp. 1–24.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.