World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:6, 2014

A New Categorization of Image Quality Metrics Based On a Model of Human Quality Perception

Maria Grazia Albanesi, Riccardo Amadeo

***Abstract*—**This study presents a new model of the human image quality assessment process: the aim is to highlight the foundations of the image quality metrics proposed in literature, by identifying the cognitive/physiological or mathematical principles of their development and the relation with the actual human quality assessment process. The model allows to create a novel categorization of objective and subjective image quality metrics. Our work includes an overview of the most used or effective objective metrics in literature, and, for each of them, we underline its main characteristics, with reference to the rationale of the proposed model and categorization. From the results of this operation, we underline a problem that affects all the presented metrics: the fact that many aspects of human biases are not taken in account at all. We then propose a possible methodology to address this issue.

***Keywords*—**Eye-Tracking, image quality assessment metric, MOS, quality of user experience, visual perception.

I. Introduction

THE switch from the Quality of Service paradigm to the Quality of Experience in multimedia content analysis is driven by the desire of the market to center products and services on the necessities and expectations of the users [1]. The consequence of this idea is that digital content should be optimized to fulfill the expectations of the users, rather than to reduce its impact on the technological platform used to create or deliver it. One of the main problems, then, is to create metrics that could capture the "human perceived quality" of multimedia content. The field of Image Quality Assessment (IQA) is one of the most prolific in delivering new and improved "human quality metrics". The usual categorization of these metrics relates to the necessity of comparing a degraded content (an image) with its original unimpaired version, to obtain a quality score. The metrics that do not require this comparison at all are called No-Reference (NR) metrics. Reduced-Reference (RR) metrics require a limited number of comparisons between degraded and original content (for example, to train the algorithm), while Full-Reference (FR) metrics need to perform the comparison between each distorted version of an image and its original. Our work proposes a new categorization scheme that does not study the performance [2], [3] of the analyzed algorithms or their input requirements. Our scheme groups the most used and reliable metrics according to their level of simulation of the human image quality judging process. The paper is organized as follows. Section 2 presents the chosen metrics. Section 3 includes the categorization itself, with details of the principle we used to develop it and with the explanation of a problem that, in our opinion, is worth to be addressed. Section 4 concludes our study with a brief summary and our proposal to address the problem we identified.

M. G. Albanesi is with the dept. of Electrical, Computer and Biomedical Engineering, University of Pavia, via Ferrata 1, I-27100, Pavia, Italy (e-mail: mariagrazia.albanesi@unipv.it).

R. Amadeo is with the dept. of Electrical, Computer and Biomedical Engineering, University of Pavia, via Ferrata 1, I-27100, Pavia, Italy (e-mail: riccardo.amadeo01@ateneopv.it).

II. Related Work

In literature, plenty of contributions refer to the problem of image quality assessment, as this issue has been currently being studied since the first experiments on image compression. Moreover, with the current pervasive presence of multimedia in different contest (network and mobile systems), the problem of image quality evaluation has become crucial in the computer vision field. In our opinion, the traditional categorizations in subjective vs. objective metrics, and, among the second ones, in full, reduced, and no reference metrics, are too reductive to fully understand which aspects of the human visual and cognitive perception are involved in the metrics. The present paper addresses this problem. However, before considering our modelization of human visual and cognitive perception of an image, it is useful to review briefly the most important contributions in literature.

A. Subjective Metrics

The aim of this study is to categorize the most used image quality objective metrics in relation to how much they take into account the human behavior. Therefore, we need to introduce the ground truth quality metric: the Mean Opinion Score (MOS). To measure the human perceived quality from an image or a set of images, the only fully reliable way is to ask directly to a group of observers how would they rate the quality of the stimulus they were subject to. This is a subjective image quality experiment. Each user gives a quality score for each stimulus he/she watches, and the average of the scores on the same stimulus becomes the MOS for that image. There are guidelines that are used to perform the most accurate and unbiased experiments in many fields [4] - [6], and there are reviews that clearly indicate what are the best methods [7] and rating scales [8] for a IQA or VQA subjective experiment. The human process of multimedia quality

evaluation is not very clear, yet. The Human Visual System has been widely studied in relation to its ability to perceive, understand and rate the quality of a visual stimulus, for example in [9]. It is widely accepted the Human Visual System (HVS) has low-level features and high-level features. Bottom-up models of the HVS mainly exploit low-level characteristics such as contrast sensitivity and perceptual decomposition [10], and they are strictly linked to involuntary, stimulus driven attention mechanisms. Top-down models, instead, rely on high-level features. These features refer to voluntary attention; they are closely connected to the experience of the users and to the assigned task.

However, the generation of the subjective quality score for a stimulus involves all these HVS characteristics, and *more*. For example, the choice of a quality score is also affected by the testers' bias toward the content of the stimulus, by their personal expectations and experience, by the user's language property and understanding, and by the entire set of factors that maybe affect human-to-human communication in general.

B. Objective Metrics

We enumerate the different objective metrics that are part of our categorization including in this list a brief description of the metrics/algorithms. The principle of our choices is to identify and categorize the metrics that are widely recognized as the best performing ones or as the most innovative. Their correlation with the MOS is the measure of their performance, which we do not take into account for the categorization. We relied on other studies [2], [11] to identify the best performing algorithms and, as we are explaining in detail in section 3, we arrange them in categories on the base of the idea behind their development.

1. Full Reference Metrics

PSNR/MSE: These two metrics are the most known and used to measure image similarity. They rely on mathematical comparison between a reference image and a degraded version of the same image. Although being excellent to measure the image similarity, they do not reflect the human perceived quality [12], [13].

DM/NQM[14]: the Distortion Measure and the Noise Quality Measure are two metrics that heavily rely on the notion that there is separation between the psychovisual effects of image filtering and noise. This leads to the creation of two metrics that use the contrast pyramid of Peli's[15] work as a base. Moreover, the contrast interaction between the spatial frequencies of the images and the contrast masking effect are also accounted using this model. The psychovisual effects of frequency distortions are quantified by using a low pass Contrast Sensitivity Function (CSF) and a Discrete Cosine Transform (DCT), a model for the HVS. The last step is the generation of a quality score.

UQI[16]*:* This Universal Quality Index exploits a modelization of the distortions as a combination of luminance distortion, contrast distortion and loss of correlation between the original image and elaborated image. These so called "quality features" are measured locally on the test images and then combined together to obtain a single score for each image.

IFC[17]: the acronym stands for Information Fidelity Criterion. It is centered on an information-theoretic framework based on Natural Scene Statistics (NSS). The inventors of this algorithm worked on the assertion that a Quality Assessment problem can be modeled as a transmission channel. The mutual information between the input and the output of this channel (the original and the elaborated test image) quantifies the amount of information that the human observer can ideally extract from the test image itself. The authors use an NSS Gaussian scale mixture to model the source (input). The distortion (transmission channel) is simply modeled by signal attenuation and additive Gaussian noise (this model captures two important and complementary distortion types on images: blur and additive noise) to create the test stimuli. The IFC criterion is then mathematically derived from the mutual information between the two stimuli. It aims to understand how similar two images are. It is a completely mathematical methodology to calculate the similarity of two images rather than the human perceived image quality.

VIF/VIFP[18]: VIF is the generalization of the IFC. The Visual Information Fidelity metric exploits the same IFC framework with the aim of rating the human perceived quality of an image. It works between two pairs of stimuli: it compares the reference image with a version of itself when the transmission channel has no distortion and when instead the distortion is present. The mutual information between these two pairs of stimuli is then extracted, and the VIF score is a simple ratio of the two. This metric is computationally expensive, so the authors created VIF in the pixel domain (VIFP), to reduce the computational burden.

SSIM[19]: The Structural SIMilarity index is one of the most known and studied image quality metrics, and of the best performing. Its core is the assumption that the HVS has evolved to extract structural information from the stimuli. The metric follows this idea: it considers the distortions as perceived changes in the structural information of an image. The SSIM metric design comes from the perspective of image formation. Two signals are decomposed in three different components, the luminance one, the contrast one and the structure one. These three components undergo a paired comparison and a similarity measure is extracted by combining the result of the comparisons.

MSSSIM [20]: Multi Scale SSIM. As the name explains quite well, this is the extension of the SSIM metric to a multi scale elaboration of the two input signals. This means that after the extraction of luminance, structure and contrast components from the image at scale 1, the algorithm iteratively applies a low pass filter and a downsamples the filtered image by a factor of 2. From this point on, only the contrast and structure component are extracted after each iteration of the algorithm. The final comparison metric is a combination of all the measures extracted at different scales.

FSIM/FSIMc[21]: Feature SIMiliarity. This metric is based on the notion that visually discernable features coincide with those points where the Fourier waves at different frequencies

have congruent phase (points with a high Phase Congruency, PC). PC is the first feature used by the FSIM/FSIMc algorithms. The second feature is the image gradient magnitude (GM). The PC is contrast invariant, so the authors of this algorithm included the GM to include the effect of local contrast in their metric. The difference between FSIM and FSIMc is simply that the first and original algorithm does not account for the chrominance component of the image (it is thought for grey scale images), while the second version does.

HLFSIM[22]: the HLFSIM metric is an enhanced version of FSIM. As explained, FSIM uses visual low-level features to extract an image quality score. HLFSIM extends this set of features by including high-level ones, i.e. the Regions of Interest (ROI). The authors calculate a fixation density map for each image, which is used to "weigh" the low-level features extracted by the FSIM algorithm.

PSNR-HVS, UQI-HVS[23]: these two metrics extend two previously used mathematical metrics for image quality, the PSNR and the UQI, integrating HVS characteristics in both of them. The PSNR-HVS uses the DCT coefficients and the JPEG quantization table to integrate the PSNR, simulating the HVS masking effect on blocks of 64x64 pixels. The UQI enhancement instead is made by a wavelet decomposition of the stimulus into four different subbands. The four subbands of both the reference and the test image undergo a paired comparison to obtain an UQI measure for each of them. These four values are then merged into a single quality score.

PSNR-HVS-M[24]: it is a further improvement of the PSNR-HVS metric. The authors slightly modify the algorithm they introduced to increase its performance. They apply a masking model on the top of the previous algorithm, increasing its similarity with the low-level behavior of the HVS and also increasing its performance in relation to the ground truth data.

PSNR-HA, PSNR-HMA[25]: these two algorithms also propose modifications to the original PSNR, in order to improve its quality prediction capabilities. The authors' contribute, in both cases, is to use PSNR-HVS or PSNR-HVS-M between the reference image and a corrected-mean/corrected-contrast version of the test image. This preprocessing correction is used to calculate a modified version of the MSE for the standard PSNR. The difference between the HA and the HMA algorithm lies in the utilization of PSNR-HVS or PSNR-HVS-M to perform the modified MSE calculation.

VSNR[26]: Visual Signal to Noise Ratio is a wavelet based metric. It addresses the issue that visual detectability of distortion does not always correlate with the human perceived quality. The metric is designed to evaluate both low-level and mid-level HVS features. The aim of this is to include the effects of suprathreshold distortions in the quality score. VSNR works in two stages: the first computes the contrast detection thresholds, while the second estimates visual fidelity by measuring the perceived contrast and the extent to which the distortions disrupt global precedence. The second stage is performed only if the first one finds visually relevant distortions in the test image. The algorithm considers global precedence as a high-level property of the HVS, which is supposed to integrate an image edges in a coarse-to-fine-scale.

2. Blind/No Reference Metrics

BRISQUE[27]: the Blind/Referenceless Image Spatial Quality Elevator utilizes a Natural Scene Statistic (NSS) model framework of locally normalized luminance coefficients. Its aim is to quantify the "naturalness" of a Natural Scene using the parameters extracted by the statistic model. The author of this metric claim that the chosen parameters are sufficient to quantify the naturalness of an image and, consequently, that it is possible to extract a quality score from that information. It requires a calibration phase to tune the algorithm to the type of distortions that are under evaluation.

NIQE[28]: the idea behind the Natural Image Quality Evaluator metric is similar to the precedent one. The authors extract a set of 'quality aware' features from an image and try with mathematical procedures to infer the human perceived image quality. They use a NSS model to derive these features. When the set is completed, the procedure elaborates the image quality as the distance between a multivariate Gaussian (MVG) fit of the NSS features taken from the impaired image, and a MVG model of the quality aware features extracted from the corpus of natural images. The deep difference between NIQE and BRISQUE is that NIQE does not require a tuning phase, and therefore it returns good results without being tied to any specific distortion types.

DIIVINE[29]: Distortion Identification-based Image Verity and INtegrity Evaluation. The statistical properties of Natural Scenes are the basis of this metric, too. The image is decomposed with a wavelet transform: the wavelet coefficients are used to extract a vector of statistical features that is description of the image distortions. This vector is used to estimate the probability that one particular distortion type in a set of many afflicts the stimulus under analysis. Then the algorithm maps the feature vector onto a quality score for each distortion category. In the end, the combination of the probability values with the mapping results returns a final quality value for the image.

BLIINDS-II[30]: it is another non distortion-specific metric based on NSS. The authors considered the NSS modeling and HVS modeling as dual problems. The framework operates entirely in the DCT domain. At first, the test image is subject to a 2D DCT coefficient computation. Then, a generalized Gaussian density model is applied to each block of DCT coefficients. The third step is to derive the generalized Gaussian model parameters, which are used in the last step (in combination with a Bayesian estimator) to obtain a quality score for the image.

III. The Categorization

In order to create a taxonomy of the metrics described in section 2, we propose a model of the MOS generation procedure: it can be divided into different steps, which roughly represent the human quality evaluation process. Our model goes from the appearance of the image to the HVS to

the "formalization" and communication of the human opinion score about the stimulus. We then categorize the metrics introduced in the previous section according to the *main* idea that drove their development, and according to a the steps of the model which are mainly involved in the metrics.

The process of MOS generation is a complex translation of the several component of quality assessment of the human visual perception/cognition into a coded number, according to the adopted scale of opinion scores[31],[32]. We propose to model this process into five steps (see Fig. 1), which ideally represent the "flow" of the subjective score from the image, "through" the user, to the database of the researchers, where the quality scores of all the images are saved before analyzing them. The steps are the following:

- The *image* itself: in this block, we consider all the features of the visual stimuli, which can be related to the pixel or frequency domain. Common characteristics, such as luminance and chrominance components, may have affect differently the human perceived quality. For example, a mobile device offers a different experience from 50' LCD. The metrics that are mainly based on this block have a strong mathematical foundation, with no simulation of the HVS.
- The *physiological* level of the HVS (or the low-level features): in this block, we consider the low-level HVS behavioral characteristics, such as those who exploit contrast sensitivity or masking effects.
- The *cognitive* level of the HVS (high-level features) in this block high level HVS behavior is explored. Instead of relying on near threshold effects, the metrics based on this block consider different biases, such as the memory or semantic conditioning.
- The *attentive* level of the HVS (high-level features): this block describes the driving mechanism of the human gaze. The metrics in this block, together with the previous ones, form the high-level HVS features based metric.
- The *human-human interaction:* in this block, we consider all the aspects involved in the communication of the user score to the researchers, i.e., the language-generated bias, the level of understanding of the rating scale and more.

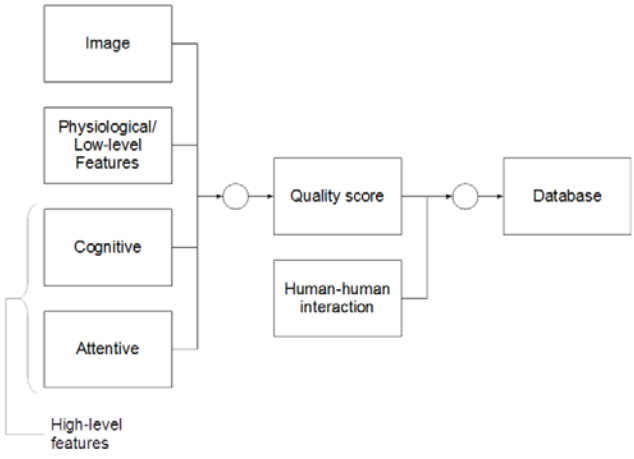

Fig. 1 The model of the MOS generation process

Fig. 2 shows a summarization of the categorization: on the lines, the several metrics are listed. On the columns, the five steps of the model, representing different component of the human perception, are listed. A black arrow means that the metric take into account, in some parts of its definition, the corresponding step. Only MOS generation of subjective metrics is able to cover all the aspects, while all the other objective metrics are very limited in catching the high-level features and in the human-human interaction process.

We call this aspect "the gap" between subjective and objective metrics (see Section IV). In the following Subsections, a deep explanation of the resulting categorization of Fig. 2 is given.

A. Image Based Metrics

Only the metrics that do no simulate directly the HVS behavior appear in this section. This means that we include in this section only the metrics that are a pure mathematical instrument to evaluate the human perceived quality. All the No-Reference metrics we introduced fall under this category: BRISQUE, NIQE, DIIVINE, BLINDS-II. To use the idea expressed in [30]: it is possible to avoid modeling poorly understood functions of the HVS, by exploiting established models of the natural environment. The four metrics rely on the assumption that a collection of "quality-aware" features extracted from a set of Natural Scene Statistics of each stimulus is enough to estimate the human perceived quality. The authors of [29] state that certain statistical properties of the natural scenes are altered in the presence of distortion, and this makes them *un-natural*. In the field of the Full-Reference metrics, instead, it is quite simple to place in this category the MSE and PSNR. These two metrics simply perform a mathematical comparison between the reference and the test image, without taking into account the HVS in any way. Their poor performance in the field of QoE evaluation has been proved in several studies. UQI is a mathematically defined universal image quality index. The authors specify that the index does not depend on the observers, the viewing condition or the test images. It models any distortion as a combination of three factors, as said, that are a form "quality-aware" features. The authors state that the success of their experimental activity is due to the ability of measuring structural distortion that happens when an image is impaired. The structural approach is also the core of SSIM (and its multi scale variation, MSSSIM). The base of this work is the assumption that the HVS is highly adapted to extract structural information from a scene. The direct consequence is that a measure of the structural change of an image can return a good measure of the perceived image distortion. This is what the metric tries to infer, with its algorithm. The philosophy considers image degradations as perceived changes in structural information. The algorithm works on signal decomposition, its final aim is to highlight the structure of the image. There is no direct involvement of any HVS-related procedure beside the hypothesis. Our choice is to consider SSIM and MSSSIM as two mathematically based metrics.

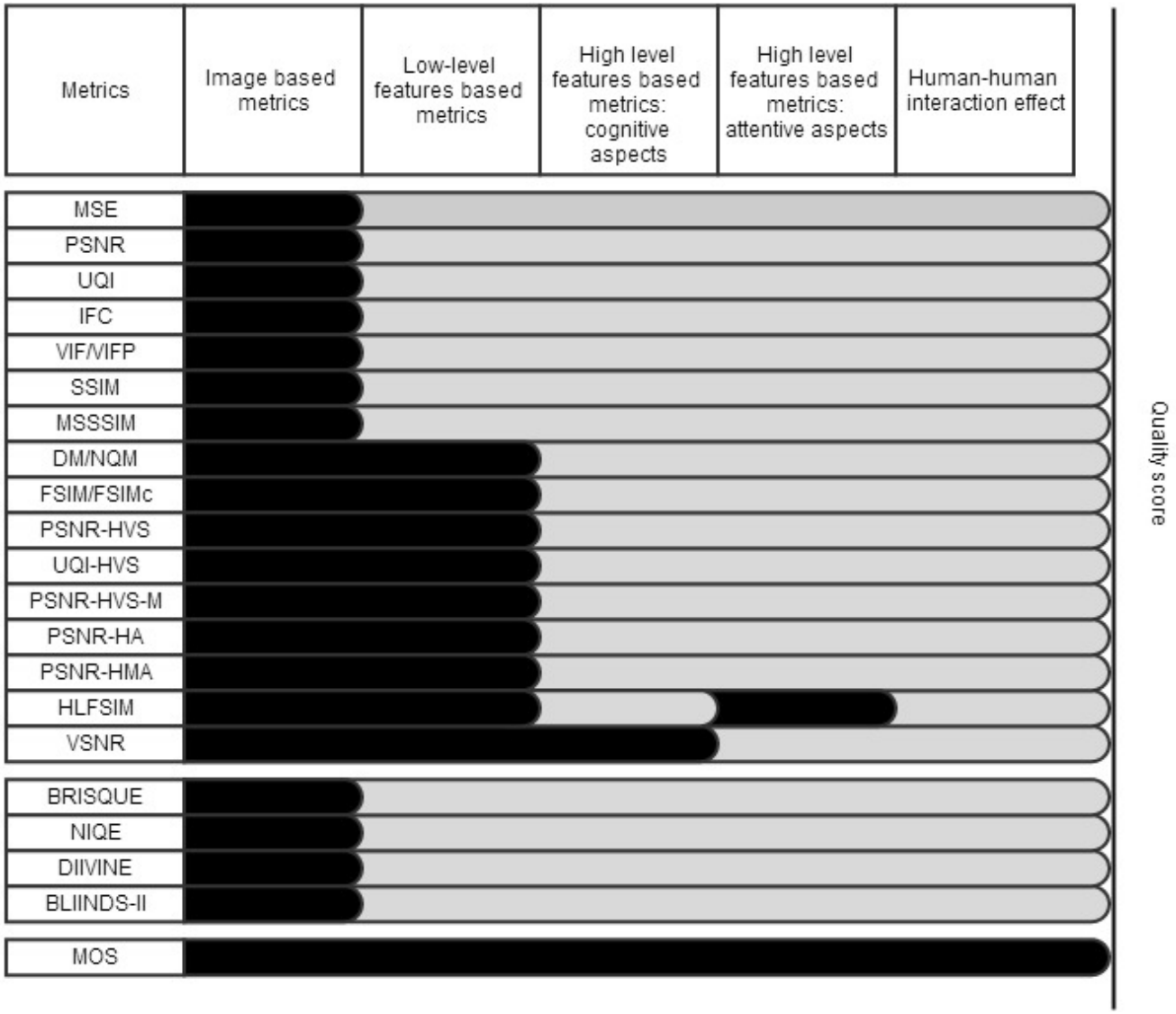

Fig. 2 The novel categorization of image quality assessment metrics in literature. On the columns, the different aspects of human perception involved in the quality assessment task involved in our model

The two metrics, in fact, account for the HVS in the initial general hypothesis (what is the "structural change" of an image? How is it rigorously defined?) but they mainly rely on signal decomposition without modeling the HVS behavior to extract the final quality score. The last algorithms we place in this category are the IFC and the consequent VIF/VIFP. IFC is a mathematical fidelity metric that, like MSE and PSNR, does not involve parameters related to the physics of the display, to the experimental set up or to psychovisual data. The VIF metric design only involves a ratio between to IFC calculated scores; therefore, it shares the same IFC theoretical framework. According to our categorization, then, it is an image based metric.

B. Low-Level HVS Features Based Metrics

This section introduces the metrics that are based on the utilization of low-level features of the HVS. Algorithms/metrics designed around the utilization of masking schemes or temporal/spatial decomposition of signals are the typical examples of the kind of metrics we think can be in this category, and probably the most widely used property of the Human Visual System in these metrics is the contrast sensitivity [9]. Let us analyze first the DM and NQM metrics. We define this two metrics as psychovisually based because, even if they work on signal decomposition, the authors simulate the behavior of the HVS by quantifying the psychovisual effects of frequency distortions using the CSF and the DCT, which are typical procedures that simulate the HVS behavior [9]. In addition to that, or maybe even more importantly, they use the contrast pyramid of Peli's work [15] to model the variation in contrast sensitivity with distance, dimensions and spatial frequency of the stimuli, and with the variation of their local luminance mean. The use of the contrast pyramid is what characterizes this metric. The simulation of nonlinear threshold characteristics of spatial vision makes the NQM and DM metrics the paradigm of low-

level HVS features based metrics. Another approach that relies on low-level HVS features and that differs from the direct simulation of HVS behavior is the one in FSIM. The authors of FSIM state that salient low-level features convey crucial information for the HVS to interpret the scene. Therefore, the image quality can be inferred by comparing low-level features between the reference and the test stimulus. Why is the Feature SIMilarity metric in this category while other "feature-based" metrics were in the previous category, then? Because FSIM algorithm analyses features that can give information about the structure of the scene, and it does not use a "classic" NSS approach. FSIM, in other words, uses the idea behind SSIM (structural similarity) and applies it to known HVS-related relevant image features (Phase Congruency and Gradient Magnitude) instead that to the features that best identify the image structure. The link between PC and GM and the psychovisual behavior of the HVS has been proved in many neurobiology studies [33], [34]. The FSIMc metric is the same metric that accounts for the chrominance component; it then follows the categorization for the original FSIM. In the set of metrics we chose, there is a subset of metrics that take its origin from the PSNR (or the UQI) and that modify it using HVS based procedures. The idea behind PSNR-HVS is to use a weighted cosine transform model to alter the calculation of the PSNR [35], which means that the purely mathematical index becomes an HVS weighted one. The same procedure, even if using the principle of high sensitivity of the HVS to distortion caused in low-frequency range, is used to modify UQI. In consideration of this, we consider both PSNR-HVS and UQI-HVS two low-level HVS features based metrics. PSNR-HVS-M, PSNR-HA, PSNR-HMA are modifications of the PSNR-HVS algorithm that do not alter its founding principle: they follow the parent algorithm in our categorization. We choose to place these metrics in this section because they focus on considering the human behavior, adding its modelization to pure mathematical methodologies. We underline that, even if the structure of the algorithms is still theoretical, the implementation of real-world procedures defines the subset of PSNR based algorithms as "HVS" oriented, which is their real "core".

C. High-Level HVS Features Based Metrics

The HVS primarily analyzes and understands the images based on its low-level features, [33], [34] so the Image Quality Evaluation metrics are usually designed around those kind of characteristics of the stimuli. Recently, though, also the high-level features of the HVS have been studied and implemented into quality algorithms [36], usually following two directions: the cognitive one and the attentive one. The cognitive aspects of the visual process are mainly the ones related to *suprathreshold* detection of the distortions. Low-level features, i.e. the CSF, are fundamental to identify *near-threshold* distortions. These distortions are barely perceived by the human eye due to their spatial frequency being close to the Contrast Sensitivity Function, but they heavily affect the human perceived quality. How to account for effect of suprathreshold distortions in a quality metric, instead, is still under evaluation. Previous research [37], [38] proved that, when a distortion is clearly visible, the low-level HVS effects (such as spatial frequency dependence and masking effects) are, to a first approximation, negligible. The meaning is that cognitive quality algorithms are an interesting topic of research and that there is room to improve the IQA metrics by adding them cognitive models. In the set of metrics we analyzed, we identified only the VSNR as a cognitive based algorithm, which, in fact, directly addresses the suprathreshold distortions problem by proposing a top-down visual model based on *global precedence*(see section 2).

The other approach is the attentive one. An algorithm that uses and attentive approach to IQA is an image that exploits the information given by the "semantically interesting" regions of the stimulus. In simple words, the regions that are "watched" by the observers. Many different psychological effects can influence these Regions of Interest (RoI), (such as change blindness [39], inattentional blindness [40], presence of semantic clue such as faces or bright spots, memory effect and many more) but there is still one only way to identify the RoI (or saliency map) of an image: to ask, directly or indirectly, to the users. From Eye-Tracking experiments and/or subjective experiments, attentive models for automatic saliency prediction have been developed [41]. Those models usually are the base for attentive quality metrics, which are best known as RoI based or saliency based. The set of metrics we identified includes one clearly attentive metric, the HLFSIM. HLFSIM is an enhancement of FSIM that introduces the high-level attention concepts to a low-level feature based metric.

IV. A Proposal to Bridge the Gap

As pointed out in the classification based on the MOS generation model, the main proposals of metrics in literature do not take into account some aspects of the model. However, several interdisciplinary studies are currently under investigation for including cognitive and attentive aspects. In our opinion, the more critic and ignored step is the human-human interaction (see Fig. 1).

A. The Human-Human Interaction Problem

In our schematization of the subjective image quality score collection process, we included a phase in which the numerical value of the perceived quality is generated by the perceptive, cognitive, and attentive processes and is coded and stored in the database of results. Looking at Fig. 1, we modeled a block that represents the personal judgment of each tester. We now focus our analysis on that block: we see that the subjective characteristics of the visual strategy of a generic tester have been studied in the previous blocks, which means that al the metrics we introduced till now cover the psychovisual process of the image quality evaluation. The research on image quality, though, usually requires a database of quality scores. Therefore, the researcher needs to collect a set of quality scores from different testers. Each tester has to communicate, somehow, his/her quality judgment of each stimulus through natural or numerical language. In

consideration of this, to model the IQA process it is necessary to account, for the task of the experiment [42], the experience of the testers, the gender, the cultural background, the psychological expectations, the language property and comprehension and so on. Subjective tests, metrics and results have been widely studied and standardized exactly with this aim but the distance between objective and subjective (ground truth) image quality metrics still exists. Because of this, we suggest that the current methodology of modeling the not-perfectly-understood HVS characteristics into objective quality algorithms might not be the only approach to close the gap between objective quality scores and the Mean Opinion Score. Work in this direction has already been done, for example in [13], [43], [44], but in our analysis of the best performing objective IQA we found that the current state-of-the-art methodologies simply account for visual processes and not judgment/communication processes of the human observers. We can define this problem, *by absurd*, with this question: if we consider the best objective IQA metrics as ground truth for the image quality, did we do enough to "move" the subjective quality evaluation methodologies towards the objective ones? In our opinion, according to the fact that we did not find any account for this in our categorization of current objective metrics, the answer is no. In the next section of this study, we introduce how we will address this topic in the future.

B. A Possible Solution to Bridge the Gap

As pointed out in the previous sections, the current developed metrics are not able to bring inside their algorithms some important aspects of human perception, because it encompass very different components, both from physiological and cognitive/psychological points of view. In particular, by referring to our model of Fig. 1, the High-Level HVS features blocks and the human-human communication blocks are seldom considered. In order to bridge this gap, new ideas are necessary. In our point of view, the proposed model of human perception and the consequent categorization of metrics, highlights one of the shortcomings of the most used image quality metrics: at the present time, it is structurally impossible for these metrics to account for all the voluntary subjectivity that is involved in the human quality evaluation process. This subjectivity is an important part of the procedure and it is one of the causes of the performance deficit of the objective metrics. We showed that the current metrics are not designed to account for these aspects of the judging process and that it might be useful to try different approaches to close the gap between subjective and objective metrics. Literature includes several contributions that underline[45], [46]the correlation between the HVS behavior(registered by an Eye-Tracking device) and the human perceived quality. Moreover, in this paper we explained how the exploitation of High-Level features such as the attentive ones is already being used to develop image-based metrics. For this reason, we suggest to *reverse the standard methodology*: by investigating the correlation between the Eye-Tracking data and the MOS it is possible to bypass the human-human interaction biases without losing information about the cognitive and attentive processes (see Fig. 3).

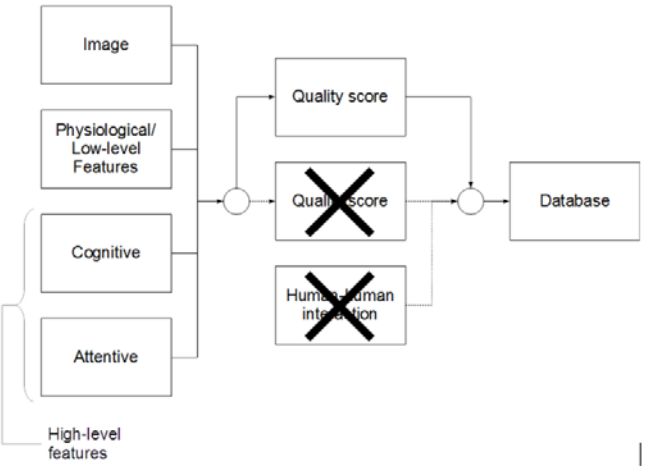

Fig. 3 Our proposed utilization of the Eye-Tracking technology

This approach is able to introduce two different improvements: one is the proposal of a new framework for *subjective* image quality evaluation experiments, where the Eye-Tracked data can be used to infer a physiological measure of the perceived quality, and, at the same time, to exclude all the human-human communication biases.

On the other hand, in the field of *objective* image quality evaluation, it would be possible to create an objective metric based on actual data from the human experience (Eye-Tracked data), instead of data from visual stimuli or statistic models. This is why we called it a "reverse methodology", because it starts from the human side and not from the stimuli side. For this reason, we consider Eye-Tracking a powerful tool to realize this innovation. Fig. 3 shows how, in our proposed approach, the quality score database is populated by data representing the processing of all the blocks of the human perception (Stimulus, Eye, Brain) but it excludes the main sources of arguable criticism (score definition and linguistic limitations). We experimented this approach in our previous studies for *video* quality assessment [47], [48], obtaining promising results. Therefore, we aim to verify this reverse methodology to IQA.

V. Conclusion and Future Work

Our study proposes a new model for the generation of the human perceived quality score of an image. We apply our model to better understand the current, most used IQA metrics. As a result of our investigation, we define a categorization of these metrics that highlights a critical aspect: it is impossible for them to map the human bias introduced by the necessary communication between the observers and the researcher. This indicates a possible reason for the performance gap existing between the current objective QoE metrics and the ground truth data (MOS).

Then, we suggest a methodology based on Eye-Tracking data to reduce this gap, which can beneficial for both objective and subjective metrics. The improvement can be the reduction

of the human-human interaction bias and by disjoining the objective quality score from the image content and from any model of the not perfectly understood HVS.

REFERENCES

[1] M. Fiedler, T. Hossfeld and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36-41, 2012.

[2] W. Lin and J. C.-C. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297-312, 2011.

[3] J. You, U. Reiter, M. Hannuksela, M. Gabbouj and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482-501, 2010.

[4] I.-T. P.910, "Subjective video quality assessment methods for multimedia applications," ITU's Telecommunication Standardization Sector, 04/2008.

[5] I.-T. P.911, "Subjective audiovisual quality assessment methods for multimedia applications," ITU's Telecommunication Standardization Sector, 12/1998.

[6] I.-T. P.913, "Methods for subjectively assessing audiovisual quality of internet video and distribution quality television, including separate assessment of video quality and audio quality, and including multiple environments," ITU's Telecommunication Standardization Sector, 01/2014.

[7] T. Tominaga, T. Hayashi, J. Okamoto and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, Trondheim, 2010.

[8] P. Corriveau, C. Gojmerac, B. Hughes and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Processing*, vol. 77, no. 1, pp. 1-9, 1999.

[9] W. R. Hendee and P. N. T. Wells, The Perception of Visual Information, New York: Springer, 1997.

[10] O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802-817, 2006.

[11] TAMPERE IMAGE DATABASE 2008 TID2008, "TID2008 database," Institute of Signal Processing, Tampere University of Technology, 22 02 2010. [Online]. Available: http://www.ponomarenko.info/tid2008.htm. [Accessed 15 1 2014].

[12] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660-668, 2008.

[13] W. Zhou and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, 2009.

[14] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans and A. C. Bovik, "Image Quality Assessment on a Degradation Model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636-649, 2000.

[15] E. Peli, "Contrast in Complex Images," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032-2040, 1990.

[16] Z. Whang and A. C. Bovik, "A Universal Image Quality Index," *IEEE Signal Processing Letters*, vol. XX, no. Y, pp. 1-4, 2002.

[17] H. R. Sheikh, A. C. Bovik and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117-2128, 2005.

[18] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, 2006.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[20] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment," in *Proceedings of the 37th IEEE Asiloma Conference on Signal, Systems and Computers*, Pacific Grove, CA, 2003.

[21] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, 2011.

[22] P. Dostal, L. Krasula and M. Klima, "HLFSIM: Objective Image Quality Metric Based on ROI Analysis," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, Boston, MA, 2012.

[23] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti and M. Carli, "Two New Full-Reference Quality Metrics Based on HVS," in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics, VPQM, vol. 4*, Chandler, AZ, 2006.

[24] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola and V. Lukin, "On Between-Coefficient Contrast Masking of DCT Basis Functions," in *Third International Workshop on Video Processing and Quality Metrics (VPQM)*, Scottsdale, AZ, 2007.

[25] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, Polyana-Svalyava, 2011.

[26] D. M. Chandler and S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284-2298, 2007.

[27] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, 2012.

[28] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a 'Completely Blind' Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, 2013.

[29] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350-3364, 2011.

[30] M. A. Saad, A. C. Bovik and C. Charrier, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339-3352, 2012.

[31] E. Karapanos, J.-B. Martens and M. Hassenzahl, "Accounting for diversity in subjective judgments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chicago, IL, 2009.

[32] T. Hoßfeld and S. E. Raimund Schatz, "SOS: The MOS is not enough," in *Third International Workshop on Quality of Multimedia Experience*, Mechelen, 2011.

[33] M. C. Morrone and R. A. Ownes, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, no. 5, pp. 303-313, 1987.

[34] P. Kovesi, "Image features from phase congruency," *Videre: A Journal of Computer Vision Research*, vol. 1, no. 3, pp. 1-26, 1999.

[35] N. B. Nill, "A Visual Model Weighted Cosine Transform for Image Compression and Quality Assessment," *IEEE Transactions on Communications*, vol. 33, no. 6, pp. 551-557, 1985.

[36] A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, "Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric," in *IEEE International Conference on Image Processing*, San Antonio, TX, 2007.

[37] D. M. Chandler and S. S. Hemami, "Suprathreshold Image Compression Based on Contrast Allocation and Global Precedence," in *Proceedings of SPIE Human Vision and Electronic Imaging VIII*, Santa Clara, CA, 2003.

[38] M. A. Georgeson and G. D. Sullivan, "Contrast constancy: deblurring in human vision by spatial frequency channels," *The Journal of Physiology*, vol. 252, no. 3, pp. 627-656, 1975.

[39] D. J. Simons and R. A. Rensink, "Change blindness: past, present, and future," *Trends in Cognitive Sciences*, vol. 9, no. 1, pp. 16-20, 2005.

[40] A. Mack and I. Rock, Inattentional Blindness, Bradford Rock, 1998.

[41] U. Rajashekar, I. van der Linde and A. C. Bovik, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564-573, 2008.

[42] Y. Zhong, I. Richardson, S. Arash and P. McGeorge, "Influence of task and scene content on subjective video quality," in *ICIAR (1)'04*, Porto, Portugal, 2004.

[43] A. Eriko, N. Koyu, F. X. Takashi and N. Nagata, "Identification of Factors Related to the Enhancement of Image-Quality for Subjective Image-Quality Assessment Model Based on Psychological Measurement," in *4th International Conference on Human System Interactions (HSI)*, Yokohama, 2011.

[44] J. Lassalle, L. Gros and G. Coppin, "Combination of Physiological and Subjective Measures to Assess Quality of Experience for Audiovisual Technologies," in *Third International Workshop on Quality of Multimedia Experience*, Mechelen, Belgium, 2011.
[45] A. Mittal, A. K. Moorthy, W. Geisler and A. C. Bovik, "Task dependence of visual attention on compressed videos: point of gaze statistics and analysis," in *Human Vision and Electronic Imaging XVI*, San Francisco, CA, 2011.
[46] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba , "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 547-558, 2010.
[47] M. G. Albanesi and R. Amadeo, "A New Algorithm for Objective Video Quality Assessment on Eye Tracking Data," in *9th International Conference on Computer Vision Theory and Applications*, Lisbon, 2014.
[48] M. G. Albanesi and R. Amadeo, "Impact of Fixation Time on Subjective Quality Metric: a New Proposal for Lossy Compression Impairment Assessment," *World Academy of Science, Engineering and Technology*, vol. 59, pp. 1604-1611, 2011.

Maria Grazia Albanesi was born in Pavia in 1962 and graduated cum laude in Electronic Engineering (University of Pavia, 1986) with a Master Thesis on image compression and visual perception. Later she obtained the Ph.D. in Electronic Engineering and Computer Science (Pavia, 1992) with a thesis on VLSI architecture for image compression. She worked at STMicroelectronics on silicon compiler to design dedicated devices for image processing. After this work experience, she joined the Computer Department of Faculty of Engineering of University of Pavia, first as senior researcher (since 1993), then as Associate Professor(since 1998). Her main actual field of interest are quality evaluation of visual media and user-experience driven application for media description and retrieval, and visual data analysis and processing in the field of computational sustainability.

Riccardo Amadeo was born in Pavia in 1985 and got his Master's in computer engineering at the University of Pavia in 2011 with a thesis on subjective and objective video quality assessment. He is currently a Ph. D. candidate (since 2012) at the same university, dept. of Electrical, Computer and Biomedical Engineering, and he continues his research on quality assessment of visual multimedia information.