Semantic-aware blind image quality assessment

Article *in* Signal Processing Image Communication · October 2017

DOI: 10.1016/j.image.2017.10.009

Ernestasia Siahaan
Centrum Wiskunde & Informatica

Judith Redi
Delft University of Technology

Semantic-Aware Blind Image Quality Assessment

Ernestasia Siahaan, Alan Hanjalic, Judith A. Redi

Delft University of Technology, Delft, The Netherlands

Abstract

Many studies have indicated that predicting users' perception of visual quality depends on various factors other than artifact visibility alone, such as viewing environment, social context, or user personality. Exploiting information on these factors, when applicable, can improve users' quality of experience while saving resources. In this paper, we improve the performance of existing no-reference image quality metrics (NR-IQM) using image semantic information (scene and object categories), building on our previous findings that image scene and object categories influence user judgment of visual quality. We show that adding scene category features, object category features, or the combination of both to perceptual quality features results in significantly higher correlation with user judgment of visual quality. We also contribute a new publicly available image quality dataset which provides subjective scores on images that cover a wide range of scene and object category evenly. As most public image quality datasets so far span limited semantic categories, this new dataset opens new possibilities to further explore image semantics and quality of experience.

Keywords: Blind image quality assessment, No-reference image quality metrics (NR-IQM), Quality of experience (QoE), Image semantics, Subjective quality datasets

1. Introduction

A recent report on viewer experience by Conviva shows that users are becoming more and more demanding of the quality of media (images, videos) delivered to them: 75% of users will give a sub-par media experience less than 5 minutes before abandoning it [1]. In this scenario, mechanisms are needed that can control and adaptively optimize the quality of the delivered media, depending on the current user perception. Such optimization is only possible if guided by an unobtrusive, automatic measure of the perceived Quality of Experience (QoE) [2] of users.

Algorithms that predict perceived quality from an analysis of the encoded or decoded bitstream of the media content are often referred to as Image Quality Metrics (IQM), and are typically categorized into full-reference (FR) or no-reference (NR) methods [3]. FR methods predict quality by comparing (features of) an impaired image with its original, pristine version. NR methods, on the other hand, do not rely on the availability of such reference images, and are therefore preferred for real time and adaptive control of visual quality.

NR methods often approach the problem of predicting quality by modeling how the human visual system (HVS) responds to impairments in images or videos [3, 4]. This approach implies that users' QoE depends mostly on the visibility of impairments, and that a measure of visual sensitivity alone is enough to predict visual quality. In this paper, we challenge this view, and we prove empirically that semantic content, besides impairment visibility, plays an important role in determining the perceived quality of images. Based on this result, we propose a new paradigm for IQM which considers semantic content information, on top of impairment visibility, to more accurately estimate perceived image quality.

The potential to exploit image semantics in QoE assessment has already been recognized in previous research that investigated the influence of various factors, besides impairment visibility, on the formation of QoE judgments. Context and user influencing factors [2, 5, 6, 7, 8], such as physical environment, task, affective state of the user and demographics, have been shown to be strong predictors for QoE, to the point that they could be used to automatically assess the perceived quality of individual (rather than average) users [6]. A main drawback of this approach is that information about context of media consumption or preferences and personality of the user may prove difficult to collect unobtrusively, or may require specific physical infrastructure (e.g., cameras) or data structure (e.g., preference records). As a result, albeit promising, this approach has limited applicability to date.

A separate but related trend has instead looked into incorporating in image quality metrics higher level features of the HVS that enable cognition, such as visual attention [9]. This has been shown to bring significant accuracy improvements without an excessive computational and infrastructural overhead, as all information can be worked out from the (decoded) bitstream.

The first steps in this direction have investigated the role of visual attention in quality assessment [9]. In [10], it was shown that impairments located in salient or visually important areas of images are perceived as more annoying by users. Because those areas are more likely to attract visual attention, the impairments they present will be more visible and therefore more annoying. Based on this rationale, a number of studies have confirmed that, by adding saliency and/or visual importance information to quality metrics, their accuracy can be significantly improved [11, 12, 13].

The study in [14] brought this concept further by identifying visually important regions with those having richer semantics, and incorporating a measure of semantic obviousness into image quality metrics. The study reasoned that regions presenting clear semantic information would be more sensitive to the presence of impairments, which may be judged more annoying by users as they hinder the content recognition. The authors therefore proposed to extract the object-like regions, and weight them based on how likely the region is actually containing an object. They would then extract local descriptors for evaluating quality from the top-N regions.

In this work, we look deeper at the role that semantics plays in image quality assessment. Our rationale relies on the widely accepted definition of vision by Marr [15]: vision is the process that allows to know what is where by looking. As such, vision involves two mechanisms: the filtering and organizing of visual stimuli (perception), and the understanding and interpreting of these stimuli through recognition of their content [16]. The earliest form of interpretation of visual content is semantic categorization, which consists of recognizing (associating a semantic category to) every element in the field of view (e.g., "man" or "bench" in the top-left picture in Figure 1). In vision studies, semantics refers to meaningful entities that people recognize as content of an image. These entities are usually categorized based on scenes (e.g., landscape, cityscape, indoor, outdoor), or objects (e.g., chair, table, person, face).

It is known that early categorization involves basic and at most superordinate semantic categories [17, 18], which are resolved within the first 500 ms of vision [19]. Most of the information is actually already processed within the first fixation (∼100 ms, [20]). Such a rapid response is motivated by evolutionary mechanisms, and is at the basis of every other cognitive process related to vision. When observing impaired images, however, semantic categories are more difficult to be resolved [21]. The HVS needs to rely on context (i.e. other elements in the visual field) to determine the semantic category of, e.g., blurred objects. This extra step (1) slows down the recognition process, and (2) reduces the confidence on the estimated semantic category. In turn, this may compromise later cognitive processes, such as task performance or decision making. Hence, visual annoyance may be a reaction to this hindrance, and may depend on the entity of the hindrance as well as on the semantic category of the content to be recognized. Some categories may be more urgent to be recognized, e.g. because of evolutionary reasons (it is known, for example, that human faces and outdoor scenes are recognized faster [20]). Images representing these categories may tolerate a different amount of impairment than others, thereby influencing the final quality assessment of the user.

It is important to remark here that the influence of semantic categories on visual quality should not be confused with the perception of utility or usefulness of an image [22, 23]. Image utility is defined as the image usefulness as a surrogate for its reference, and so relates with the amount of information that a user can still draw from an image despite any impairment present. The idea that image usefulness can influence image quality perception has been exploited in some work on no-reference image quality assessment such as in [14], although there are studies that argue the relationship between utility and quality perception is not straightforward [22]. Instead of looking at the usefulness of an image content, we look at users' *internal* bias toward the content category, and show in this paper the difference between the two and their respective relationship with quality perception.

In our previous research, we conducted a psychophysical experiment to verify whether the semantic content of an image (i.e., its scene and/or object content category) influences users' perception of quality [24]. Our findings suggest that this is the case. Using JPEG impaired images, we found that users are more critical of image quality for certain semantic categories than others. The semantic categories we used in our study are *indoor, outdoor natural* and *outdoor manmade* for scene categories, and *inanimate* and *animate* for object categories. In [25], we then showed initial results that adding object category features to perceptual quality features significantly improves the performance of existing no-reference image quality metrics (NR-IQMs) on two well-known image quality datasets. Based on these studies, in this work we look into improving NR-IQMs by injecting semantic content information in their computation.

In this paper, we extend our previous work to include (1) different types of impairments and (2) scene category information in NR-IQM. As a first step, we collect subjective data of image quality for a set of images showing high variance in semantic content. Having verified the validity of the collected data, we then use it as ground truth to train our semantic-aware blind image quality metric. The latter is based on the joint usage of perceptual quality features (either from Natural Scene statistics [26], or directly learned from images [27]), and semantic category features. We then show the added value of semantic information in image quality assessment, and finally propose an analysis of the interplay between semantics, visual quality and visual utility.

Our contribution through this paper can be summarized as follows.

1. We introduce a *new image quality dataset comprising a wide range of semantic categories*. In the field of image quality research, several publicly available datasets exist. However, most (if not all) of these datasets do not cover the different semantic categories extensively or uniformly. To open more possibilities of research on visual quality and semantics, we set up an image quality dataset which spans a wider and more uniform range of semantic categories than the existing datasets.
2. We show how *using scene and object information in NR-IQMs improves their performance across impairments and image quality datasets*. We perform experiments to analyze how different types of semantic category features would be beneficial to use in improving NR-IQM. We also compare the performance of adding semantic features to improve NR-IQMs on different impairments and image quality datasets.

This paper is organized as follows. In the following section, we review existing work on blind image quality assessment, creation of subjective image quality datasets, and automatic methods for categorizing images semantically. In Section 3, we introduce our new dataset, SA-IQ, detailing the data collection, reliability and analysis to prove that semantic categories do influence image quality perception. In Section 4, we describe the experiments proposing our semantic-aware objective metrics, based on the addition of semantic features to the perceptual quality ones. In addition, in Section 5, we compare the relationship of semantic categories with image utility and image quality. We conclude our paper in Section 6.

2. Related Work

2.1. No-Reference Image Quality Assessment

Blind or No-reference image quality metrics aim at predicting perceived image quality without the use of a reference image. Many algorithms have been developed to perform this task, and usually fall into one of two categories: impairment-specific or general purpose NR-IQMs. As the name suggests, impairment-specific NR-IQMs rely on prior knowledge of the type of impairment present in the test image. Targeting one type of impairment at a time, these metrics can exploit the characteristics of the particular impairment and how the HVS perceives it to design features that convey information on the strength and annoyance of such impairments. Examples of these metrics include those for assessing blockiness in images [28, 29], blur [30, 7], and ringing [31].

General purpose NR-IQMs deal with multiple impairment types, and do not rely on prior information on the type of impairment present in a test image. This of course allows for a wider applicability of the metrics, but also requires a more complex design of the quality assessment problem. To develop these metrics, usually a set of features is selected that can discriminate between different impairment types and strengths, followed by a mapping (pooling) of those features into a range of quality scores that matches human perception as closely as possible [32].

Handcrafted features are often used to develop general purpose NR-IQMs, one of the most common being natural scene statistics (NSS), although other types of features have also been proposed, such as the recent free-energy based features [33, 34]. NSS assume that pristine natural images have regular statistical properties which are disrupted when the image is impaired. Capturing this disruption can reveal the extent to which impairments are visible (and thus annoying) in the image. To do so, typically the image is transformed to a domain (e.g. DCT or wavelet), that better captures frequency or spatial changes due to impairments. The transform coefficients are then fit to a predefined distribution, and the fitting coefficients are taken as the NSS features.

Different NSS-based NR-IQMs have used various image representations to extract image statistical properties. In [26], for example, the NSS features were computed from the subband coefficients of an image's wavelet transform. Beside fitting a generalized Gaussian distribution to the subband coefficients, some correlation measures on the coefficients were also used in extracting the features. The study aimed at predicting the quality of images impaired by either JPEG or JPEG 2000 compression, white noise, Gaussian blur, or a Rayleigh fading channel. Saad et al. [35] computed NSS features with a similar procedure, but in the DCT domain. Mittal et al. [36] worked out NSS features in the spatial domain instead. They fitted a generalized Gaussian distribution on the image's normalized luminance values and their pairwise products along different orientations. In this case, the parameters of the fit were used directly as features. Another study in [37] took the Gradient Map (GM) of an image, and filtered it using Laplacian of Gaussian (LOG) filters. The GM and LOG channels of the image were then used to compute statistical features for the quality prediction task.

Lately, the IQM community has also picked up on the tendency of using learned, rather than handcrafted (e.g., NSS and free energy-based), features. A popular approach is to first learn (in an unsupervised way) a dictionary or codebook of image descriptors from a set of images. Using another set of images, the codebook will then be used as the basis for extracting features to learn a prediction model. To extract these features, an encoding step is performed on the image descriptors, followed by a pooling step. The study in [38] used this approach. The codebook was built based on normalized image patches and K-means clustering. To extract features for training and testing the model, a soft-assignment encoding was then performed, followed by max-pooling on the training and testing images. In [39], image patches underwent Gabor filtering before being used as descriptors to build the codebook. Hard assignment encoding was then performed, after which average pooling was used to extract the image features. To limit the computational burden yield by the large size of codebooks, a more recent study [27] proposed using a small sized codebook, built using K-means clustering based on normalized image patches. Smaller sized codebook usually decreases the prediction performance, and so to compensate for that, the study proposes to calculate the differences of high order statistics (mean, covariance and co-skewness) between the image patches and corresponding clusters as additional features.

Finally, the research on NR-IQMs has also recently started looking at features learned through convolutional neural networks (CNNs). CNNs [40] are multilayer neural networks which contain at least one convolutional layer. The network structure already includes parts that extract features from input images and a regression part to output a prediction for the corresponding input. The training process of this network not only optimizes the prediction model, but also the layers responsible for extracting representative features for the problem at hand. The study in [41] is an example of NR-IQMs using this approach, which brings promising results. However, one should be aware that, when learning features especially through CNNs, their interpretability is mostly lost. The high dimensionality of learnable CNN parameters also makes those features to be prone to overfitting of the training data, which

Table 1: Properties of Several Publicly Available Image Quality Datasets

Dataset	**Number of Images**	**Number of Reference Images**	**Impairment Types * Levels**	**Semantic Categories (of Reference Images)**	
				Scene	**Object**
TID2013 [42]	3000	25	24 * 5	21 Outdoors 3 Indoors 1 N/A	7 Animate 14 Inanimate 1 N/A
CSIQ [43]	900	30	6 * 5	30 Outdoors 0 Indoors	13 Animate 17 Inanimate
LIVE [44]	982	29	5 * 6-8	28 Outdoors 1 Indoor	8 Animate 20 Inanimate
MMSPG HDR with JPEG XT [45]	240	20	3 * 4	12 Outdoors 6 Indoors 2 N/A	4 Animate 14 Inanimate 2 N/A
IRCCyN-IVC on Toyama [46]	224	14	2 * 7	14 Outdoors 0 Indoors	3 Animate 11 Inanimate
UFRJ Blurred Image DS [47]	585	N/A	N/A	412 Outdoor 173 Indoors	198 Animate 387 Inanimate
ChallengeDB [48]	1163	N/A	N/A	759 Outdoor 403 Indoors	321 Animate 842 Inanimate
SA-IQ	474	79	2 * 3	39 Outdoors 40 Indoors	25 Animate 54 Inanimate

is especially a risk when the size of data is small, as in the case of Image Quality Assessment databases (see more details in sec. 2.2).

The NR-IQMs described earlier, which are based on features representing perceptual changes in an image due to the presence of impairments, have higher interpretability and can still obtain acceptable accuracy. In this paper, we aim at improving accuracy while maintaining interpretability. Therefore, we focus on this category of metrics and on enabling them to incorporate features that account for semantic content understanding.

2.2. Subjective Image Quality Datasets

Over the years, the IQM community has developed a number of datasets for metric training and benchmarking. Such datasets usually consist of a set of reference (pristine) images, and a larger set of impaired images derived from the pristine ones. Impaired images are typically obtained by injecting different types of impairments (e.g., JPEG compression or blur) at different levels of strength. Each image is then associated with a subjective quality score, usually obtained from a subjective study conducted with a number of users. Individual user judgments of Quality are averaged per image across users into Mean Opinion Scores, which represent the quantity to be predicted by Image Quality Metrics.

Most Subjective Image Quality datasets are structured to have a large variance in terms of types and level of impairments, as well as perceptual characteristics of the reference images, such as Spatial Information or Colorfulness [49]. On the other hand, richness in semantic content of the reference images is often disregarded, nor information is provided on categories of objects and scenes represented there. This limits the understanding and assessment of image quality as it excludes users' higher-level interpretation of image content in their evaluation of quality. Table 1 gives an overview of the semantic diversity covered by several well-known and publicly available image datasets. The semantic categorization follows that proposed by Li et al. in their work related to pre-attentional image content recognition [20] (note that these categories were not provided with the datasets and were manually annotated by the authors of this paper).

From the table, we can see that most datasets do not have a balanced number of scene or object categories to allow for further investigation of the relationship between semantic categories and image quality. Two datasets are quite diverse in their semantic content: the UFRJ Blurred Image dataset [47], and the Wild Image Quality Challenge dataset [48]. On the other hand, these datasets lack structured information on the impairment types and levels of impairments present in the images. The images were collected "in the wild", meaning that they were collected in typical real-world settings with a complex mixture of multiple impairments, instead of being constructed in the lab by creating well-modeled impairments on pristine reference images.

These datasets were created to simulate the way impairments typically appear in consumer images. An impaired image in these datasets thus does not correspond to any reference image, and there is no clear framework to refer to in order to obtain information about how the impairments were added to the images. This makes it difficult to systematically look into the interplay between image semantics, impairments, and perceived quality.

In this work we propose a new dataset rich in semantic content diversity. We look at 79 reference images with contents covering different object and scene categories. These images are further impaired to obtain blur and JPEG compression artifacts at different levels. The proposed dataset SA-IQ can be seen as the last entry in Table 1, and we explain details of how the dataset was constructed in Section 3.

2.3. Image Semantics Recognition

One of the most challenging problem in the field of computer vision has long been that of recognizing the semantic content of an image. A lot of effort has been put by the research community to improve image scene and object recognition performances: creating larger datasets [50], designing better features, and training more robust machines [51]. In the past five years, wider availability of high-performance computation machines and labelled data has allowed for the rise of Convolutional Neural Networks (CNNs) [40], and resulted in vast progress in the field of image semantic recognition.

One of the pioneering attempts of deploying CNNs for object recognition was AlexNet by Krizhevsky et al. [52]. Based on five convolutional and three fully connected layers, the AlexNet processes 224x224 images to map them into a 1000-dimensional vector, the elements of which represent the probability values that the input image belongs to any of a thousand predefined object categories. Since AlexNet, current state-of-the-art systems include VGG [53], and GoogleNet [54]. For a more comprehensive overview of state-of-the-art systems, readers are referred to [51].

Along with object recognition, scene recognition has also had its share of rapid development with the advent of CNNs. One recently proposed trained CNN for scene recognition is the Places-CNN [55, 56]. This CNN is trained on the Places image dataset, which contains 2.5 millions images with a scene category label. 205 scene categories are defined in this dataset. The original Places-CNN was trained using similar architecture as the Alexnet mentioned above. Further improvements of the original Places-CNN were obtained by training on the VGG and GoogleNet architectures [56].

The implementation we use in this paper is the PlacesVGG. The architecture has 13 convolutional layers, with four pooling layers among them, and a fifth pooling layer after the last convolutional layer. Three fully connected layers follow afterwards. The network outputs a 205-dimensional vector with elements representing the probability that the input image belongs to any of the 205 scene categories.

3. Semantic-Aware Image Quality (SA-IQ) Dataset

As mentioned in Section 2, most publicly available image quality datasets do not cover a wide range of semantic categories. This limitation does not allow us to look deeper into how users evaluate image quality in relation with their interpretation of the semantic content category. For this reason, we created a new image quality dataset with not only a wider range of semantic categories included in it, but also a more even distribution of these categories. We describe our proposed dataset in the following subsections.

3.1. Stimuli

We selected 79 images that were 1024x768 in size from the LabelMe image annotation dataset [57]. The images were selected such that there was a balanced number of images belonging to each of the scene categories indoor, outdoor natural, and outdoor manmade, and within each scene category, enough number of animate and inanimate objects. Animate objects include humans and animals, whereas inanimate objects include objects in nature (e.g., body of water, trees, hill, sky) and objects that are manmade (e.g., buildings, cars, roads).

To have an unbiased annotation of the image categories, we asked five users to categorize the image scenes and objects. They were shown the pristine or unimpaired version of the images, and asked to assign the image to either of the three scene categories and either of the two object categories. The images were presented one at a time, and we did not restrict the time for users to view each image. Each image was then assigned the scene and object category which had the majority vote from the five users. In the end, we have 39 indoor images, 19 outdoor natural images, and 21 outdoor man-made images. In terms of object categories, we have 25 images with animate objects and 54 with inanimate objects. Figure 1 shows examples of the images in the dataset.

Image texture and luminance analysis. A possible concern in structuring a subjective quality dataset based on semantic, rather than perceptual, properties of the reference images is that certain semantic categories could include a majority of images with specific perceptual characteristics, and be more or less prone to visual alterations caused by the presence of impairments. For example, outdoor images may have higher luminance than indoor ones, and risk incurring luminance masking of artifacts. If that were the case, outdoor images would mask impairments better, thereby resulting in higher quality than indoor one; this difference, though, would not be due to semantics.

Texture and luminance are two perceptual properties that are known to influence and possibly mask impairment visibility [58, 59]. We therefore

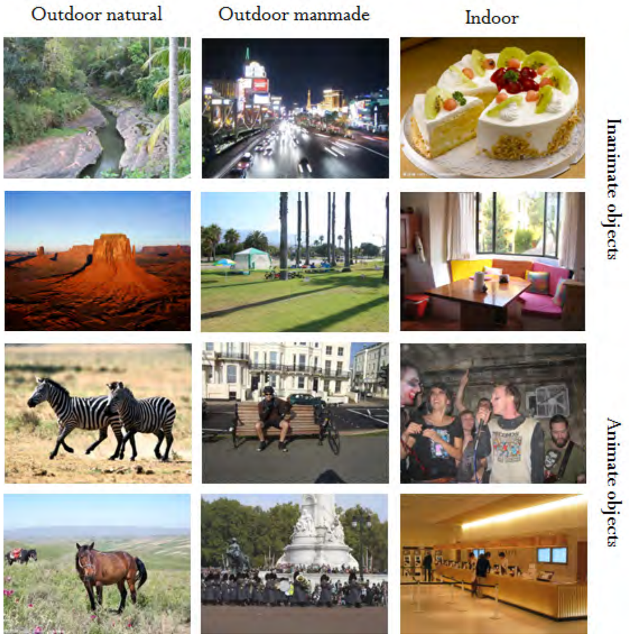


Figure 1: Examples of images in the SA-IQ dataset; the dataset contains images with indoor, outdoor natural, and outdoor manmade scenes, as well as animate and inanimate objects.

verified that the images included in the dataset had similar texture and luminance levels across semantic categories. Although this does not make our image set bias-free with respect to other possible perceptual characteristics, as luminance and texture play a major role in the visibility of artifact (and, consequently, on perceptual quality) [58, 59], this ensures that we rule out possible major effects of potential biases on our results so that we can ascribe differences in perceptual quality (in our study) to differences in semantics.

We used a modified version of Law's texture energy filter based on [28] to measure texture in horizontal and vertical directions. For each image, we computed the average mean and standard deviation of texture measures in both horizontal and vertical directions. Similarly, we used a weighted low-pass filter based on [28] to measure luminance in horizontal and vertical directions. We then calculated the average mean and standard deviation in both directions as our image luminance measure.

We compared the luminance and texture values of the images in the different scene categories using a one-way ANOVA. To compare the values across the different object categories, we used a T-Test. Our analysis showed that there is no significant difference in luminance or texture among the indoor, outdoor natural, and outdoor manmade images ($p<0.05$). Similarly, no significant difference was found for the two perceptual characteristics among the images belonging to animate or inanimate object categories ($p<0.05$). Hence, we can conclude that perceptual properties of the images are uniform across semantic categories.

Impairments. We impaired the 79 reference images with two different types of impairments, namely JPEG compression and Gaussian blur. We chose these two impairment types, as they are typically found in practical applications [60]. Moreover, most image quality assessment studies typically include these two impairment types, giving us the possibility to easily compare our results with previous studies. Of course, other types of impairments may be added in further studies. The impairments were introduced as follows.

1. *JPEG compression.* We impaired the original images through Matlab's implementation of JPEG compression. We set the image quality parameter Q to 30 and 15, to obtain images with visible artifacts of medium and high strength, respectively.
2. *Gaussian blur.* We applied Gaussian blur to the original images using Matlab's function *imgaussfilt*. To obtain images with visible artifacts of medium and high strength, the standard deviation parameter was set to 1.5 and 6, respectively. As for the choice of parameters for our JPEG compression, we also considered the parameters for our Gaussian blur to represent medium and low quality images.

Eventually, we obtained 316 impaired images. JPEG and blur images were then evaluated in two separate subjective experiments.

3.2. Subjective Quality Assessment of JPEG images

To collect subjective quality scores for the JPEG compressed images, we conducted an experiment in a laboratory setting. 80 naive participants (28 of them were females) evaluated each 60 images. The 60 images were selected randomly from the whole set of 237 images (79 reference + 158 impaired), such that no image content would be seen twice by a participant, and at the end of the test rounds, we would obtain 20 ratings for each image.

The environmental conditions (*e.g.,* lighting and viewing distance) followed those recommended by the ITU in [61]. Images were shown in full resolution on a 23" Samsung display. At the beginning of each experiment session, participants went through a short training to familiarize themselves with the task and experiment interface. Participants were then shown the test images one at a time, in a randomized order, to avoid fatigue or learning effects in the responses. There was no viewing time restriction. Participants could indicate that they were ready to score the image by clicking on a button; this would make a discrete 5-point Absolute Category Rating (ACR) quality scale appear, on which they could express their judgment of perceived quality.

3.3. Subjective Quality Assessment of Blur images

For the images impaired with Gaussian blur, we decided to conduct the experiments in a crowdsourcing environment. Crowdsourcing platforms such as AMTurk[1], Micorworkers [2] and Crowdflower [3] have become an interesting alternative environment to perform subjective tests as it is more cost and time-friendly compared with its lab counterpart. A consistent body of research has shown that crowdsourcing-based subjective testing can yield reliable results, as long as a number of precautions are taken to ensure that the scoring task is properly understood and carried out properly [62]. For example, evaluation sessions should be short (no longer than 5 minutes) and control questions (honey pots) should be included in the task to monitor the reliability of the execution.

We used Microworkers as the platform to recruit participants for our test. We randomly divided our 237 images into 5 groups consisting of 45-57 images each, such that we could set up 5 tasks/campaigns on Microworkers. Each campaign would take 10-15 minutes to complete. A user on Microworkers could only participate in one campaign out of the five, and would be paid $0.40 for completing the campaign. To avoid misunderstanding of the task, and since our experiment was presented in English, we constrained our participants to those coming from countries with fluency in English. The aim here was to prevent users from misinterpreting the task instructions, which is known to impact task performance ([23, 63, 62]). Users were directed to our test page through a link in Microworkers. We obtained 337 participations over all of the campaigns.

[1] http://mturk.com

[2] http://microworkers.com

[3] http://crowdflower.com

Protocol. When a Microworkers user chose our task, s/he was directed to our test page, and shown instructions explaining the aim of the test (to rate image quality), and how to perform evaluations. To minimize the risk of users misunderstanding their task, we were careful to provide detailed instructions and training for our users. In the first part of our training session (as recommended by [62]), we gave a definition of what we intended as impaired images in the experiment (i.e., images with blur impairments). Example images of the worst and best quality that could be expected in the experiment were provided. Afterwards, participants were asked to rate an example image to get acquainted with the rating interface. The test started afterwards. Images were shown at random order, along with the rating scale at the bottom of the screen.

We used a continuous rating scale with 5-point ACR labels in this experiment. In [64], it was shown that both discrete 5-point ACR and continuous scale with 5-point ACR labels in crowdsourcing experiments would yield results with comparable reliability. We decided to use the continuous scale in this experiment, to give users more flexibility to move the rating scale. The continuous scale range was [0..100]. In our analysis, we will normalize the resulting mean opinion scores (MOS) into the range [1..5] using a linear normalization, so that we can easily compare the results on blurred images with those on JPEG images.

To help filter out unreliable participants, we included two control questions in the middle of the experiment. For these control questions, we would show a high quality image with a rating scale below it. After the user rates that image, a control question would appear, asking the users to indicate what they saw in the last image. A set of four options were given from which the users could select an answer.

3.4. Data overview and reliability analysis

For the lab experiment on the JPEG images, we ended up with a total of 4618 ratings for the whole 237 images in the dataset after performing an outlier detection. One user was indicated as an outlier, and was thus removed for subsequent analysis. After this step, as customary, individual scores were pooled into Mean Opinion Scores (MOS) across participants per image, resulting in 237 MOS now provided along with the images.

For the crowdsourcing experiments on blurred images, we first filtered out unreliable users based on incorrect answers to the content questions in the experiment, and incomplete task executions. We also performed outlier

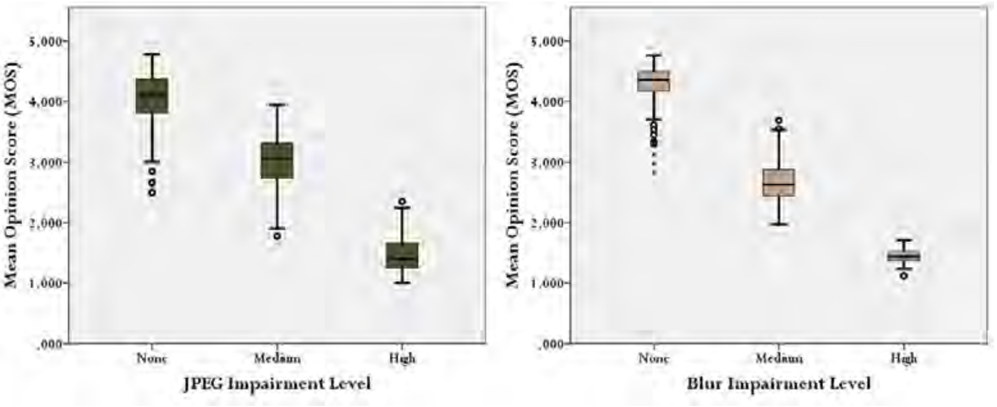


Figure 2: Overview of MOS across impairments for the two impairment types: Blur and JPEG compression

detection on the filtered data. From the 337 total responses that we received across all campaigns, we removed almost half of them due to incorrect answers to content questions, and failure to complete the whole task in one campaign. We did not find any outliers from the filtered data. In the end, we had 179 users whose responses were considered in our data analysis, with on average 37 individual scores per image. These were further pooled in MOS as described above. Figure 2 shows the collected MOS values across all impairment levels for the two impairment types: JPEG compression and Blur.

Given the diversity in the data collection method, and the concerns in terms of faithfulness of the evaluations obtained in crowdsourcing, we performed a reliability analysis. Our aim was to establish whether the obtained MOS were estimated with sufficient confidence, i.e., whether different participants expressed sufficiently similar evaluations for the same image. To do so, based on our and other previous work [25, 65], we chose the following measures to compare data reliability:

1. *SOS hypothesis alpha.* The SOS hypothesis was proposed in [66], and models the extent to which the standard deviation of opinion scores (SOS) changes with the mean opinion scores (MOS) values. This change is represented through a parameter α. A higher value of α would indicate higher disagreement among user scores. The hypothesis for an image i is expressed as in Eq. 1 below.

$$SOS^2(i) = \alpha(-MOS^2(i) + (V_1 + V_K)MOS(i) - V_1V_K), \qquad (1)$$

Table 2: SOS hypothesis alpha and average confidence interval (CI) across datasets

Dataset	**Rating Methodology**	**Number of Ratings per Image**	**Experiment Environment**	**SOS Hypothesis Alpha**	**Average CI**
SA-IQ (JPEG images)	discrete 5-point ACR scale	19-20	Lab	0.200	0.316
SA-IQ (Blur images)	continuous with 5-point ACR labels	37 on average	Crowdsourcing	0.2473	0.3182
CSIQ	multistimulus comparison by positioning a set of images on a scale	n/a	Lab	0.065	n/a
IRCCyN-IVC on Toyama	discrete 5-point ACR scale	27	Lab	0.1715	0.1680
UFRJ Blurred Image DS	continuous 5-point ACR scale	10-20	Lab	0.1680	0.5011
MMSPG HDR with JPEG XT	DSIS 1 [61], 5-grade impairment scale	24	Lab	0.201	0.273
ChallengeDB	continuous with 5-point ACR labels	175 on average	Crowdsourcing	0.1878	2.85 (100-point scale)
TID2013	tristimulus comparison	>30	Lab and Crowdsourcing	0.001	n/a

where V_1 and V_K indicate, respectively, the lowest and highest end of a rating scale.

2. *Average 95% confidence interval.* We calculate the average confidence interval over all images in a dataset to indicate user's average agreement on their ratings across the images. The confidence interval of an image i, rated by N users, is given as follows.

$$CI(i) = 1.96 * \frac{SOS(i)}{\sqrt{N}} \tag{2}$$

Table 2 gives a comparison of SOS hypothesis alpha and average CI values across different image quality studies and datasets. We also note in the table the different experiment setups used in the studies to construct the datasets. From the table, we see that the highest user agreement is obtained in studies that use comparison methods (i.e. double stimulus [61]) as their rating methodology. This was not a feasible option for us, as comparison methods on quite a large number of images as we have would be very costly. Nevertheless, our laboratory and crowdsourcing studies obtained highly comparable reliability measures. Moreover, our studies showed comparable reliability to that of other studies that also employ single stimulus rating methodology, and have the number of ratings per image proportionate to ours (i.e. the datasets IRCCyN-IVC on Toyama, UFRJ Blurred Image DS, and MMSPG HDR with JPEG XT as shown in Table 2).

3.5. Effect of Semantics on Visual Quality

Having established that our collected data is reliable, we proceeded to check how semantic categories influence visual quality ratings at different levels and types of impairments. Perception studies have looked into the relation of scene versus objects with respect to human interpretation of image content. Questions such as whether users recognize scenes or objects first when looking at images have been asked and explained. In [20], it was found that even in pre-attentive stages, users do not have the tendency to recognize scenes or objects one faster than the other. Both are processed simultaneously to form an interpretation of the image content. Here, we attempt to check if one holds more significance than the other in influencing the user assessment of image quality.

Figures 3 and 4 show bar plots of the mean opinion scores (MOS) across impairment levels and semantic categories for JPEG and blurred images, respectively. From the plots, we see that images with no perceptible impairments are rated similarly in both cases: indoor images are rated more critically than outdoor images, and images with animate objects are rated more critically than those with inanimate objects. From the figures, we see that in the case of JPEG compressed images, this tendency of being more critical towards indoor images and images with animate objects continues for images with lower quality. However, the reverse seems to happen in the case of blurred images. It seems that with the presence of blur impairments, indoor images and images with animate objects are rated higher than other scene and object categories.

To check how semantic categories influence visual quality ratings, we fit a Generalized Linear Mixed Model (GLMM) to Visual Quality ratings, where semantic categories (scene and object) and impairment levels act as fixed factors, and users are considered as a random factor. Due to the different rating scale used to evaluate the two different impairment types, the model for JPEG images uses a multinomial distribution with logit link function,

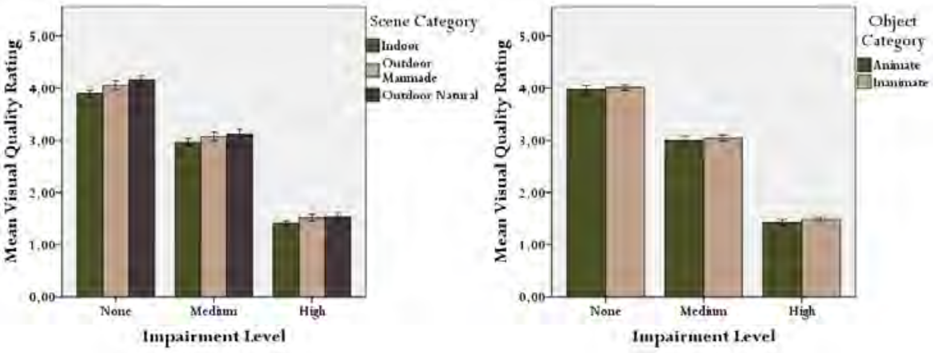


Figure 3: Bar plots of mean visual quality rating of JPEG compressed images across impairment levels and scene categories (right), and object categories (left)

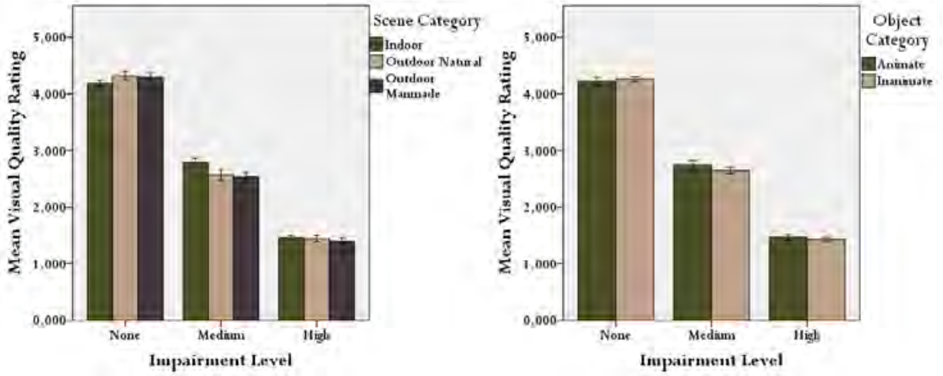


Figure 4: Bar plots of mean visual quality rating of blurred images across impairment levels and scene categories (left), and object categories (right)

while that for blurred images uses a linear distribution with an identity link function. We use the following notation to describe the output of our statistical analysis. Next to each independent variable that we looked into, we indicate the degrees of freedom (*df1, df2*), the F-statistic evaluating the goodness of the model's fit (F), and the p-value representing the probability that the variable is not relevant to the model (p). A *p-value* that is less than or equal to 0.05 indicates a statistically significant influence of a variable to predicting visual quality ratings.

For images with JPEG impairments, we find that all three independent variables, as well as the interaction of the three of them significantly influence user rating of visual quality (impairment level: *df1=2, df2=4.657,*

Table 3: Comparison of p-values for semantic category variables obtained through GLMM fitting

Impairment Type	**Independent Variables to Predict Visual Quality Ratings**	***p-value***
JPEG	Scene category	p=0.00
	Object category	p=0.00
	Scene category and object category (interaction of the two)	p=0.00
Blur	Scene category	p=0.015
	Object category	p=0.086
	Scene category and object category (interaction of the two)	p=0.00

F=1193.54, p=0.00; scene category: *df1=2, df2=4.657, F=28.35, p=0.00*; object category: *df1=1, df2=4.657, F=13.35, p=0.00*; impairment level*scene category*object category: *df1=6, df2=4.657, F=18.28, p=0.00*). This shows us that in judging images with JPEG compression impairments, users are significantly influenced by both scene and object category content.

Interestingly, for blurred images, a different conclusion is found. When we consider both scene and object categories, our model shows that scene category and impairment level has a significant effect on visual quality rating, while object category only significantly influences visual quality rating in interaction with scene category and impairment level (impairment level: *df1=2, df2=8.717, F=1880.8, p=0.00*; scene category: *df1=2, df2=8,717, F=4.18, p=0.01*; scene category*impairment level: *df1=4, df2=8.717, F=9.74, p=0.00*; impairment level*scene category*object category: *df1=6, df2=4.657, F=6.722, p=0.00*). Unlike images with JPEG compression impairments, the visual quality rating of images with blur impairments are more significantly influenced by their scene category content than their object category content. For a clear overview of the *p-values* for the different (semantic category) independent variables, a summary is given in Table 3.

4. Improving NR-IQMs using Semantic Category Features

In this section, we show how the performance of no-reference image quality metrics can significantly improve when taking semantic category information into consideration. We do this by concatenating features that represent image semantic category (as extracted, for example, by large convolutional networks trained to detect objects and scenes in images) to perceptual quality features. Figure 5 illustrates this idea. A no-reference image quality metric (NR-IQM) typically consists of two building blocks [32]. The first is a feature

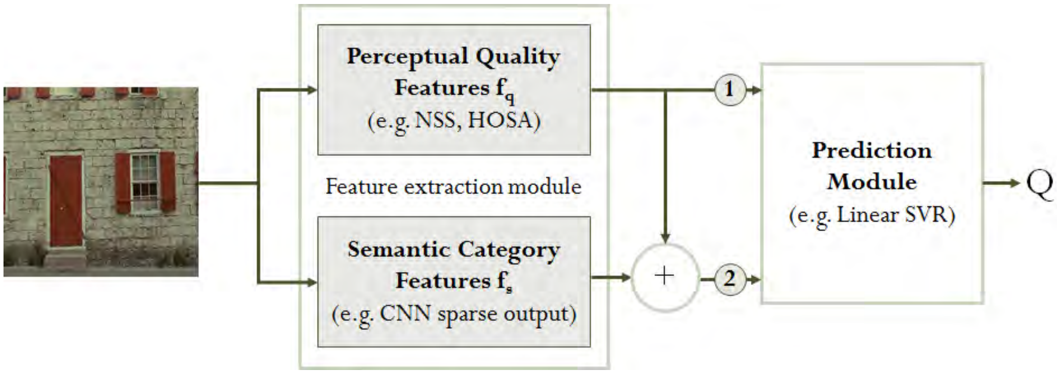

Perceptual Quality Features f_q (e.g. NSS, HOSA)
Feature extraction module
Semantic Category Features f_s (e.g. CNN sparse output)
+
1
2
Prediction Module (e.g. Linear SVR)
Q

Figure 5: Block diagram of no-reference image quality metrics (NR-IQM): (1) using only perceptual quality features, and (2) using both perceptual quality features and semantic category features

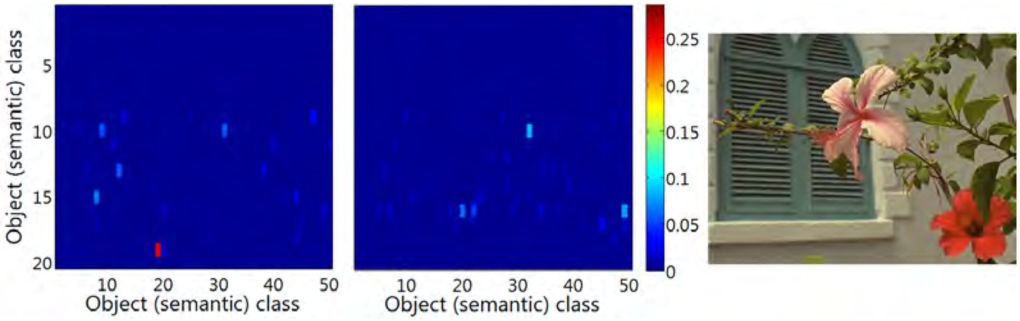


Figure 6: Heat map of probability values for the 1000 semantic classes output by AlexNet for two impaired images (with JPEG compression) taken from the TID2013 dataset, and the corresponding reference image on the right.

extraction module, which produces a set of features that represent the image, as well as any artifacts present in it. The next block is the prediction or pooling module, which translates the set of features from the previous block into a quality score Q. In the following subsections, we compare the performance of image quality prediction when using only well-known perceptual quality features (condition 1 in Figure 5), with that of using a combination of perceptual quality features and semantic category features (condition 2 in the figure).

4.1. Perceptual and Semantic Features for Prediction

We used the following perceptual quality features in our experiment:

1. *NSS features.*
   As mentioned in Section 2, NSS features are hand-crafted, and designed based on the assumption that the presence of impairments in images disrupts the regularity of an image's statistical properties. We used three different kinds of NSS features in our experiment, *BLIINDS* [35], *BRISQUE* [36], and *GM-LOG* [37]. These three metrics were chosen such that we would have NSS features extracted in different domains (DCT, spatial, and GM-LOG channels, respectively).
2. *Learned features.*
   We also chose to perform our experiment using learned features (codebook-based features). As these features are learned directly from image patches, it is possible that the features themselves already have semantic information embedded. It is therefore interesting to check how our approach would add to this type of metrics. We used HOSA features [27] to represent learned features in this paper.

To extract *semantic category features*, we fed the test images to the AlexNet [52] to obtain object category features, and to PlacesVGG [56] to obtain scene category features. We used the output of the last softmax layer of each CNN as our semantic category features. This led to a 1000-dimensional vector resulting from AlexNet, and a 205-dimensional vector resulting from PlacesVGG. Each element k in these vectors represents the probability that the corresponding image content depicts the k-th semantic category (scene or object). Each of these semantic category feature vectors would then be appended to the one containing the perceptual quality features. Adding object category features would result in an additional 1000-dimensional feature vector to the perceptual quality feature vector, while adding scene category features would result in an additional 205-dimensional feature vector to the perceptual quality features. Consequently, adding both to evaluate the benefit of considering jointly scene and object information in the IQM, would increase the feature count of 1205.

In Figure 6, we show heatmaps of the 1000-object category probability values that were output by AlexNet for two images with different levels of JPEG compression impairment. From the image, we can observe that most of the probability values of the 1000 object categories are very small. Given that quality prediction is a regression problem, we decided to use a sparse representation of these semantic feature vectors to improve on computational complexity. With a sparse representation, the number of non-zero multiplica-

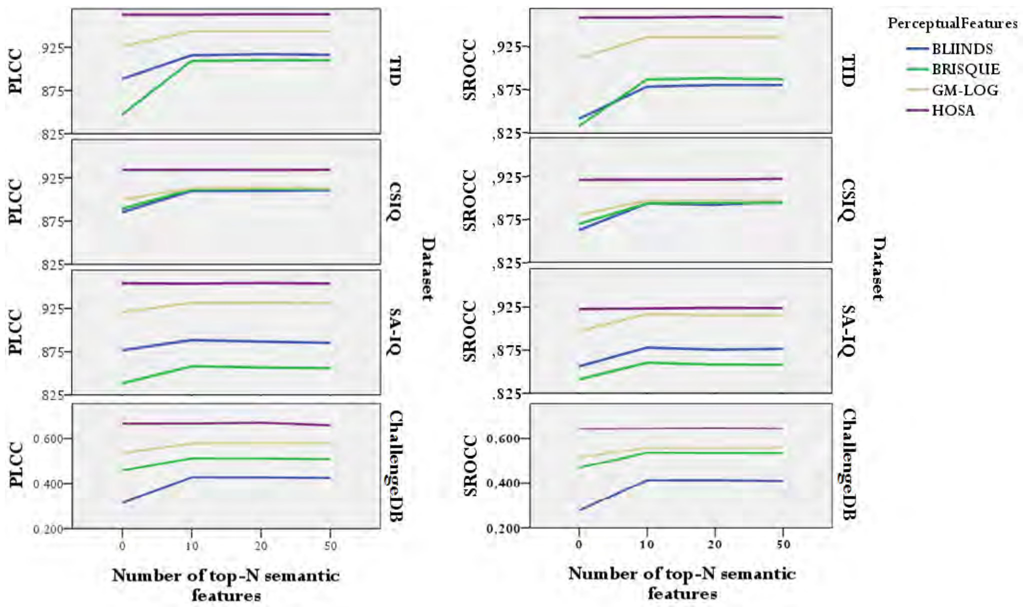


Figure 7: Impact of the number of top-N semantic categories condsidered in the IQM, in terms of Pearson and Spearman Correlation Coefficients (PLCC and SROCC respectively), between teh IQM prediction and the subjetive quality scores of different datasets. When the number of semantic features is 0, no semantic information is attached to the perceptual features, and the metric is calculated purely on perceptual feature information.

tions to be performed by our regression model is significantly smaller, thereby reducing the computation time. To make the semantic feature vector sparse, we set to zero the values of all but the top-N semantic categories in each vector.

In our previous study [25], we compared the performance of using only the top-10, 20, and 50 probability values in the object feature vector in addition to perceptual quality features. Our results showed no significant difference in performance among the three choices of N for top-N object category features. Given this result, we proceeded with using the top-20 object category features in subsequent experiments. In the next subsection, we investigate whether these results also hold for scene category features.

4.2. Augmenting NR-IQM with Semantics

To investigate the added value of using semantic category information in NR-IQM, we first compared metrics with and without using semantic information in a simplified setting. We first concatenated the sparsified semantic feature vectors with *10, 20* and *50* top-N scene semantic features to the NSS and HOSA features described in the previous section. Then, we fed the perceptual + semantic feature vector to a prediction module as depicted in Figure 5. For the sake of comparison, we also added a condition with $N=0$, corresponding to not adding semantic features. This condition represents, for this specific test, our baseline.

With reference to Figure 5, we used the same prediction module: a Support Vector Regression (SVR) with a linear kernel. This means that here we discarded the prediction modules used in the original studies proposing the perceptual quality features (i.e. BLIINDS uses a bayesian probabilistic inference module [35], BRISQUE and GM-LOG use linear SVR with RBF kernel [36, 37], white HOSA uses a linear kernel SVR [27]). This step is necessary to isolate the benefit that adding semantic information brings in terms of prediction accuracy: using different learning methods to implement the prediction module would be a confounding factor for our result here.

We performed our experiments on four datasets, TID2013, CSIQ, our own SA-IQ, and ChallengeDB. The subjective scores of these datasets were collected in different experiment setups, e.g. display resolution, impairment types and viewing distance, such that our experiment results not be limited to images viewed in one particular setting. The TID2013 dataset [42] and CSIQ dataset [43] originally contains images with 5 to 24 different types of image impairments. As most perceptual quality metrics (including those used in this paper) are constructed to evaluate images impaired with JPEG, JPEG2000 compression, additive white noise, and Gaussian blur, we limited our experiments to images with these impairments only.

The ChallengeDB dataset contains images with impairments present in the wild (typical real world consumption of images), and we included this dataset in our experiments to see how our approach would perform on said impairment condition. We used all the images in the ChallengeDB dataset in our experiment. We ran 1000-fold cross validation to train the SVR, with data partitioned into an 80%:20% training and testing set. The resulting median Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coeffient (SROCC) values between subjective and predicted quality scores are reported in Figure 7 for performance evaluation.

In our previous work [25], we observed that the addition of object category features in combination with NSS perceptual quality features (BLIINDS, BRISQUE, and GM-LOG) improved the performance of quality prediction. These improvements in both PLCC and SROCC were statistically significant

(T-test, $p < 0.05$). However, using object category features in combination with learned features (in this case, HOSA), did not bring significant added value. A similar result can be seen for the case of combining scene category features with perceptual quality features. In Figure 7, we see that for the NSS perceptual quality features, prediction performance increased with the addition of scene semantic categories. Conducting T-tests on the resulting PLCC and SROCC values showed that the improvements were statistically significant ($p < 0.05$). On the other hand, combining scene category features with HOSA features did not contribute to significant performance improvement. The average PLCC and SROCC values for the TID2013 dataset without scene features, for example, were 0.962 and 0.959, respectively, while the values when using scene features were 0.963 and 0.959, respectively.

A possible reason for the lack of improvement of the HOSA-based metric is that, unlike the handcrafted NSS features that specifically capture impairment visibility, HOSA features are learned directly from image patches. The features learned in this way may also capture semantic information, beside the impairment characteristics. Thus, the addition of semantic category features to these features may be redundant. Despite this observation, it is worth noting that the addition of semantic categories (either object or scene) could bring NSS-based models' performances close to that of HOSA while keeping the input dimensionality and thus model complexity lower (HOSA uses 14700 features, whereas NSS models use less than 100).

From the figure, we also notice that prediction performance did not change significantly among the *N=10, 20* and *50* for top-N scene features (further confirmed using one-way ANOVA, giving $p=0.05$). This applies for all four datasets and four perceptual quality metrics used in the experiment, and is aligned with our previous study on object category features [25].

4.3. Full-stack Comparison

In the previous section, we used a uniform prediction module (i.e. linear kernel SVR) across combinations of perceptual and semantic features to isolate the effect of adding semantic information on the performance of IQM. Referring once again to Figure 5, most image quality metrics in literature are optimized using a specific prediction module. For example, BLIINDS uses a Bayesian inference model, while BRISQUE and GM-LOG use SVR with an RBF kernel, and HOSA uses a linear kernel. In this subsection, we compare our approach of combining semantic category features with perceptual

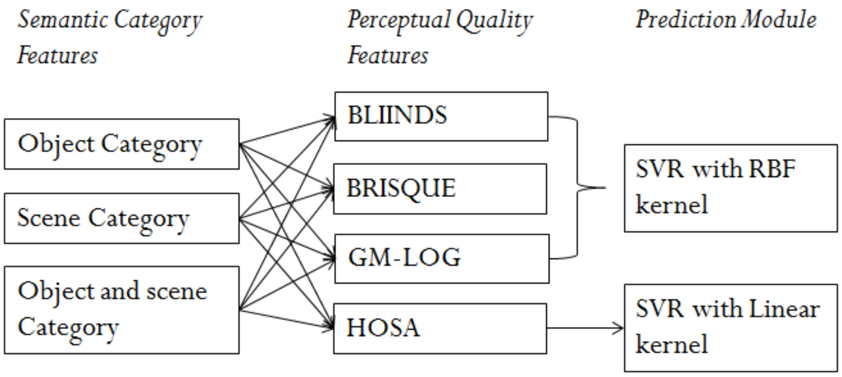


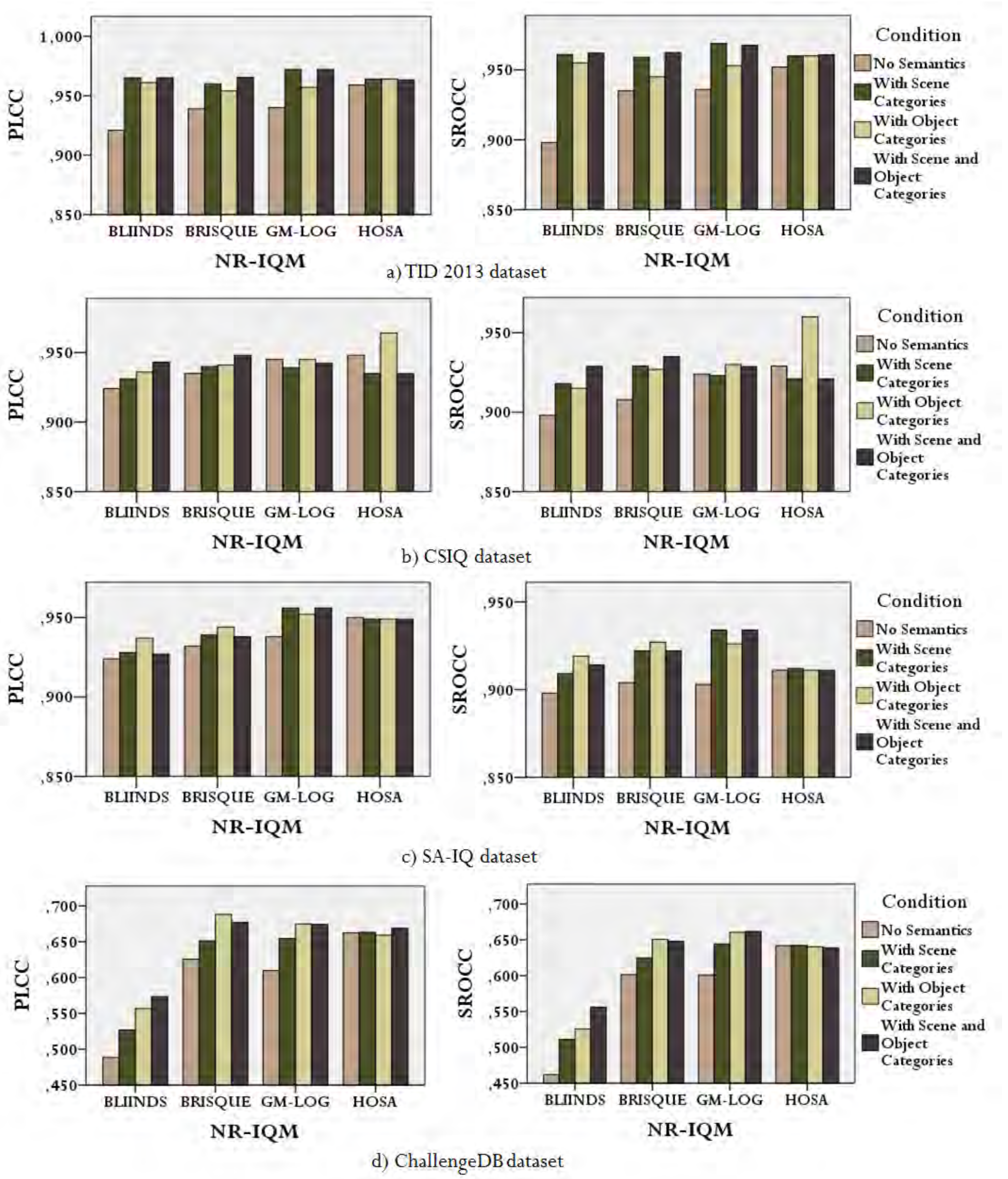
Figure 8: Features and prediction module combinations for blackbox comparison

quality features within the metrics original implementations (i.e. using their proposed perceptual quality features along with their prediction module).

Figure 8 shows the semantic category feature combinations that we used in our experiments, along with the prediction module that we use for each perceptual quality feature. There are three types of semantic category features that we looked into: object category features, scene category features, and the combination of both. Each of these were combined with each of the four perceptual quality metric, and trained using the corresponding learning method as shown in the table. We used RBF kernel SVR as learning method for the combination of semantic features with BLIINDS, BRISQUE and GM-LOG features. For the combination of semantic features with HOSA features, we used linear kernel SVR as our learning method.

As we used optimized prediction modules for each combination of features, we report here the performance of each original NR-IQM also when optimized for each dataset separately. The performance of the NSS metrics optimized for TID2013 and CSIQ that we report here are as per [37], while the HOSA metric performance optimized for the two datasets corresponds to that in [27]. For SA-IQ and ChallengeDB, we used grid search to optimize the SVR parameters of the four metrics. For performance evaluation, again we took the median PLCC and SROCC between the subjective and predicted quality scores across a 1000 folds cross-validation. Figure 9 gives an overview of the prediction performance for each feature combination on the four datasets TID2013, CSIQ, SA-IQ and ChallengeDB.

A look into the results on the TID2013 dataset reveals that the addi-



a) TID 2013 dataset

b) CSIQ dataset

c) SA-IQ dataset

d) ChallengeDB dataset

Figure 9: Full-stack comparison of the different NR-IQMs and semantic category feature combination on datasets TID2013, CSIQ, SA-IQ, and ChallengeDB

tion of semantic category features generally improved the performance of no-reference image quality assessment. As expected, based on our observation in section 4.2, the NSS-based metrics showed larger improvement in predicting quality when combined with semantic category features. Nevertheless, the combination of semantic category features with learned features (HOSA) also improved prediction performance in this case.

Results on the CSIQ dataset showed improvement particularly when the perceptual quality features were combined with object category features. If we refer back to Table 1, which gives an overview of semantic categories spanned by the different datasets used in this work, we see that the CSIQ dataset does not have any variance in scene category (all images are outdoor images), whereas there seems to be more diversity in terms of objects. We argue that this could make object category features more discriminative than scene category features.

The figure further shows results on the SA-IQ dataset. We can see that adding semantic features results in a prediction improvement compared with only using NSS features. However, as also observed in Section 4, adding semantic features did not improve prediction performance for codebook-based features (*i.e.* HOSA). Furthermore, we also note that adding scene and object category features together did not result in higher prediction performance than when using only scene or only object category features.

Similarly for the ChallengeDB dataset, we observe improvement of quality prediction with the addition of semantic category features across the three NSS-based IQMs. On the other hand, the addition of semantic category features did not improve the performance of learning-based metric, HOSA, similar to our results for the TID2013, and SA-IQ datasets.

As mentioned briefly in Section 4.2, the four datasets that we use in our experiments were constructed through subjective experiments with different experiment setups, including viewing condition and type of impairments. For example, the TID2013 study suggested users to use a viewing distance from the monitor that is comfortable to them [42], while the CSIQ study maintained a fixed viewing distance from the monitor for all its participants [43]. All the datasets use different monitors and display resolutions in their studies. And while the datasets TID2013, CSIQ, and SA-IQ have images with one impairment type per image, the ChallengeDB dataset images contain multiple impairments per image. Considering these differences across the datasets, our results here and in Section 4.2 indicates that our proposed approach to improve NR-IQMs could be applied across multiple impairments and viewing conditions.

Performance with other type of perceptual quality features. So

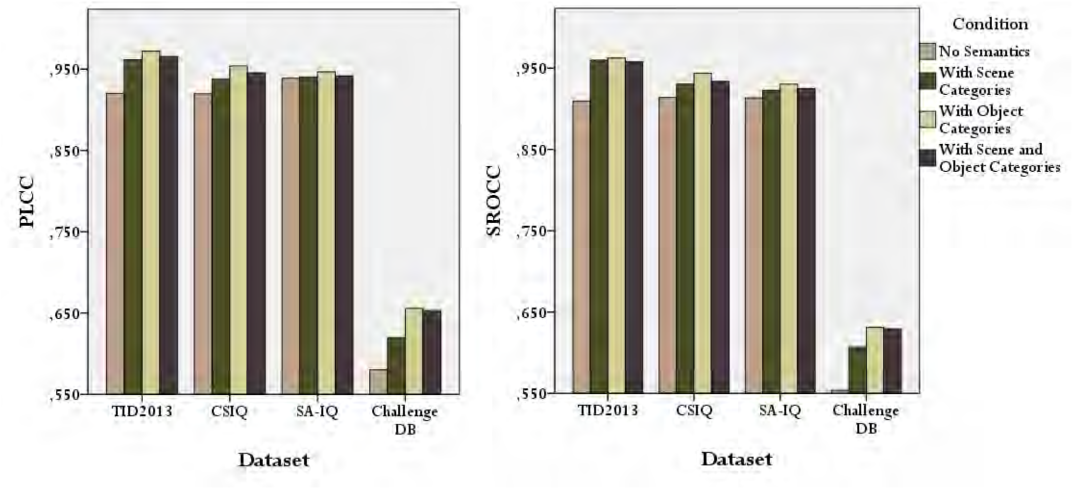


Figure 10: Full-stack comparison of the NFERM IQM and semantic category feature combination on datasets TID2013, CSIQ, SA-IQ, and ChallengeDB

far, our experiment results show that the addition of semantic category features alongside perceptual quality features can improve the performance of quality prediction, especially for NR-IQMs with handcrafted (*i.e.* NSS-based) features. We would like to show here that these results still hold for NR-IQMs based on different types of handcrafted features, such as free energy-based features ([33, 34]).

We performed a full-stack comparison using the NFERM metric on the datasets TID, CSIQ, SA-IQ and ChallengeDB. We used grid search to optimize the prediction modules for each combination of features, including when no semantic feature is used. We show our results in figure 10, which plots the median PLCC and SROCC between the subjective and predicted quality scores across 1000 folds cross-validation. The figure shows that our previous results for NSS-based NR-IQMs still hold for non NSS-based NR-IQMs such as NFERM, that is, the addition of either scene or object category features, or both, helps improve the performance of blind image quality prediction

4.4. Performance on Specific Impairment Types

In the previous experiments, we performed our evaluation on datasets consisting of different impairment types: JPEG and JPEG2000 compression, blur, and white noise in the TID2013 and CSIQ datasets, and JPEG and blur in the SA-IQ dataset. As shown through our analysis in section 3.5,

Table 4: Comparison of the different NR-IQMs and semantic category features on different impairment types in the SA-IQ dataset

		BLIINDS	**BLIINDS+S**	**BLIINDS+O**	**BRISQUE**	**BRISQUE+S**	**BRISQUE+O**
SA-IQ	**JPEG**	0.8717	**0.8941**	**0.8938**	0.885	**0.9086**	**0.9084**
	BLUR	0.8925	**0.9093**	**0.9068**	0.9029	**0.9219**	**0.9222**
TID	**JPEG**	0.8853	**0.9383**	**0.9391**	0.9103	**0.9478**	**0.9530**
	JP2K	0.9118	**0.9591**	**0.9529**	0.9044	**0.9487**	**0.9504**
	BlUR	0.9176	**0.9665**	**0.9696**	0.9059	**0.9635**	**0.9696**
	WN	0.7314	**0.9417**	**0.9409**	0.8603	**0.9524**	**0.9509**
CSIQ	**JPEG**	0.9115	0.9052	**0.9300**	0.9253	**0.9342**	**0.9292**
	JP2K	0.8870	**0.9147**	**0.9416**	0.8934	**0.9056**	**0.9262**
	BlUR	0.9152	0.9003	0.9148	0.9143	0.8781	0.9018
	WN	0.8863	**0.9248**	**0.9246**	0.9310	**0.9398**	**0.9416**

		GM-LOG	**GM-LOG+S**	**GM-LOG+O**	**HOSA**	**HOSA+S**	**HOSA+O**
SA-IQ	**JPEG**	0.8843	**0.9218**	**0.9099**	0.9149	0.9140	0.9151
	BLUR	0.9048	**0.9262**	**0.9228**	0.9029	0.9034	0.9030
TID	**JPEG**	0.9338	**0.9478**	**0.9403**	0.9283	0.9288	0.9271
	JP2K	0.9263	**0.9539**	**0.9548**	0.9453	0.9283	0.9265
	BlUR	0.8812	**0.9635**	**0.9604**	0.9538	**0.9604**	0.9562
	WN	0.9068	**0.9513**	**0.9524**	0.9215	0.9273	0.9243
CSIQ	**JPEG**	0.9328	0.8927	0.9220	0.9254	0.9062	0.9071
	JP2K	0.9172	**0.9249**	**0.9316**	0.9244	0.9032	0.9036
	BlUR	0.9070	0.8752	0.8969	0.9266	0.8848	0.9037
	WN	0.9406	0.9342	0.9237	0.9192	**0.9232**	0.9038

semantic categories influence the assessment of visual quality in both JPEG compressed and blurred images, but in a different way. It is therefore interesting to look at the prediction performance on different impairment types individually. Our setup for this experiment is similar to that of Section 4.3, i.e. the SVR parameters of the NR-IQMs were optimized for evaluating each of the three datasets. The datasets were split into subsets with specific impairment types, and the prediction models were re-trained for each impairment type. We again refer to [37] for the performance of NSS metrics optimized for TID2013 and CSIQ, and [27] for the HOSA metric performance.

Table 4 shows the results of our experiments. We report only the SROCC values due to limited space, however we note here that the resulting PLCC values yielded similar conclusions. The bold numbers in the table indicate the conditions in which the prediction performance improved with the addition of semantic category features. From the table, we see that the addition of semantic category features, whether they are scene or object features, improved significantly the performance of NSS-based no-reference metrics on all impairment types presented for the SA-IQ and TID datasets. However, for the CSIQ dataset, only images with JP2K compression and white noise impairment consistently showed similar improvement. It is interesting to note that the improvement in performance were not significantly different between the addition of object and scene categories. For the codebook-based metric, HOSA, as we have seen in the previous sections, we again observe that the addition of semantic category features did not bring improvement, even for specific impairment types, on any of the three datasets.

5. Image Utility and Semantic Categories

Image quality has often been associated with image usefulness or utility. Nevertheless, studies have shown that perceived utility does not linearly relate to perceived quality [22]. In this section, we show that bias on image content category can influence utility and perceived quality differently, and thus further confirm that an image usefulness cannot always explain perceived image quality. We do this by comparing the relationship between image semantic categories and image utility with the relationship between image semantic categories and image quality. We perform this comparison on our image dataset, SA-IQ.

To perform the comparison, we calculated image utility scores for each image in the dataset. We refer to [67] for image utility metric NICE. The metric calculates image utility based on image contour. For every image, we used an edge detection algorithm (e.g., Canny) to obtain the binary of the test image and its reference, which we denote as B_T and B_R, respectively. We then performed a morphological dilation on the two binary images using a 3x3 plus-sign shaped structural element. We further assumed that the result of this morphological dilation is I_R for the reference image and I_T for the test image. We then obtained the utility score NICE for the image by taking the Hamming distance of I_R and I_T, and dividing it by the number of non-zero elements in B_R, to account for the variability of contours across the reference images. The utility metric NICE gives an estimation of how

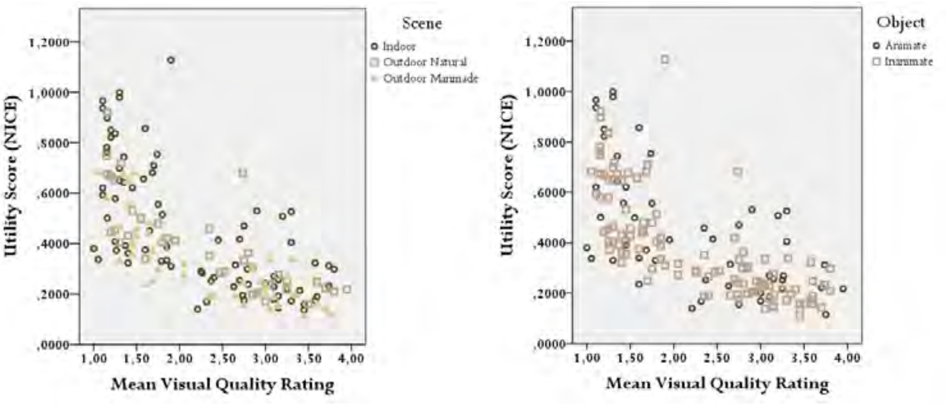


Figure 11: Image utility vs. quality scores of JPEG images across semantic categories (left: scene categories, right: object categories

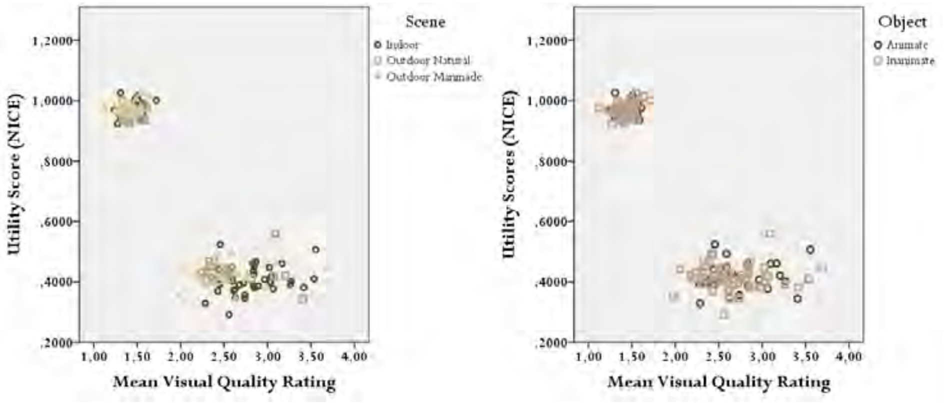


Figure 12: Image utility vs. quality scores of blurred images across semantic categories (left: scene categories, right: object categories

degraded an image's contours have become due to impairments compared with its reference, and is thus inversely related with image utility.

In Figures 11 and 12 we show plots of perceived quality mean opinion scores (MOS) against NICE utility scores for JPEG compressed and blurred images in our datasets. If we compare our plots with the perceived utility vs. perceived quality plot found in [22], we can observe that our blurred images span the lower range of image quality and higher range of image quality, in which utility doesn't grow or change with the change of perceived quality. However, our JPEG images seem to span a middle-range quality, in which perceived quality has a linear relationship with perceived utility.

Table 5: Significance level of semantic categories' influence on image utility and quality across Blurred and JPEG image clusters

Impairment Type	**Image Cluster**	**Semantics on Utility**		**Semantics on Quality**	
		Scene	**Object**	**Scene**	**Object**
BLUR	HQ cluster	p = 0.098	p = 0.971	**p = 0.009**	p = 0.324
	LQ cluster	p = 0.054	p = 0.469	p = 0.177	p = 0.228
JPEG	HQ cluster	**p = 0.03**	**p = 0.049**	p = 0.851	p = 0.866
	LQ cluster	**p = 0.003**	p = 0.219	p = 0.307	p = 0.365

In general, we can see that our data represented the different relationships between perceived quality and utility across the range of quality.

We ran K-means on the blurred and JPEG image data, to isolate the different clusters as shown in the plots, and conducted statistical analysis to check how semantic categories influence utility and quality in these clusters. We set the number of clusters k to two for both the blurred and JPEG data. We then performed several one-way ANOVA for each cluster. Specifically, we first conducted one-way ANOVAs with semantic categories (either scene or object categories) as independent variables, and utility as dependent variables. Similarly, we then conducted one-way ANOVAS with quality MOS as dependent variables instead of utility.

Table 5 shows the results of our analysis. We label the two clusters for each image sets as HQ for clusters with images having higher quality range, and LQ for clusters with images having lower quality range. The numbers in bold indicate cases in which semantics has a significant influence on either utility or quality. From the table, we can see that semantic categories influence image utility and quality differently. Moreover, the influence of semantics on utility seems to be more significant in JPEG images than in blurred images.

6. Conclusion

In this paper, we showed that an image's semantic category information can be used to improve its quality prediction to align better with human perception. Through subjective experiments, we first observed that an image's scene and object categories influence users' judgment of visual quality for JPEG compressed and blurred images. We then performed experiments on different types of no-reference image quality metrics (NR-IQMs), and showed that blind/no-reference image quality predictions can be improved by incorporating semantic category features into our prediction model. This applied across different image quality datasets representing diverse viewing condition (e.g. display resolution, viewing distance), and image impairments, including multiple impairments. We also provided a comparison of how semantics influences image utility and image quality, and conclude that semantics has more significant influence on image utility for JPEG images than for blurred images.

Another contribution of this paper is a new image quality dataset, SA-IQ, consisting of images spanning a wide range of scene and object categories, with subjective scores on JPEG compressed and blurred images. The dataset can be accessed through `http://ii.tudelft.nl/iqlab/SA-IQ.html`. Future work on these findings would include looking into better representations or methods to combine semantic information and perceptual quality features in NR-IQMs.

7. Acknowledgement

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

8. References

[1] Conviva, Viewer experience report (2015).
URL: `http://www.conviva.com/conviva-viewer-experience-report/vxr-2015/`

[2] P. Le Callet, S. Möller, A. Perkis, 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) 3 (2012).

[3] S. S. Hemami, A. R. Reibman, No-reference image and video quality estimation: applications and human-motivated design, Signal Processing: Image Communication 25 (7) (2010) 469–481 (2010).

[4] W. Lin, C.-C. J. Kuo, Perceptual visual quality metrics: a survey, Journal of Visual Communication and Image Representation 22 (4) (2011) 297–312 (2011).

[5] J. Xue, C. W. Chen, Mobile video perception: New insights and adaptation strategies, IEEE Journal of Selected Topics in Signal Processing 8 (3) (2014) 390–401 (2014).

[6] Y. Zhu, A. Hanjalic, J. A. Redi, QoE prediction for enriched assessment of individual video viewing experience, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 801–810.

[7] K. Gu, G. Zhai, W. Lin, X. Yang, W. Zhang, No-reference image sharpness assessment in autoregressive parameter space, IEEE Transactions on Image Processing 24 (10) (2015) 3218–3231 (2015).

[8] S. C. Guntuku, J. T. Zhou, S. Roy, W. Lin, I. W. Tsang, Understanding deep representations learned in modeling users likes, IEEE Transactions on Image Processing 25 (8) (2016) 3762–3774 (2016).

[9] U. Engelke, R. Pepion, P. L. Callet, H.-J. Zepernick, Linking distortion perception and visual saliency in H. 264/AVC coded video containing packet loss, in: Visual Communications and Image Processing, SPIE, 2010.

[10] H. Alers, J. Redi, H. Liu, I. Heynderickx, Studying the effect of optimizing image quality in salient regions at the expense of background content, Journal of Electronic Imaging 22 (4) (2013).

[11] W. Zhang, A. Borji, Z. Wang, P. Le Callet, H. Liu, The application of visual saliency models in objective image quality assessment: A statistical evaluation, IEEE transactions on neural networks and learning systems 27 (6) (2016) 1266–1278 (2016).

[12] K. Gu, L. Li, H. Lu, X. Min, W. Lin, A fast reliable image quality predictor by fusing micro-and macro-structures, IEEE Transactions on Industrial Electronics 64 (5) (2017) 3903–3912 (2017).

[13] D. Temel, G. AlRegib, Resift: Reliability-weighted sift-based image quality assessment, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 2047–2051.

[14] P. Zhang, W. Zhou, L. Wu, H. Li, Som: Semantic obviousness metric for image quality assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2394–2402.

[15] D. Marr, Vision: A computational approach (1982).

[16] S. Edelman, S. Dickinson, A. Leonardis, B. Schiele, M. Tarr, On what it means to see, and what we can do about it, Object Categorization: Computer and Human Vision Perspectives (2009) 69–86 (2009).

[17] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cognitive psychology 8 (3) (1976) 382–439 (1976).

[18] A. Rorissa, H. Iyer, Theories of cognition and image categorization: What category labels reveal about basic level theory, Journal of the American Society for Information Science and Technology 59 (9) (2008) 1383–1392 (2008).

[19] I. Biederman, R. C. Teitelbaum, R. J. Mezzanotte, Scene perception: a failure to find a benefit from prior expectancy or familiarity., Journal of Experimental Psychology: Learning, Memory, and Cognition 9 (3) (1983) 411 (1983).

[20] L. Fei-Fei, A. Iyer, C. Koch, P. Perona, What do we perceive in a glance of a real-world scene?, Journal of vision 7 (1) (2007) 10–10 (2007).

[21] A. Torralba, K. P. Murphy, W. T. Freeman, Using the forest to see the trees: exploiting context for visual object detection and localization, Communications of the ACM 53 (3) (2010) 107–114 (2010).

[22] D. M. Rouse, R. Pepion, S. S. Hemami, P. Le Callet, Image utility assessment and a relationship with image quality assessment, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2009, pp. 724010–724010.

[23] H. Ridder, S. Endrikhovski, 33.1: Invited paper: image quality is fun: reflections on fidelity, usefulness and naturalness, in: SID Symposium Digest of Technical Papers, Vol. 33, Wiley Online Library, 2002, pp. 986–989.

[24] E. Siahaan, A. Hanjalic, J. A. Redi, Does visual quality depend on semantics? A study on the relationship between impairment annoyance and image semantics at early attentive stages, Electronic Imaging 2016 (16) (2016) 1–9 (2016).

[25] E. Siahaan, A. Hanjalic, J. A. Redi, Augmenting blind image quality assessment using image semantics, in: 2016 IEEE International Symposium on Multimedia (ISM), IEEE, 2016, pp. 307–312.

[26] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: from natural scene statistics to perceptual quality, IEEE transactions on Image Processing 20 (12) (2011) 3350–3364 (2011).

[27] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, D. Doermann, Blind image quality assessment based on high order statistics aggregation, IEEE Transactions on Image Processing 25 (9) (2016) 4444–4457 (2016).

[28] H. Liu, I. Heynderickx, A perceptually relevant no-reference blockiness metric based on local image characteristics, EURASIP Journal on Advances in Signal Processing 2009 (1) (2009) 263540 (2009).

[29] S. Ryu, K. Sohn, Blind blockiness measure based on marginal distribution of wavelet coefficient and saliency, in: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 1874–1878.

[30] P. V. Vu, D. M. Chandler, A fast wavelet-based algorithm for global and local image sharpness estimation, IEEE Signal Processing Letters 19 (7) (2012) 423–426 (2012).

[31] H. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing artifacts in images, IEEE Transactions on Circuits and Systems for Video Technology 20 (4) (2010) 529–539 (2010).

[32] P. Gastaldo, R. Zunino, J. Redi, Supporting visual quality assessment with machine learning, EURASIP Journal on Image and Video Processing 2013 (1) (2013) 1–15 (2013).

[33] K. Gu, G. Zhai, X. Yang, W. Zhang, Using free energy principle for blind image quality assessment, IEEE Transactions on Multimedia 17 (1) (2015) 50–63 (2015).

[34] K. Gu, J. Zhou, J. Qiao, G. Zhai, W. Lin, A. Bovik, No-reference quality assessment of screen content pictures, IEEE Transactions on Image Processing (2017).

[35] M. A. Saad, A. C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Transactions on Image Processing 21 (8) (2012) 3339–3352 (2012).

[36] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708 (2012).

[37] W. Xue, X. Mou, L. Zhang, A. C. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Transactions on Image Processing 23 (11) (2014) 4850–4862 (2014).

[38] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1098–1105.

[39] P. Ye, D. Doermann, No-reference image quality assessment using visual codebooks, IEEE Transactions on Image Processing 21 (7) (2012) 3129–3138 (2012).

[40] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (10) (1995).

[41] L. Kang, P. Ye, Y. Li, D. Doermann, 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.

[42] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database tid2013: peculiarities, results and perspectives, Signal Processing: Image Communication 30 (2015) 57–77 (2015).

[43] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (1) (2010).

[44] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on image processing 15 (11) (2006) 3440–3451 (2006).

[45] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, T. Ebrahimi, Subjective quality assessment database of hdr images compressed with jpeg xt, in: Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on, IEEE, 2015, pp. 1–6.

[46] S. Tourancheau, F. Autrusseau, Z. P. Sazzad, Y. Horita, Impact of subjective dataset on the performance of image quality metrics, in: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, IEEE, 2008, pp. 365–368.

[47] A. Ciancio, A. L. N. T. da Costa, E. A. da Silva, A. Said, R. Samadani, P. Obrador, No-reference blur assessment of digital pictures based on multifeature classifiers, IEEE Transactions on image processing 20 (1) (2011) 64–75 (2011).

[48] D. Ghadiyaram, A. C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Transactions on Image Processing 25 (1) (2016) 372–387 (2016).

[49] S. Winkler, Analysis of public image and video databases for quality assessment, IEEE Journal of Selected Topics in Signal Processing 6 (6) (2012) 616–625 (2012).

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252 (2015).

[52] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[53] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[55] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in neural information processing systems, 2014, pp. 487–495.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[57] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, Labelme: a database and web-based tool for image annotation, International Journal of Computer Vision 77 (1-3) (2008) 157–173 (2008).

[58] T. N. Pappas, R. J. Safranek, J. Chen, Perceptual criteria for image quality evaluation, Handbook of image and video processing (2000) 669–684 (2000).

[59] C.-H. Chou, Y.-C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile, IEEE Transactions on circuits and systems for video technology 5 (6) (1995) 467–476 (1995).

[60] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, Tid2008-a database for evaluation of full-reference visual quality assessment metrics, Advances of Modern Radioelectronics 10 (4) (2009) 30–45 (2009).

[61] I. REC, Bt. 500-12, Methodology for the subjective assessment of the quality of television pictures (2009).

[62] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, P. Tran-Gia, Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing, IEEE Transactions on Multimedia 16 (2) (2014) 541–558 (2014).

[63] B. L. Jones, P. R. McManus, Graphic scaling of qualitative terms, SMPTE journal 95 (11) (1986) 1166–1171 (1986).

[64] E. Siahaan, A. Hanjalic, J. Redi, A reliable methodology to collect ground truth data of image aesthetic appeal, IEEE Transactions on Multimedia 18 (7) (2016) 1338–1350 (2016).

[65] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, A. Raake, Study of rating scales for subjective quality assessment of high-definition video, IEEE Transactions on Broadcasting 57 (1) (2011) 1–14 (2011).

[66] T. Hoβfeld, R. Schatz, S. Egger, Sos: The mos is not enough!, in: Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on, IEEE, 2011, pp. 131–136.

[67] D. M. Rouse, S. S. Hemami, R. Pépion, P. Le Callet, Estimating the usefulness of distorted natural images using an image contour degradation measure, JOSA A 28 (2) (2011) 157–188 (2011).