Group MAD Competition − A New Methodology to Compare Objective Image Quality Models

Kede Ma[1,4], Qingbo Wu[1,2], Zhou Wang[1], Zhengfang Duanmu[1]
Hongwei Yong[3,4], Hongliang Li[2] and Lei Zhang[4]
[1]University of Waterloo, [2]University of Electronic Science and Technology of China,
[3]Xi'an Jiaotong University, [4] The Hong Kong Polytechnic University

Abstract

Objective image quality assessment (IQA) models aim to automatically predict human visual perception of image quality and are of fundamental importance in the field of image processing and computer vision. With an increasing number of IQA models proposed, how to fairly compare their performance becomes a major challenge due to the enormous size of image space and the limited resource for subjective testing. The standard approach in literature is to compute several correlation metrics between subjective mean opinion scores (MOSs) and objective model predictions on several well-known subject-rated databases that contain distorted images generated from a few dozens of source images, which however provide an extremely limited representation of real-world images. Moreover, most IQA models developed on these databases often involve machine learning and/or manual parameter tuning steps to boost their performance, and thus their generalization capabilities are questionable. Here we propose a novel methodology to compare IQA models. We first build a database that contains 4,744 source natural images, together with 94,880 distorted images created from them. We then propose a new mechanism, namely group MAximum Differentiation (gMAD) competition, which automatically selects subsets of image pairs from the database that provide the strongest test to let the IQA models compete with each other. Subjective testing on the selected subsets reveals the relative performance of the IQA models and provides useful insights on potential ways to improve them. We report the gMAD competition results between 16 well-known IQA models, but the framework is extendable, allowing future IQA models to be added into the competition.

1. Introduction

Digital images undergo many transformations in their lifetime during acquisition, processing, compression, storage, transmission and reproduction. Any transformation may introduce distortions that result in degradations in visual quality [24, 30]. Being able to assess image quality is of fundamental importance in many image processing and computer vision applications. Since the human visual system (HVS) is the ultimate receiver in most applications, subjective evaluation is the most reliable way of quantifying image quality but is time-consuming, cumbersome and expensive. In recent years, there has been a rapidly growing interest in developing objective image quality assessment (IQA) models that can automate the process [30, 31]. Depending on the availability of a distortion-free reference image, objective IQA models may be categorized into full-reference (FR), reduced-reference (RR) and no-reference (NR) approaches, where the reference image is fully, partially, and completely not accessible.

With a significant number of IQA models proposed recently, how to fairly compare their performance becomes a challenge. The standard approach in the literature is to first build databases of images with various content and distortions, and then collect subjective evaluation scores for all images. Widely recognized image databases with subjective ratings include LIVE [27], TID2008 [23], TID2013 [22], CSIQ [12], IVC [13], Toyama-MICT [9] and VCL@FER [41]. Crowdsourcing techniques [2, 15] were also adopted to construct subjective databases of real world Internet images [4]. Given these databases, correlations between subjective mean opinion scores (MOSs) and objective model predictions can then be computed. Higher correlations suggest better model performance.

A major problem with this conventional evaluation methodology is the conflict between the enormous size of the image space and the limited scale of affordable subjective testing. Subjective testing is expensive and time-consuming. As a result, a typical "large-scale" subjective test allows for a maximum of several hundreds or a few thousands of test images to be rated. Given the combination of source images, distortion types and distortion levels, realistically only a few dozens of source images (if not fewer) can be included, which is the case in all well-known databases. Moreover, many source test images are repeated in the current databases, and the distortion types being used are also similar. By contrast, digital images live in an extremely high dimensional space, where the dimension equals the number of pixels, which is typically in the order of hundreds of thousands or millions. Therefore, a few thousands of samples that can be evaluated in a typical subjective test are deemed to be extremely sparsely distributed in the space. Furthermore, it is difficult to justify how a few dozens of source images can provide a sufficient representation of the variations of real-world image content. It is also worth noting that most state-of-the-art IQA models were developed after the above-mentioned image databases became publicly available. These models often involve machine learning or manual parameter tuning steps to boost their performance on these databases. In particular, recent IQA models based on sophisticated machine learning approaches employ a very large number of image features together with large-scale learning networks to improve quality prediction performance. It is thus questionable if the reported highly competitive performance of recent IQA models can be generalized to the real-world, where images have much richer content and undergo a much broader variation of quality degradations.

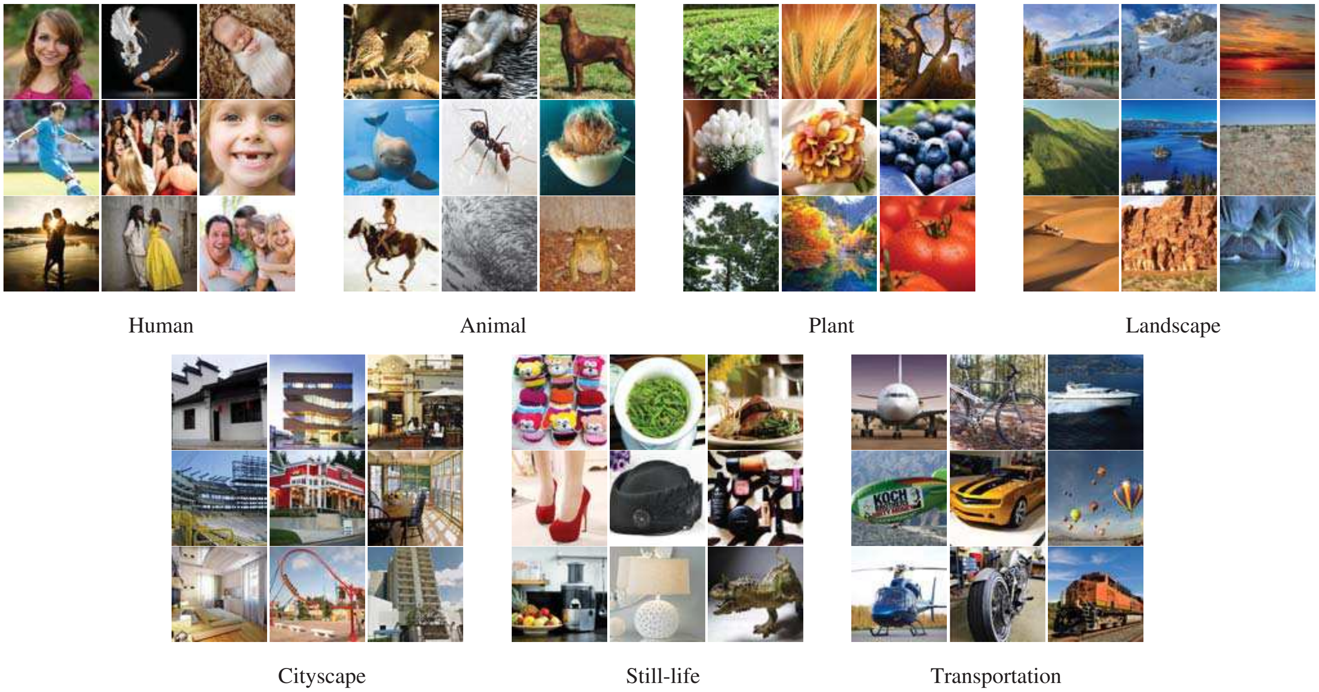
We believe that to provide a fair comparison of IQA models and to test their generalization capability, a much larger test database (e.g., thousands of source images and tens of thousands distorted images) must be used. Apparently, the main difficulty here is how to make use of such a database to compare IQA models under the constraint of very limited resource for subjective testing, knowing that rating all test images by human subjects is impossible.

In this paper, we propose a substantially different methodology to address the problem. We first build a database that contains 4,744 source natural images, together with 94,880 distorted images created from them. Assuming a group of IQA models are available for testing, we propose a novel mechanism, namely group MAximum Differentiation (gMAD) competition, that automatically selects subsets of image pairs from the database that provide the strongest test to let the IQA models compete with each other. The key idea behind gMAD is to minimize the number of required subjective tests in order to most efficiently falsify a "defender" model by selecting and testing on image pairs that maximally differentiate the *defender* model using multiple "attacker" models. In other words, instead of trying to prove a model using a set of pre-defined images from subject-rated databases, we attempt to disprove the model in the most efficient way using a small set of deliberately selected, model-dependent image pairs. The process is applied to every IQA model in the group as the "defender" model. Subjective testing on the selected subsets of test image pairs reveals the relative strengths and weaknesses of the IQA models and also provides useful insights on potential ways to improve them. This work is inspired by the idea behind the MAD competition approach [33], but unlike gMAD, the MAD method includes only two models in the competition, relies on gradient computations of the models, and is not structured to explore a database of images.

2. Related Work

Well-known subject-rated image databases include LIVE [27], TID2008 [23], CSIQ [12], IVC [13], Toyama-MICT [9], VCL@FER [41] and TID2013 [22]. They have been extensively employed in the training and testing processes in the development and benchmarking of a majority of state-of-the-art IQA models. The specific subjective testing methodologies vary, but eventually each image in the databases is labeled with an MOS, which represents the average subjective opinion about the quality of the image and is often referred to as the "ground truth" quality score of the image. The most common distortion types shared by these databases are JPEG compression, JPEG2000 compression, white Gaussian noise contamination and Gaussian blur. The typical size of these databases is in the order of hundreds or a few thousands images. Among them, the TID2013 database [22] is the largest and contains 25 source and 3000 distorted images in total. By contrast, the current database we created in this work contains 190 times more source images and 30 times more distorted images, respectively.

Depending on how the test images are presented to human subjects and how human subjects are instructed to rate the images, subjective testing may be carried out in three ways: 1) single-stimulus method, where one test image is shown at any time instance and the subjects directly give quality scores to the image; 2) paired comparison method (also known as two-alternative forced choice, or 2AFC approach), where a pair of images are shown to the subjects, who are instructed to choose a preferred image from the two; and 3) multiple-stimulus method, where multiple images are shown simultaneously and the subjects rank all images based on their quality or give quality scores to all images. Assume that there are n test images in total. $O(n)$ evaluations are needed in single-stimulus and multiple-stimulus methods, while $O(n^2)$ evaluations are needed in a full paired comparison experiment. Although paired comparison method is often preferred to collect reliable subjective evaluations, exhaustive paired comparison requires a very large number of evaluations, which are often impractical when the total number of test images is large. A number of approaches have been proposed to improve the efficiency. Four types of balanced sub-set designs were developed in the 1950's [1], among which the square design method became popular and was later further improved [14]. Another method was to randomly select a small subset of pairs for each subject [3], and it was



Human

Animal

Plant

Landscape

Cityscape

Still-life

Transportation

Figure 1. Sample source images in the new image database.

shown that $O(n \log n)$ distinct pairs are needed for large random graphs to guarantee graph connectivity and thus to achieve any global ranking using HodgeRank [10]. In the construction of the TID2013 [22] database, a Swiss competition principle was adopted to decrease the evaluations to $O(n \log n)$. Recently, an active sampling strategy for subjective quality assessment was proposed [39] with a complexity of $O(n)$. Different from all the above strategies to improve testing efficiency, gMAD requires a fixed number of paired comparisons and thus does not scale with the number of images in the database. This feature allows it to exploit large-scale databases with low and manageable cost.

3. Image Database Construction

We construct a new image database, which currently contains 4,744 high quality source natural images with a great amount of image content. An important consideration in selecting the images is that they need to be representative of the images we see in real-world applications. Therefore, we resort to the Internet, and more specifically, we elaborately select 196 keywords to search for images via Google Images [5]. The keywords can be broadly classified into 7 categories: human, animal, plant, landscape, cityscape, still-life and transportation. As a result, we initially obtain more than $200,000$ images. Many of these images contain significant distortions or inappropriate content, and thus a sophisticated manual process is applied to refine the selection. In particular, we first delete those images that have obvious distortions, including heavy compression artifacts, strong motion blur or out of focus blur, low contrast, underexposure or overexposure, substantial sensor noise, visible watermarks, artificial image borders, and other distortions due to improper operations during acquisition. Next, images of too small and too large sizes, cartoon and computer generated content, and inappropriate content are excluded. After this step, about $7,000$ images are left in the database. To make sure that the images have pristine or nearly pristine quality, we further carefully investigate each of the remaining images multiple times by zooming in and delete those images with visible compression distortions. Eventually, we end up with 4744 high quality natural images in our database. Sample images are shown in Fig. 1.

Four distortion types, namely JPEG and JPEG2000 compression, white Gaussian noise and Gaussian blur, each with five distortion levels are used to generate $94,880$ distorted images. The four distortion types are common in existing IQA databases [27, 23] and many IQA models are claimed to be able to properly handle these distortions [19, 20, 25, 17, 40, 18, 38, 37, 26, 42, 36, 7, 35]. The distortion generation process follows the method in [27], and the parameters that control the distortion levels for each type are optimized in order to uniformly cover the subjective quality scale. Once determined, the parameters are kept unchanged for all images.

Overall, our new image database contains a total of $99,624$ images which is the largest image database so far in the IQA research community. It will be made publicly available.

4. gMAD Competition

The underlying principle in traditional approaches of IQA model evaluation is to *prove* a model. This requires the model to be validated using a sufficient number of test samples in the applicable space of the model. Applying such a principle in IQA model evaluation is a major challenge because the applicable space (i.e., the space of all possible images) is enormous (millions of dimensions), but the total number of test samples (subject-rated images) that can be obtained in a realistic subjective experiment is only in the order of thousands (if not fewer). It is extremely difficult to justify that these test samples are sufficient to represent the population of real-world images.

The most fundamental idea behind the MAD [33] and the current gMAD competition approaches is to give up the traditional principle. Instead of trying to *prove* a model, here we attempt to *disprove* a model, and a model that is more difficult to be disproved is regarded as a relatively better model. This new principle gives us an opportunity to largely reduce the required number of test samples because ideally even one "counter-example" is sufficient to disprove a model. Another important ingredient in the gMAD approach is to use an efficient and automatic way to find potential "counter-examples". When attempting to disprove a model (denoted as the "defender"), instead of trying to hand design or manually search for the best counter-examples, gMAD makes use of a group of other models (denoted as the "attackers") to search for the counter-examples in the database that are optimal with regard to the *attacker* models such that if the attack is successful, the *defender* model is simply disproved. If instead, the *defender* survives from such an attack, it is a strong indicator that it is likely to be a robust and reliable model. gMAD runs this game using all available models with all possible combinations of *defender-attacker* roles of the models before performing overall statistics that help summarize the relative performance of the competing models.

The details of the gMAD competition procedure are as follows: We are given a database $\mathbf{D}$ that contains N images with different distortion types and levels. Also given are a group of M objective IQA models.

- **Step 1**. Apply all M IQA models to all N images in $\mathbf{D}$. This results in a score matrix $\mathbf{S}$ of M rows and N columns, where each entry is the quality score given by one specific IQA model to one specific image;

- **Step 2**. Choose the first model as the *defender* by setting $i = 1$. The rest of the $M - 1$ models are the *attackers*;

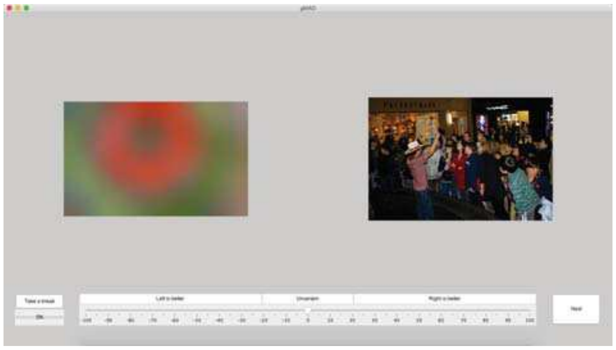


Figure 2. User interface for subjective testing.

- **Step 3**. Choose the first quality level $k = 1$ from a total of K quality levels, where $k \in \{1, 2, \cdots, K\}$;

- **Step 4**. At the i-th row in $\mathbf{S}$, find all images at quality level k (based on the *defender* model i). This results in a subset of images $\mathbf{D}_{ik}$, where all images have the same or similar quality scores according to the *defender* model i;

- **Step 5**. Choose one model j from the *attacker* models ($j \neq i$).

- **Step 6**. Within $\mathbf{D}_{ik}$, find a pair of images I^l_{ijk} and I^u_{ijk} that correspond to the minimal and maximal quality scores on the j-th row of matrix $\mathbf{S}$, respectively. This image pair is referred to as the MAD counterexample suggested by model j to attack model i at quality level k;

- **Step 7**. Carry out a subjective quality discriminative test on I^l_{ijk} and I^u_{ijk} (details given in Section 5.1);

- **Step 8**. Choose another *attacker* model j and repeat **Steps 6-7** until all *attacker* models are exhausted;

- **Step 9**. Choose the next quality level by setting $k = k + 1$ and repeat **Steps 4-8** until $k = K$ (all quality levels are exhausted);

- **Step 10**. Choose the next *defender* model by setting $i = i + 1$ and repeat **Steps 3-9** until $i = M$ (all IQA models are exhausted);

- **Step 11**. Carry out statistical analysis on the subjective quality discriminative test results (details given in Section 5.2).

Several useful features of the gMAD competition method are worth mentioning here. First, the process does not depend on the specific image database being explored.

The same approach can be applied to any collection of images of any content and distortion types. Second, the number of image pairs selected by gMAD for subjective testing is $M(M-1)K$, which is independent of the size N of the image database $\mathbf{D}$. As a result, applying gMAD competition to a larger database has no impact on the cost of subjective testing. Third, each selected pair of images are associated with two IQA models, which hold highly different opinions on their perceived image quality; one believes the pair have the same quality while the other suggests that they have very different quality. If the pair are easily differentiated by human subjects, they constitute strong evidence against the *defender* model. On the other hand, if the pair indeed have similar perceived quality, they provide strong evidence to support the *defender* model against the *attacker* model. Fourth, it is easy and cost-effective to add new IQA models into the competition. No change is necessary on all the selected pairs and their corresponding subjective testing. The only additional work is to select a total of $2MK$ new image pairs for subjective testing, half of which are for the case that the new model acts as a *defender* and the other half as an *attacker*. A MATLAB program will be made publicly available to facilitate the future usage of the gMAD competition approach.

5. IQA Models Comparison

5.1. Subjective Testing

The construction of the test image database has been described in detail in Section 3. A total of sixteen IQA models are selected in the gMAD competition process to cover a wide variety of IQA methodologies with an emphasis on NR models. These include FR models 1) PSNR, 2) SSIM [32], 3) MS-SSIM [34], 4) FSIM [43] and NR models 5) BIQI [19], 6) BLINDS_II [25], 7) BRISQUE [17], 8) CORNIA [40], 9) DIIVINE [20], 10) IL-NIQE [42], 11) LPSI [36], 12) M3 [37], 13) NFERM [7], 14) NIQE [18], 15) QAC [38] and 16) TCLT [35]. The implementations of all algorithms are obtained from the original authors. For IQA models that involve training, we use all images in the LIVE database [27] to train the models. To compensate the nonlinearity of model predictions on the human perception of image quality and to make the comparison more consistent, we adopt a logistic nonlinear function as suggested in [29] to map the prediction scores of each model to the MOS scale of the LIVE database [27]. As a result, the score range of all algorithms spans between $[0, 100]$, where a higher value indicates a better perceptual quality.

For each *defender* model, we define six quality levels evenly spaced on the quality scale, so that the selected subsets of images have a good coverage from low to high quality levels. The quality range within each subset of images is set to be within 1 standard deviation (std)[1]of MOSs in the LIVE database [27]. Thus the images within the same subsets have approximately the same or similar quality by the *defender* model. The *attacker* models then choose pairs of images from each of the 6 subsets, as described in Section 4. After the gMAD image pair selection process, a total of $16 \times (16-1) \times 6 = 1440$ image pairs are chosen for the subsequent subjective testing.

A subjective quality discrimination test is conducted in an office environment with normal indoor illumination levels and without reflecting ceiling walls and floor. The display is a Truecolor LCD monitor at a resolution of 2560×1600 pixels and is calibrated in accordance with the recommendations of ITU-R BT.500 [29]. A customized MATLAB interface is adopted to render a pair of images simultaneously at their original pixel resolutions but in random spatial order. A scale-and-slider applet is used for assigning a quality score, as shown in Fig. 2. For each pair of images, the subject assigns a score between -100 and 100 to indicate his/her preference to either the left image [-100, -20] (labeled as “left is better”) or the right image [20,100] (labeled as “right is better”). In case the subject is uncertain about his/her decision, he/she can also assign a score between [-20, 20] (labeled as “uncertain”), where a score 0 indicates completely neutral. This approach is different from a typical paired comparison method where the subjects can only make a binary decision on his/her preference even when he/she is uncertain about the answer. The benefit of the current approach is to better capture the subjects' confidence when expressing his/her preferences. During the experiment, the subjects are allowed to move their positions to get closer or further away from the screen for better observation. We divide the experiment into 4 sessions, each of which is limited to a maximum of 30 minutes to minimize the influence of fatigue effect. Furthermore, in order to inspect if subjects are using consistent scoring strategies throughout the experiment, we repeat 10% of the total number of image pairs (144 pairs) during the test.

A total of 31 naïve subjects, including 16 males and 15 females, participate in the subjective experiment. The subjects do not have any experience in the area of IQA and all have a normal or correct-to-normal visual acuity. Each subject is first introduced about the goal of the experiment and is then given an introduction on the experimental procedure and the user interface. They are also shown pairs of sample images (independent of the test images) in a training session so as to become familiar with the test process and image distortions. All subjects participate in all sessions.

[1] Every image in the LIVE database has a MOS and an std associated with it, computed from all valid subjects. The std used here is in fact an average of stds for all images.

5.2. Analysis

After the raw subjective data are collected, we employ the outlier detection and subject rejection algorithm suggested in [29]. Specifically, the raw score for an image is considered to be an outlier if it is outside 2 stds about the mean score of that image for Gaussian case or outside $\sqrt{20}$ stds for non-Gaussian case. A subject is removed if more than 5% of his/her evaluations are outliers. Moreover, a consistency check is conducted for each subject by making use of the image pairs that have been repeated. We define the consistency measure as the average of stds of scores given by one subject to the repeated pairs. A subject is rejected if his/her consistency measure is more than 2 stds of consistency measures for all subjects. As a result, one subject is rejected due to inconsistent judgements. Among all scores given by the remaining valid subjects, about 1.4% of the total subjective evaluations are identified as outliers and are subsequently removed.

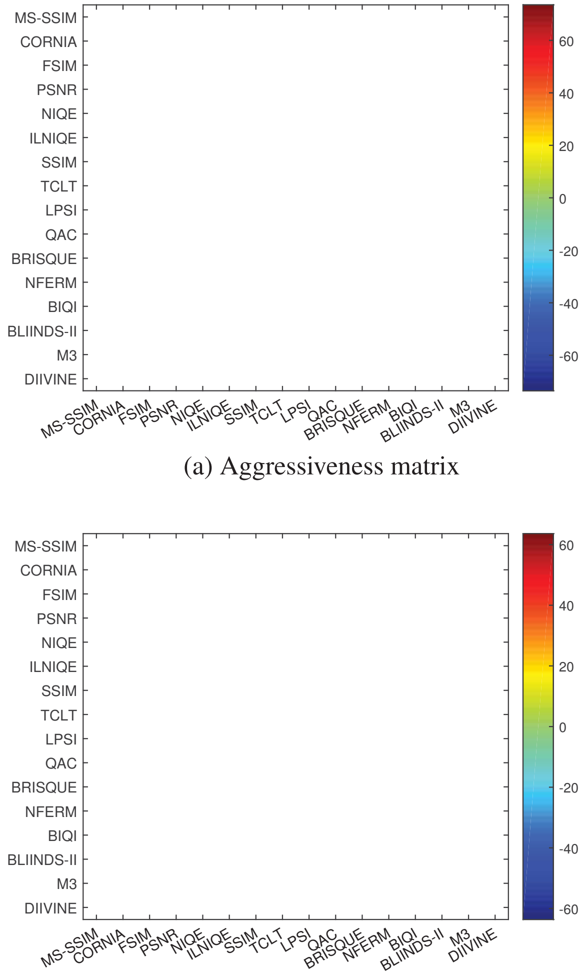
Since every pair of images in the gMAD competition are associated with two IQA models, we first compare these models in pairs and then aggregate the pairwise comparisons into a global ranking using mature rank aggregation tools such as maximum likelihood for multiple options [28], hodgeRank [10] and ranking by eigenvectors [16]. We define an aggressiveness matrix $\mathbf{A}$ and a resistance matrix $\mathbf{R}$, within which an entry a_{ij} represents the *aggressiveness* of the i-th model as an *attacker* against the j-th model as a *defender*, and an entry r_{ij} represents the *resistance* of the i-th model as a *defender* against the j-th model as an *attacker*, respectively. The *aggressiveness* measure indicates how strong an *attacker* in disproving a *defender* and is evaluated by

$$a_{ij} = \frac{\sum_{k=1}^{K} p_{jk} s_{ijk}}{\sum_{k=1}^{K} p_{jk}}\,, \tag{1}$$

where s_{ijk} is the subjective score averaged over all valid subjects on the image pair selected from the k-th subset. p_{jk} is the number of samples in the k-th subset. The value of a_{ij} ranges between $[-100, 100]$ with a larger value indicating stronger *aggressiveness* of the i-th model. In general, a_{ij} is expected to be positive for a competitive model, but it may also be negative which means that the order of the test image pair selected by the i-th model is the opposite of the average subjective judgements. A negative a_{ij} is a strong indication of a failure of the i-th model. On the other hand, the *resistance* measure indicates how resistent of a *defender* to be defeated (disproved) by an *attacker*. It is evaluated by

$$r_{ij} = \frac{\sum_{k=1}^{K} p_{ik} (100 - |s_{jik}|)}{\sum_{k=1}^{K} p_{ik}}\,. \tag{2}$$

r_{ij} ranges between $[0, 100]$ with a higher value indicating better *resistance* of the i-th model as a *defender* against the j-th model as an *attacker*. The matrices $\mathbf{A}$ and $\mathbf{R}$ are computed by comparing all pairs of IQA models and the results are shown in Fig. 3, where the higher value of an entry (warmer color), the stronger the *aggressiveness* and *resistance* of the corresponding row model against the column model.



(a) Aggressiveness matrix

(b) Resistance matrix

Figure 3. Pairwise gMAD competition matrices: Each entry indicates the Aggressiveness (a) or the Resistance (b) of the row IQA model against the column IQA model. $\mathbf{A} - \mathbf{A}^T$ and $\mathbf{R} - \mathbf{R}^T$ are drawn here for better visibility.

We aggregate the pairwise comparison results into a global ranking via a maximum likelihood method for multiple options [28]. The results are shown in Fig. 4. Using other ranking aggregation algorithms such as hodgeRank [10] and ranking by eigenvectors [16] gives very similar results. From the figure, we have several useful observations. First, an IQA model that has a stronger *aggressiveness* generally also has a stronger *resistance*. The Kendall's rank-order

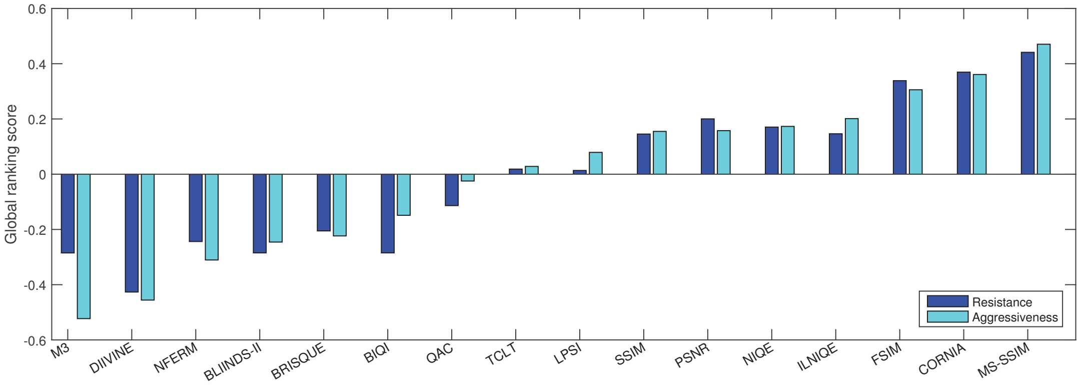


Figure 4. Global ranking of IQA models in terms of Resistance and Aggressiveness.

correlation coefficient (KRCC) between them is 0.87. Second, in general, FR-IQA algorithms are more competitive than NR-IQA methods. This is not surprising because FR algorithms make use of more information. Third, the best performance overall is obtained by MS-SSIM [34], which is a multi-scale version of SSIM [32] and significantly improves upon it. This suggests that multi-scale approaches are important in improving the performance of IQA models. Fourth, CORNIA [40], NIQE [18] and its feature enriched version ILNIQE [42] perform the best among all NR-IQA algorithms. It is worth mentioning that these methods are based on perception- and distortion-relevant natural scene statistics (NSS) features either hand-crafted or learned from data. This reveals the power of the NSS features, which help map images into a perceptually meaningful space for comparison. Fifth, a model that is worth noting is LPSI [36], which essentially reduces the feature space to one dimension and without using MOS for training, but it outperforms sophisticated machine learning-based approaches such as BRISQUE [17] and DIIVINE [20] which use many features for training. Sixth, machine learning based IQA models, though performed very well in existing publicly available databases, generally do not perform well in the current gMAD competition. This may be because the training samples are not sufficient to represent the population of real-world natural images and thus the risk of over fitting is high.

Furthermore, we perform a rational test to evaluate the robustness of IQA models when rating images with the same content and the same distortion type but different distortion levels. The underlying assumption is that the quality of an image degrades monotonically with the increase of the distortion level for all distortion types, and a good IQA model should rank the images in the same order. An example is given in Fig. 5, where the quality scores given by a good IQA model is supposed to decrease monotonically with the increase of the level of JPEG2000 compression. We use KRCC to check the consistency of the rankings between the distortion levels and the predicted scores of a given model. An overall consistency measure is defined as

$$C = \frac{1}{HT}\sum_{i=1}^{H}\sum_{j=1}^{T}\mathrm{KRCC}(\mathbf{l}, \mathbf{q}_{ij}), \tag{3}$$

where $H = 4744$ and $T = 4$ are the numbers of source images and distortion types in the database, respectively. $\mathbf{l} = [1, 2, 3, 4, 5, 6]$ represents the 6 distortion levels and $\mathbf{q}_{ij}$ is a 6×1 vector that contains the corresponding quality scores given by a model to 6 images, which have the same (i-th) source image and the same (j-th) distortion type but different distortion levels. Fig. 6 shows the overall consistency results of 16 IQA models, from which we have several observations. First, it is not surprising that FR models generally perform better than NR approaches because they are fidelity measures on how far away a distorted image departs from the source image, and such fidelity typically decreases monotonically with increasing distortion levels. Second, the NR model CORNIA [40], NIQE [18] and its feature enriched extension ILNIQE [42] outperform all other NR-IQA models, which coincides with the results of the gMAD competition shown in Fig. 4. Third, training based models generally have a lower overall consistency value and a larger error bar, suggesting lower robustness.

6. Conclusion and Future Work

In this paper, we attempted to address the IQA model comparison problem to overcome the conflict between the enormous size of image space and the limited resource for subjective testing. Our major contributions are threefold. First, we built a database of $4,744$ high quality source natural images and $94,880$ distorted images, which is the largest in the literature of IQA research. Second, we proposed a substantially different methodology named gMAD competition to evaluate the relative performance of multiple IQA models. Different from conventional methods that attempt to prove a model, gMAD focuses on disproving a model in the most efficient way using automatically selected

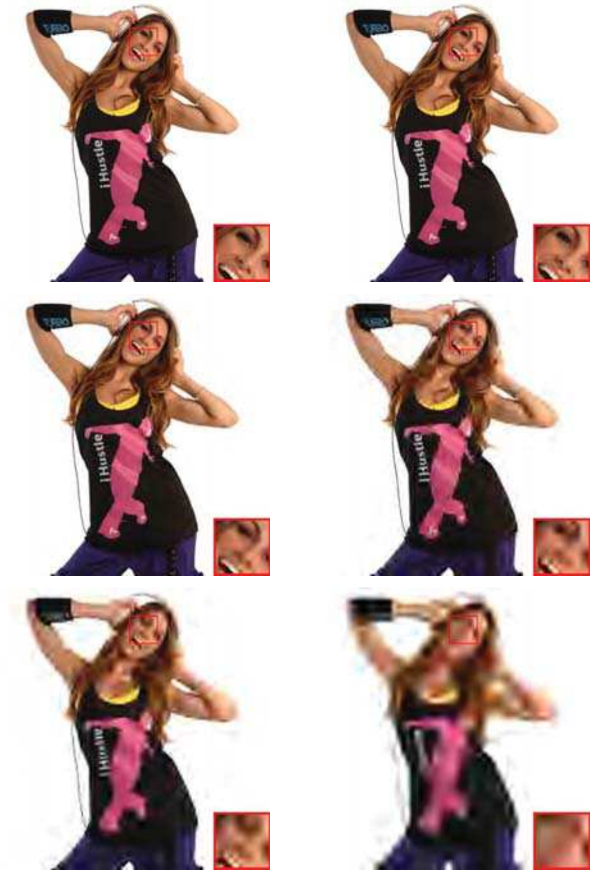


Figure 5. Illustration of the rational test on the "Hip-hop Girl" image under JPEG2000 compression. Obviously, the image quality degrades with the distortion level from left to right and from top to bottom ($\mathbf{l} = [1, 2, 3, 4, 5, 6]$). A good IQA model (ILNIQE [42] for example) ranks the images in exactly the same order. By contrast, a less competitive model may give a different order, e.g., the QAC model [38] ranks the image as $\mathbf{q} = [4, 3, 1, 5, 2, 6]$.

and model-dependent image pairs. The number of selected image pairs does not scale with the number of images, allowing it to exploit image databases of any size without increasing its complexity. Third, applying gMAD to the new database, we performed a systematic comparison of 16 well-known IQA models and made a number of useful observations.

The current work can be extended in many ways. First, the current database can continuously grow to include more image contents and more distortion types, and future IQA models can be added into the gMAD competition. We will make the database as well as the gMAD competition protocol and source code available online to facilitate future broader usage by researchers in the IQA community. Second, many useful observations have been made through the gMAD competition process. They can be used to facilitate further improvement of existing IQA models or future development of new IQA models. Third, the application scope of the fundamental idea behind gMAD is far beyond IQA model comparison. It is indeed a general methodology that can be used to compare any group of computational models used to predict certain continuous-scale quantities that need to be validated by expensive testing such as human subjective evaluation. To give a few examples, these may include comparisons of image/video interestingness predictors in the field of cognitive vision [8], the relative attributes (sportiness, furriness) estimators in the field of sematic image search [11], machine translation quality estimators in the field of computational linguistics [6], and thermal comfort models in the field of thermal environment of buildings [21].

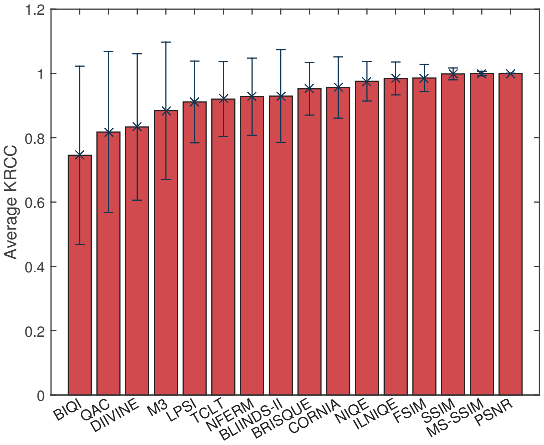


Figure 6. Overall KRCC consistency of IQA models.

Acknowledgments

This work was supported in part by the NSERC Discovery Grant and Steacie Memorial Fellowship programs of Canada, the National Natural Science Foundation of China (No. 61525102) and the Hong Kong RGC GRF grant (PolyU 5313/13E). Kede Ma was also partially supported by the CSC.

References

[1] W. H. Clatworthy. Partially balanced incomplete block designs with two associate classes and two treatments per block. *Journal of Research of the National Bureau of Standards*, 54:177–190, 1955. 2

[2] A. Donath and D. Kondermann. Is crowdsourcing for optical flow ground truth generation feasible? In *Computer Vision Systems*, pages 193–202. Springer Berlin Heidelberg, 2013. 1

[3] A. Eichhorn, P. Ni, and R. Eg. Randomised pair comparison: an economic and robust method for audiovisual quality assessment. In *ACM international workshop on Network and*

operating systems support for digital audio and video, pages 63–68, 2010. 2

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

[5] Google. https://images.google.com/. 3

[6] Y. Graham. Improving evaluation of machine translation quality estimation. In *53rd Annual Meeting of the Association for Computational Linguistics*, pages 1804–1813, 2015. 8

[7] K. Gu, G. Zhai, X. Yang, and W. Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, 2015. 3, 5

[8] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *IEEE International Conference on Computer Vision*, pages 1633–1640, 2013. 8

[9] Y. Horita, K. Shibata, Y. Kawayoke, and Z. M. Parvez. Toyama-MICT image quality evaluation database 2010 [online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb. 1, 2

[10] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. 3, 6

[11] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012. 8

[12] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *SPIE Journal of Electronic Imaging*, 19(1), 2010. 1, 2

[13] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database 2005 [online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/. 1, 2

[14] J. Li, M. Barkowsky, and P. Le Callet. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In *IS&T/SPIE Electronic Imaging*, 2013. 2

[15] L. Maier-Hein, D. Kondermann, T. Roß, S. Mersmann, E. Heim, S. Bodenstedt, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, et al. Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences. *International journal of computer assisted radiology and surgery*, pages 1–12, 2015. 1

[16] C. D. Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2000. 6

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 3, 5, 7

[18] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 3, 5, 7

[19] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010. 3, 5

[20] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. 3, 5, 7

[21] J. F. Nicol and M. A. Humphreys. Adaptive thermal comfort and sustainable thermal standards for buildings. *Energy and buildings*, 34(6):563–572, 2002. 8

[22] 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*, 30:57–77, 2015. 1, 2, 3

[23] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 1, 2, 3

[24] A. Rosenfeld. Picture processing by computer. *ACM Computing Surveys*, 1(3):147–176, 1969. 1

[25] 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*, 21(8):3339–3352, 2012. 3, 5

[26] A. Saha and Q. M. J. Wu. Utilizing image scales towards totally training free blind image quality assessment. *IEEE Transactions on Image Processing*, 24(6):1879–1892, 2015. 3

[27] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 1, 2, 3, 5

[28] K. Tsukida and M. R. Gupta. How to analyze paired comparison data. Technical Report UWEETR-2011-0004, University of Washington, 2011. 6

[29] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment 2000 [online]. Available: http://www.vqeg.org. 5, 6

[30] Z. Wang and A. C. Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2006. 1

[31] Z. Wang and A. C. Bovik. Mean squared error: love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 1

[32] 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*, 13(4):600–612, 2004. 5, 7

[33] Z. Wang and E. P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8, 2008. 2, 4

[34] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003. 5, 7

[35] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng. Blind image quality assessment based on multichannel features fusion and label transfer. *to appear in IEEE*

Transactions on Circuits and Systems for Video Technology, 2016. 3, 5

[36] Q. Wu, Z. Wang, and H. Li. A highly efficient method for blind image quality assessment. In *IEEE International Conference on Image Processing*, pages 339–343, 2015. 3, 5, 7

[37] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014. 3, 5

[38] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, 2013. 3, 5, 8

[39] P. Ye and D. Doermann. Active sampling for subjective image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4249–4256, 2014. 3

[40] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012. 3, 5, 7

[41] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić. VCL@FER image quality assessment database. *AUTOMATIKA*, 53(4):344–354, 2012. 1, 2

[42] L. Zhang, L. Zhang, and A. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 3, 5, 7, 8

[43] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: a feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 5