ASSP: An adaptive sample statistics-based pooling for full-reference image quality assessment

Yurong Ling[a], Fei Zhou[b,*], Kun Guo[c], Jing-Hao Xue[a]

[a] *Department of Statistical Science, University College London, UK*
[b] *College of Information Engineering, Shenzhen University, China*
[c] *School of Psychology, University of Lincoln, UK*

Abstract

Most full-reference image quality assessment (IQA) models first compute local quality scores and then pool them into an overall score. In this paper, we develop an innovative pooling strategy based on sample statistics to adaptively make the IQA more consistent with human visual assessment. The innovation of this work is threefold. First, we identify that standard sample statistics and robust sample statistics could provide complementary information about the degree of degradation in distorted images. Second, an effective IQA metric is proposed by adaptively integrating robust sample statistics and standard sample statistics via excess kurtosis. Third, instead of using the statistics directly, we adjust them by taking into account the global change of image gradients to avoid exaggerating the degradation degree. Experiments conducted on five well-known IQA databases demonstrate the effectiveness of the proposed pooling strategy in terms of high prediction accuracy and monotonicity.

Keywords: IQA, statistics-based pooling, robust sample statistics, excess kurtosis

[*] Corresponding author

Email addresses: `yurong.ling.16@ucl.ac.uk` (Yurong Ling), `flying.zhou@163.comm` (Fei Zhou), `kguo@lincoln.ac.uk` (Kun Guo), `jinghao.xue@ucl.ac.uk` (Jing-Hao Xue)

1. Introduction

Image quality assessment (IQA) plays a crucial role in numerous applications, such as image acquisition, transmission and restoration, and a variety of distorted images and subjective evaluations are introduced to facilitate the development of three types of objective IQA approaches [1–8]: full-reference (FR) IQA, reduced-reference IQA and no-reference IQA. This paper focuses on FR-IQA where the reference image is fully known.

Among FR-IQA methods, mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) are two widely used metrics. However, they do not correlate well with the human visual system (HVS) [9]. Several IQA metrics, such as SSIM [10], MS-SSIM [11], IW-SSIM [12] and MvSSIM [13], are developed as the HVS perceive the structural loss. Most FR-IQA methods share two main steps (Fig. 1): first calculate local quality scores from image features, and then pool these local scores into a single overall score.

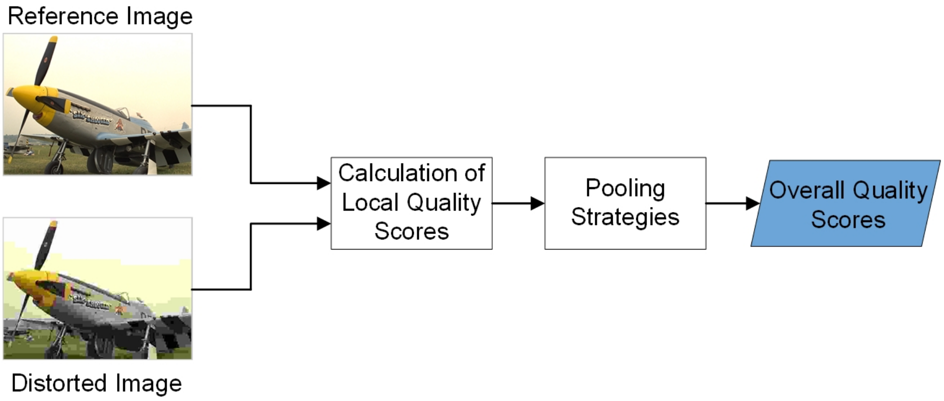


Figure 1: Flowchart of common FR-IQA models.

In the first step, various image features have been adopted, including gradient similarity in GSM [14], gradient magnitude in the feature similarity index (FSIM and FSIMc) [15], phase congruency [16], coding coefficients in the most apparent distortion model [17], two-dimensional mel-cepstrum features [18], local linear model (LLM) [19], sparse coding significance features [20], superpixel-based regional gradient consistency in SPSIM [21] and deep features [22, 23]. These methods focus on various structural rather than statistical features. In contrast, many methods prefer statistics-based features. Mutual information between reference and distorted images is calculated to construct the fidelity in the metrics of information fidelity criterion (IFC) [24] and visual information fidelity (VIF) [25]. Moment correlation between reference and distorted images is considered in [26].

In the second step, various pooling strategies have been adopted. An intuitive way to gain the overall score is to use the mean of local quality scores [10]. Using the mean implies that all local regions contribute equally to the perception; however, it is inconsistent with the HVS [27, 28]. Hence, the weighted mean is proposed since humans tend to look at a selection of local image regions in image-viewing [12, 29–31]. Although using weighted mean can improve IQA, it is not easy to propose a proper weighting function. The weighted mean would fail to boost the performance of an IQA model if the constructed weighting function is inconsistent with the HVS. Apart from mean and weighted mean, standard deviation (SD) is used for pooling in the method of gradient magnitude similarity deviation (GMSD) [32], which is inspired by the fact that the dispersion of local quality scores increases with the degradation degree. However, as we will discuss later, solely using SD to measure the degradation is insufficient in some cases (Section 2.2.3), and the SD pooling strategy is not always consistent with the HVS (Section 2.2.2).

Recently, statistical techniques have been widely investigated for IQA [33] including our previous efforts on hypothesis testing for comparison between different IQA metrics [34] and on robust statistics for NR-IQA [35]. Inspired by the popularity of statistical approaches, we study sample statistics in this work which have not been fully explored for the FR-IQA pooling [36], in particular the statistics of gradient similarity, although it is a popular metric in the first step. Therefore, in this paper, we investigate the integration of various sample statistics of gradient similarity to propose an innovative adaptive sample statistics-based pooling (ASSP) strategy for FR-IQA. It is worth mentioning that there are numerous proposals of NR-IQA methods [37–40] using deep neural networks to extract features for calculating local distortion scores and to pool these scores. We believe our pipeline of investigating the use of sample statistics can also be extended to link the sample statistics of the features extracted by neural networks and the HVS, enhancing the interpretation of these approaches.

Two groups of sample statistics are explored in the proposed ASSP: standard sample statistics and robust sample statistics. Here the robust sample statistics refer to the sample statistics that are not unduly affected by outliers. Typical robust sample statistics include the median, trimmed mean, median absolute deviation, and interquartile range (IQR) [41, 42]. Here we exploit a robust measure of dispersion, RD, derived from a robust measure of skewness [43] and designed for skewed distributions. In total, four sample statistics are explored in ASSP: two standard statistics (SD, mean) and two robust statistics (RD and median).

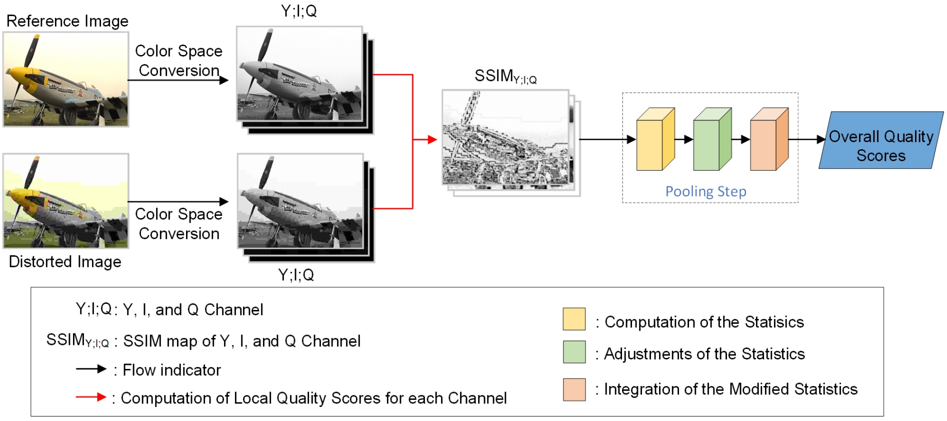
We first observe that if the proportion of outliers among local quality scores is relatively large, local regions of outliers can have major impacts on HVS, and standard sample statistics could reflect the impact of outliers better. Conversely, if the outlier proportion is small, HVS is barely influenced by outliers, and therefore robust statistics are more suitable as they can exclude the influence of outliers during pooling. In other words, it will be beneficial to IQA by considering both robust sample statistics and standard sample statistics for pooling (see Section 2.2.2 for the detailed discussion of adaptive use of these two types of sample statistics). A popular way to determine whether the proportion of outliers is large is to evaluate the excess kurtosis, a descriptor of the tail of distributions. Hence we use the excess kurtosis to construct a weighting function between robust and standard sample statistics to adaptively integrate them.

Then we propose an adjustment of the statistics to make them more consistent with HVS. The adjustment aims to avoid the statistics' exaggeration of degradation degree. This is achieved by using the gradient of the image to transform the statistics, as detailed in Section 2.3.

In summary, the contributions of our work are threefold. First, we identify that standard sample statistics and robust sample statistics could provide complementary information about the degree of degradation in distorted images. Second, we propose an effective IQA metric to adaptively integrate these two types of sample statistics via excess kurtosis. Third, we also adjust these statistics by considering the global change of image gradients to avoid exaggerating the degradation degree. Moreover, in Section 3 we conduct extensive experiments on five well-known IQA databases, the results of which demonstrate the effectiveness of the proposed pooling strategy in terms of high prediction accuracy and monotonicity.

2. Proposed ASSP framework

The flowchart of our proposed ASSP is illustrated in Fig. 2. First, the reference image and the distorted image are converted into the YIQ colour space. Local quality scores for each channel are then computed separately (Section



Reference Image
Color Space Conversion
Y;I;Q
SSIM_Y;I;Q
Overall Quality Scores
Pooling Step
Color Space Conversion
Distorted Image
Y;I;Q
Y;I;Q : Y, I, and Q Channel
SSIM_Y;I;Q : SSIM map of Y, I, and Q Channel
: Flow indicator
: Computation of Local Quality Scores for each Channel
: Computation of the Statisics
: Adjustments of the Statistics
: Integration of the Modified Statistics

Figure 2: Flowchart of the proposed ASSP framework.

2.1). The proposed sample statistics-based pooling strategy is composed of three steps: 1) calculate the statistics of local quality scores (Section 2.2); 2) adjust the statistics (Section 2.3); and 3) integrate the statistics within and across channels (Section 2.4).

2.1. Local quality scores

In this paper, we focus on the pooling strategy, and the local quality scores are borrowed from FSIMc [15]. There are two types of local quality scores in the proposed method: the similarity between gradient magnitudes and that between chromatic features.

Image gradients have been successfully applied to image quality assessment [14, 15, 32, 44, 45]. In this paper, the Prewitt filters are used for their simplicity to calculate gradients. $\mathbf{X}_r(i)$ and $\mathbf{X}_d(i)$ denote the gradients magnitudes of reference image $\mathbf{r}$ and distorted image $\mathbf{d}$ respectively at position i. We use the similarity between $\mathbf{X}_r(i)$ and $\mathbf{X}_d(i)$ as a local quality score:

$$S_g(i) = \frac{2\mathbf{X}_r(i)\mathbf{X}_d(i) + C_1}{\mathbf{X}_r^2(i) + \mathbf{X}_d^2(i) + C_1}, \tag{1}$$

where C_1 is a positive constant for numerical stability.

For colour images, since the chrominance information is valuable for detecting the degradation related to colour [15, 21, 46], we not only calculate the similarity of gradient magnitudes in the luminance channel but also compute the similarity between chromatic channels. To obtain the chrominance information, we convert the images into the YIQ colour space [47], which has been successfully applied to IQA models [15, 46]. Let Y denote the luminance channel and I and Q denote the two chromatic channels. The similarity of chromatic channels is measured in a way similar to that for the gradient similarity:

$$S_I(i) = \frac{2\mathbf{I}_r(i)\mathbf{I}_d(i) + C_2}{\mathbf{I}_r^2(i) + \mathbf{I}_d^2(i) + C_2}, \quad S_Q(i) = \frac{2\mathbf{Q}_r(i)\mathbf{Q}_d(i) + C_2}{\mathbf{Q}_r^2(i) + \mathbf{Q}_d^2(i) + C_2}, \tag{2}$$

where $\mathbf{I}_r(i)$ ($\mathbf{I}_d(i)$) and $\mathbf{Q}_r(i)$ ($\mathbf{Q}_d(i)$) are the I and Q chromatic channels of reference image $\mathbf{r}$ (distorted image $\mathbf{d}$) at position i, respectively. Since I and Q components have similar dynamic ranges, we use the same constant C_2 to avoid numerical instability.

2.2. Statistics of local quality scores

Table 1: Statistics used in this paper

	Robust statistics	Standard statistics
Central tendency	Median	Mean
Dispersion	RD	SD

In the pooling stage, four sample statistics (Table 1) of local quality scores in each channel are first calculated. For brevity, we shall omit the subscript of channel unless it might raise confusion.

2.2.1. Multiple statistics

The four statistics used can be divided into two groups in two ways. According to the type of information that the statistics provide, we have the mean and median as the measures of central tendency (typical values) of local quality scores, and the SD and RD as the measures of the dispersion of scores. According to the robustness of the statistics, we have the mean and SD as standard non-robust statistics, and the median and RD as robust statistics [41–43]. In this subsection, we shall discuss the proper identification of outliers among local quality scores and the effectiveness of the statistics used for characterising the distortion degree.

Boxplots are widely-used in the exploration of data, and data points are neither within 1.5 IQR of the lower quartile nor within 1.5 IQR of the upper quartile, where IQR is the difference between the upper quartile and the lower quartile, are always flagged as outliers which behave differently from the majority of the data [48]. This way of identifying outliers, however, assumes that regular data points are symmetrically distributed, and thus results in incorrect detection of outliers if the assumption is violated. A well-founded modification regarding outliers is proposed in [43] to distinguish better between regular observations and outliers especially for skewed distributions. In this paper, we adopt the definition of outliers in [43], because the distributions of local quality scores are mostly skewed and meanwhile, the definition is the same as that with regular boxplots if the distributions are symmetric.

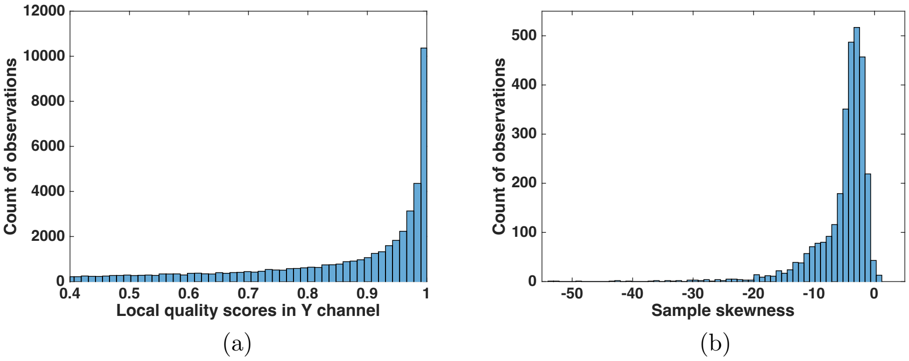


Figure 3: Histograms of local quality scores and sample skewness for TID2013. (a) Histogram of local quality scores in the Y channel obtained from one TID2013 distorted image, which is contaminated by the JPEG compression. (b) Histogram of the sample skewness of local quality scores in the Y channel for all TID2013 distorted images; the sample skewness of 83.73% distorted images is less than -1.

To visualise the skewness of distributions of local quality scores, two histograms are shown in Fig. 3. It is found that the distribution is skewed to the left in Fig. 3(a) for a distorted image from TID2013. Furthermore, we calculate the sample skewness for each distorted image in TID2013, the histogram of which is provided in Fig. 3(b). It is noticeable that the sample skewness of most distorted images is less than -1, i.e. the distributions of local quality scores are skewed in most cases. It is therefore more appropriate to use the definition of outliers in [43] rather than that in [48].

In [43], the authors propose three models to identify outliers: linear model, quadratic model and exponential model. The model parameters are determined by fitting a whole range of distributions (12605 distributions) such that the expected percentage of marked outliers is close to 0.7%. We adopt the exponential model in our paper, since it performs the best in identifying outliers across different distributions. The set of outliers $S_{outlier}$ in the exponential model is defined as the extreme scores outside the interval (the fence):

$$[Q_1 - 1.5e^{a\mathrm{MC}}\mathrm{IQR}, Q_3 + 1.5e^{b\mathrm{MC}}\mathrm{IQR}], \tag{3}$$

where Q_1 and Q_3 are the 25th and 75th percentiles of local quality scores respectively, $IQR = Q_3 - Q_1$, and MC is the medcouple which is a robust measure of skewness [49]. MC is defined as

$$\mathrm{MC} = \mathrm{median}_{S(i) \leq Q_2 \leq S(j)} h(S(i), S(j)) \tag{4}$$

with $S(\cdot)$ and Q_2 representing a local quality score and the sample median, respectively, where for all $S(i) \neq S(j)$, the kernel function h is given by

$$h(S(i), S(j)) = \frac{(S(j) - Q_2) - (Q_2 - S(i))}{S(j) - S(i)}. \tag{5}$$

For $S(i) = S(j)$, a different definition applies, see [49]. The range of MC is $[-1, 1]$. Note that a negative value of MC indicates that the distribution is skewed to the left while a positive value indicates being right-skewed. If underlying distributions are symmetric, MC $= 0$. Meanwhile, $a = -4, b = 3$ if MC ≥ 0; $a = -3, b = 4$ if MC < 0. The adjusted boxplot accounts for skewness by adsorbing the value of MC in (3). Specifically, a positive value of MC indicates that the distribution is skewed to the right. Hence, the minimum distance from identified outliers in the left tail to Q_1 should not be equal to that from the outliers in the right tail to Q_3, otherwise many observations would be erroneously flagged as outliers due to the asymmetry. The asymmetry also indicates that the typical boxplot is not suitable in such a case since the whiskers, which define the outliers, are symmetric around the median. On the contrary, the minimum distance $1.5e^{-4\mathrm{MC}}\mathrm{IQR}$ in the left tail is less than the minimum distance $1.5e^{3\mathrm{MC}}\mathrm{IQR}$ in the right tail in terms of (3) in the adjusted boxplot, which indicates that the adjusted boxplot offers a better alternative in such cases since the mass of the distribution is concentrated on the left. It is worth mentioning that a and b play the same role as MC in accounting for skewness when identifying outliers. Similar conclusions can be reached for MC < 0. Note that the fence in (3) is equivalent to that in traditional boxplots for MC $= 0$.

By taking advantage of (3), we propose to use RD concerning the dispersion of regular data points, and it is defined by

$$RD = \max_{S(i) \leq Q_3 + 1.5e^{b\mathrm{MC}}\mathrm{IQR}} S(i) - \min_{S(j) \geq Q_1 - 1.5e^{a\mathrm{MC}}\mathrm{IQR}} S(j). \tag{6}$$

It is clear that the RD is not being much affected by outliers which are not taken into account in the measure. In contrast, the SD can be largely affected by outliers as it covers all extreme scores.

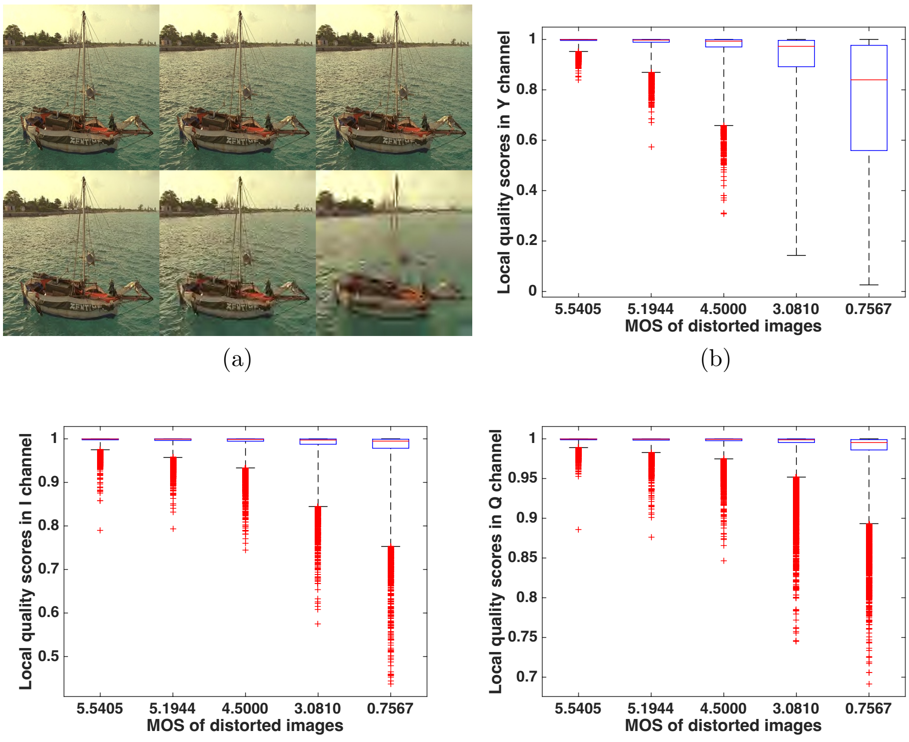
To our best knowledge, the usefulness of the RD and median to reflect the degradation degree in the pooling step for FR-IQA has not been explored. Here we take two examples to illustrate their usefulness. The first example is shown in Fig. 4, where the distortion type is JPEG2000 compression and the degradation lies in the Y channel. It is observed that the RD increases and the median decreases with perceptual quality decreases, and this pattern is much more remarkable in the Y channel as it should be. That is, both the RD and the median are indicative of different perceptual qualities.

The second example is shown in Fig. 5, where the distorted images are contaminated by the change of colour saturation, and the degradation exists in the I and Q channels. The RD and median show similar patterns as those in Fig. 4, but this time are much more clearly in the I and Q channels as they should be.

These examples indicate that both the RD and the median of local scores can be highly correlated to the HVS. Meanwhile, the SD pooling, proposed in GMSD, is similar to the RD in the sense that it also quantifies the dispersion of local quality scores. The good performance achieved by GMSD indicates its correlation with the HVS. Mean reflecting the central tendency as median does is also taken into account in the proposed statistics-based pooling, since it shows the different aspect of local quality scores as a standard sample statistics compared with the SD, which will be discussed in Section 2.2.3. We found that the robust statistics and standard statistics should be used adaptively in the pooling strategy, which we shall discuss in Section 2.2.2.

2.2.2. Standard statistics vs. robust statistics

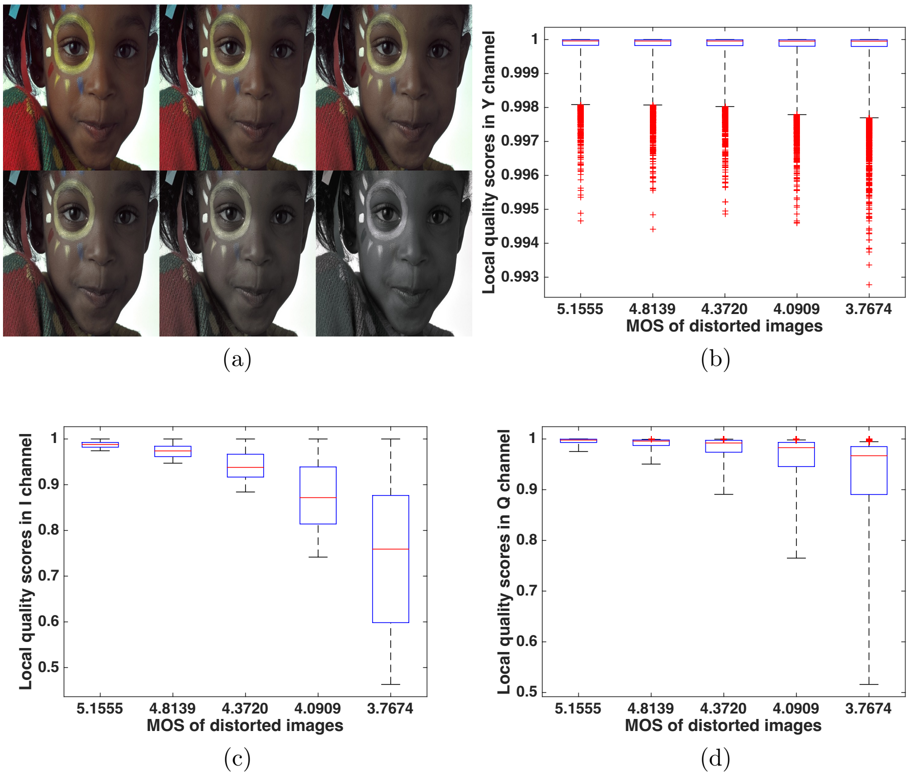
We observe that the HVS responds differently to outliers in local quality scores. The HVS has difficulty perceiving the outliers if they are few, indicating that in such cases the robust statistics can be more representative of the degradation degree. On the contrary, the HVS would pay attention to outliers if outliers are of a large proportion and thus informative, which



(a) (b)

(c) (d)

Figure 4: Adjusted boxplots of local quality scores for five distorted images in the (b) Y, (c) I and (d) Q channels and corresponding reference image and distorted images. (a) An TID2013 reference image and five distorted images. The leftmost image in the first row from the top is the reference image and the remaining images are distorted images. The perceptual quality decreases from top to bottom and from left to right in terms of the mean opinion score (MOS). The distortion type for all distorted images is JPEG2000 which affects the Y channel. The red line inside each box is the median. The bottom and top of each box are the 75th and 25th percentiles, respectively, and thus the box height is the IQR. The whiskers extend to the most extreme data points that are not outliers. The difference between the top end whisker and the bottom end whisker is the dispersion (RD) that we use. These adjusted boxplots indicate that the (lower) median and the (larger) RD can reflect well the (increased) Y-channel perceptual quality to the HVS.



(a)

(b)

(c)

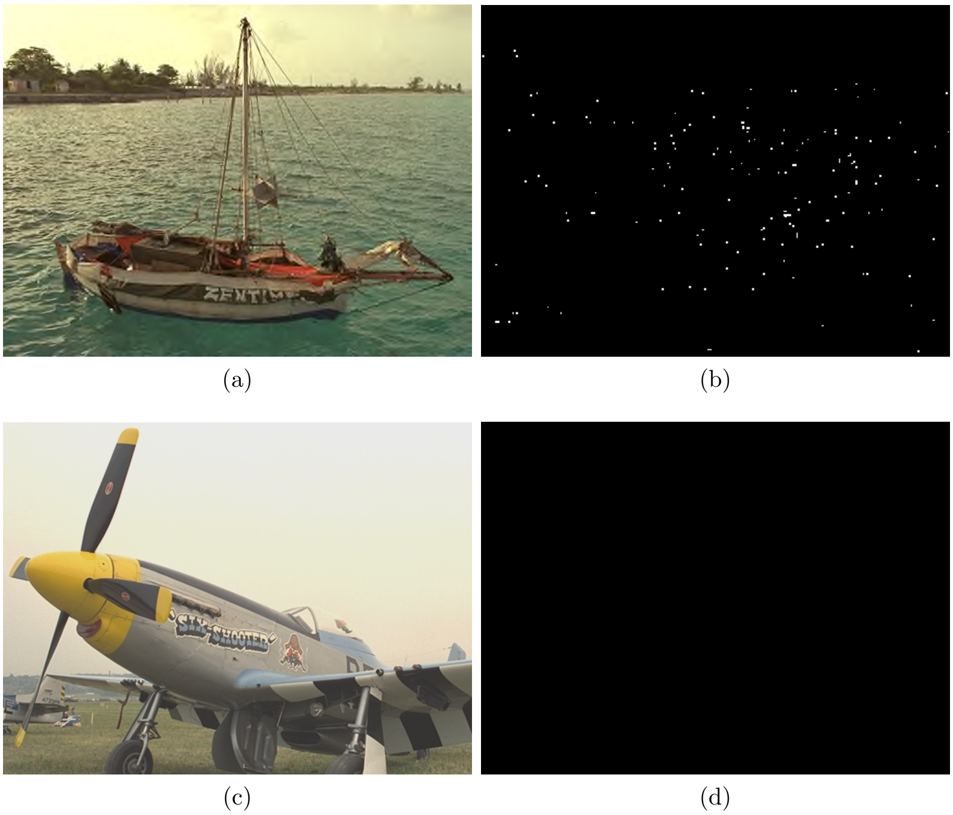
(d)

Figure 5: Adjusted boxplots of local quality scores for five distorted images in the (b) Y, (c) I and (d) Q channels and corresponding reference image and distorted images. (a) An TID2013 reference image and five distorted images. The distortion type is change of colour saturation which affects the I and Q channels.

suggests that in this case we should rely more on the non-robust statistics for pooling in order to extract and reflect the useful information from the outliers.

Since outliers are defined as local quality scores that are far away from most local quality scores, their values are either much higher or much lower than most scores. For example, if most local quality scores computed from one distorted image lie in the range $[0.8, 0.9]$, the outliers could be much closer to 0 or 1 compared with other local quality scores. Outliers that are of much lower scores than most local quality scores should correspond to pixels with relatively large degrees of degradation, and outliers that are of much higher scores should result from much less distorted pixels. When the outlier ratio is low, it is very likely pixels corresponding to outliers are not structured as informative local image region. In such cases, the HVS would ignore the outliers since pixels are only meaningful to the HVS when they correspond to regions. Indeed, human perceptual studies have implied that HVS may ignore the outliers if the proportion of outliers is low in the image. Since the HVS often needs to select and extract visual information from a noisy environment (e.g. due to out-of-focus, foggy or raining weather), it has evolved and/or learned over time to process these visual signals embedded in natural distortions, and developed certain tolerance to the degradation in image quality [50]. We could correctly understand nature scene gist and classify scene categories in low-resolution or blurred images [50–52]. Furthermore, the HVS often has an invariant representation of familiar/learned scenes or objects [53]. That is, these scenes/objects are still recognisable even though they appear in very different forms under different viewing conditions, such as varying optical quality and optical aberrations, brightness, viewing angle or viewing distance. Taken together, it is reasonable to assume that the perceptual assessment (e.g. judging image quality) of the distorted image would not be affected by few outliers with either lower or higher local quality scores from the rest of the image regions because these outliers have little impact on scene gist understanding, and the HVS may ignore these outliers (especially) from relatively large degrees of degradation. In contrast, when the proportion of outliers is high, the HVS should be able to perceive these outliers due to the large size of the corresponding regions, so in this case we should resort to the standard non-robust sample statistics which can reflect the influence of the outliers.

One example that illustrates the negative impact resulting from the outliers on the non-robust statistics is provided in Fig. 6, in which the outlier



(a) (b)

(c) (d)

Figure 6: Distorted images and their outlier maps. Distorted images contaminated by (a) JPEG2000 compression and (c) contrast change respectively. (b) Outlier map of (a). (d) Outlier map of (c). The outliers in the local quality scores in the Y channel are shown in white in the outlier maps while other scores in black. MOS of (a) is 4.5000 and that of (c) is 4.3243, indicating (a) is of better perceptual quality than (c). However, the mean (0.9720) of (a) is worse than that (0.9880) of (c); and the SD (0.0520) of (a) is also worse (i.e. larger) than that (0.0180) of (c).

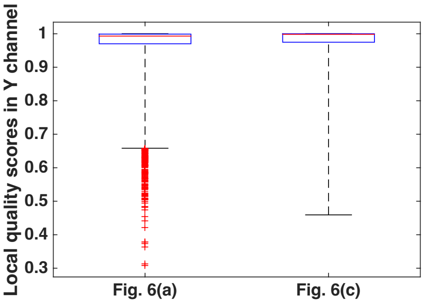
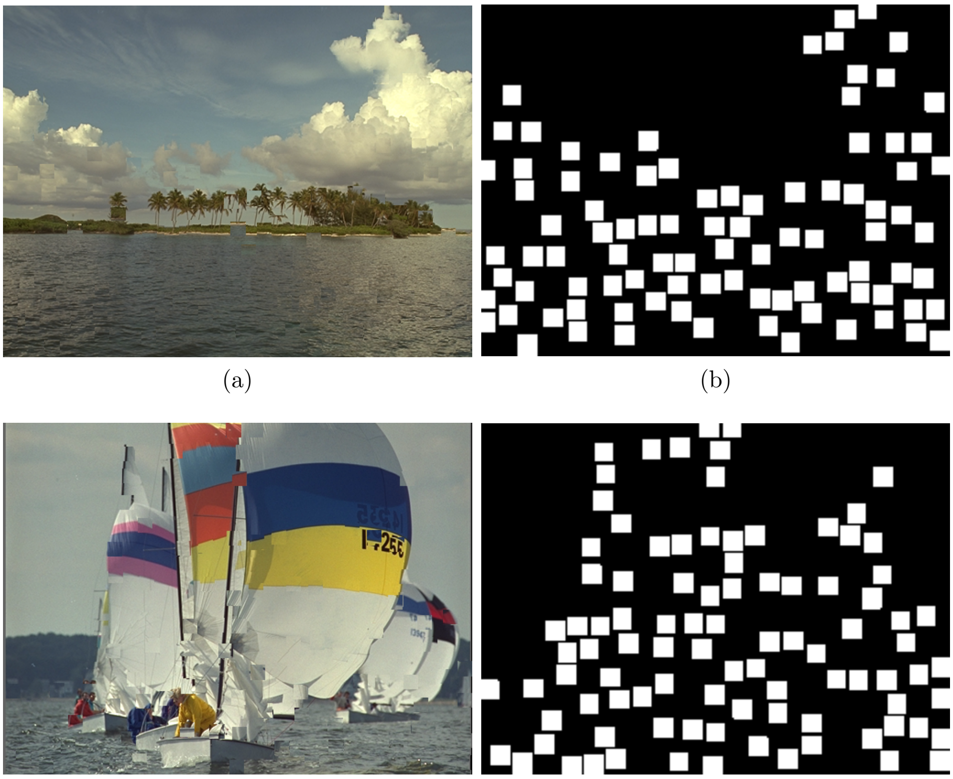


Figure 7: Adjusted boxplots of local quality scores computed from (a) and (c) in Fig. 6

ratios of (a) and (c) are 0.0033 and 0 respectively. It is clear that the quality of Fig. 6(a) is superior to that of Fig. 6(c) in terms of the MOS. However, the image quality of Fig. 6(a) would be considered to be inferior to that of Fig. 6(c) due to a larger SD and a smaller mean if we use these non-robust statistics to measure the degradation. Hence the use of non-robust statistics is inappropriate in this case. As a matter of fact, it is the outliers in the local quality scores that result in the improper quantification of degradation of Fig. 6(a). In order to better understand the influence of outliers on the non-robust statistics, we provide in Fig. 6(b) and (d) the corresponding outlier maps and in Fig. 7 the adjusted boxplots of local quality scores. It can be seen that the boxes that most local quality scores lie in are the same for Fig. 6(a) and Fig. 6(c). The median of Fig. 6(a) is roughly the same as that of Fig. 6(c), while RD of Fig. 6(a) is smaller, which is more consistent with the HVS. One difference between their local quality scores is that Fig. 6(a) has more outliers, corresponding to lower quality scores, than Fig. 6(c) does. These outliers inevitably lower the mean of Fig. 6(a). Meanwhile, Fig. 6(a)-(b) show that the outliers should be ignored by the HVS. If the outliers in Fig. 6(a) could be excluded when computing the statistics, we should be able to provide a better (more MOS-consistent) quantification of its degree of degradation.

To further illustrate the principle that we should put more emphasis on the non-robust statistics if the outlier ratio is high, we provide two distorted images with the corresponding MOS and statistics of the local quality scores in Fig. 8. The ratios of outliers are both high for Fig. 8(a) and Fig. 8(c), which are 0.2078 and 0.2160 respectively. As the degradation in Fig. 8 exists in the Y channel, we only provide the statistics in the Y channel to save



(a)

(b)

(c)

(d)

Figure 8: Distorted images contaminated by non eccentricity pattern noise and corresponding outlier maps (b) and (d). Outliers in the Y channel are first identified by the adjusted boxplots. We then use white squares to frame irregular areas that contain many outliers for clearer visualisation. (a) Distorted images with MOS 4.5750. (c) Distorted images with MOS 3.3125. The SD for (a) is 0.1030 and that for (c) is 0.1274. The mean for (a) is 0.9699 and that for (c) is 0.9630. Both medians for (a) and (c) are 1, and both RD are 0.

space. In terms of MOS, the perceptual quality of Fig. 8(a) is superior to that of Fig. 8(c). However, we are unable to infer the superiority of Fig. 8(a) based on the same RD and medians for both Fig. 8(a) and Fig. 8(c). On the contrary, a higher mean and a smaller SD of Fig. 8(a) indicate that the non-robust statistics are more consistent with the HVS in this case compared with the robust statistics. By comparing the outlier maps with the corresponding distorted images, it is observed that the regions corresponding to the outliers are of inferior quality, hence the proper quality metric should take into account of these outliers. Robust statistics, which exclude the outliers, are hence at odds with the HVS.

The above observations inspire us to adaptively integrate robust statistics and standard statistics, leading to a new integration method proposed in Section 2.4.

2.2.3. Central tendency vs. dispersion

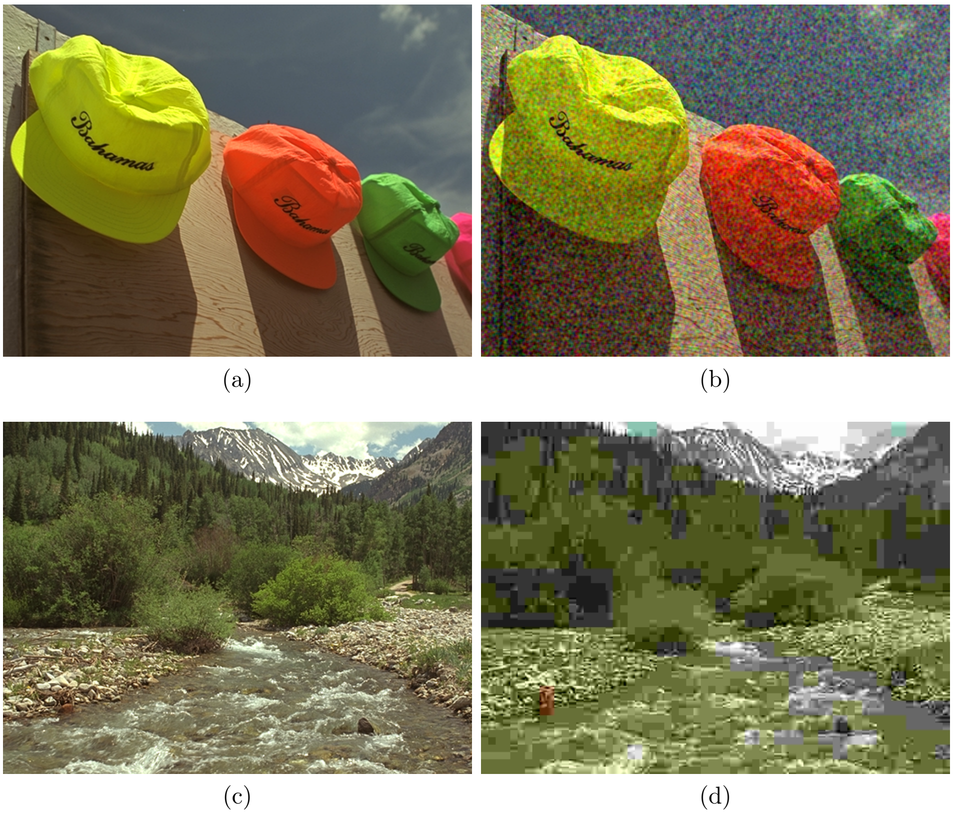
It is shown that the RD and SD are able to quantify the degradation in the pooling step. They both account for the dispersion of local quality scores without indicating the typical values (central tendency). On the contrary, the mean/median can represent the typical value of the scores. The image qualities of two distorted images, with the same degree of dispersion but entirely different typical values, or with roughly the same typical values but large difference in dispersion, should be different. If we only use the measure of central tendency or the measure of dispersion to quantify the degradation, we cannot differentiate these two distorted images from the perspective of quality. We hence use both the measure of central tendency (mean/median) and the measure of dispersion (SD/RD) in the pooling step.

2.3. Adjustment of the statistics using gradient information

Although the robust/non-robust statistics of local quality scores are capable of quantifying the degradation, they are not exactly consistent with the visual perception of distorted images. We therefore adjust the statistics by taking the characteristics of HVS into account, as detailed in this section.

For some types of degradation, the direct use of sample statistics could exaggerate the degradation. An example is shown in Fig. 9.

One reason for the exaggeration is that the statistics do not consider different perceptual effects on positive or negative changes of gradient magnitude. Let us take an extreme example. Suppose we have two distorted images of the same reference image, the gradient magnitude for each pixel in



(a) (b)

(c) (d)

Figure 9: Two pairs of reference and distorted images: (b) and (d) are distorted versions of (a) and (c), respectively. The MOS of (b) is 2.8947, larger than 1.1627 of (d), indicating that (b) has a lower degree of degradation and thus a better perceptual quality than (d). However, regarding the local quality scores of (b), the mean (0.6898) and median (0.7252) are lower than those (0.7904 and 0.8930) of (d), indicating that (b) has a poorer quality than (d), a conclusion opposite to that drawn from the MOS. The same exaggeration of statistics can be observed too from the SD (0.2545) and RD (0.9461) of (b), which are larger than those (0.2347 and 0.9376) of (d), indicating poorer quality of (b).

one distorted image increases while in the other one decreases. Meanwhile, the changes in the gradient magnitude at the same position for both distorted images are equal. This will lead to the same objective score for the two distorted image in terms of all statistics due to the symmetric way of calculating local quality scores. However, the distorted image with increasing gradient magnitudes should be assessed with higher quality from the perspective of visual perception. Loosely speaking, the increased gradient magnitudes from the original image generally result from two cases in the context of IQA: one case is the noise, and the other case is the contrast change [54]. Both the noise and the contrast change lead to components with a higher spatial frequency. In [50], the authors find that the perceived quality of natural scene images with a higher spatial frequency are less affected by the noise distortion than those images with lower spatial frequency, since complex structures in the images with a high spatial frequency offer people more local information for the extraction and analysis. Regarding the contrast change, higher contrast can also improve the image quality perceived by the HVS [55, 56].

In order to alleviate the exaggeration problem, we incorporate the global change of gradient magnitudes between the reference and distorted images in the statistics for further adjustment. The relatively global change (gc) of gradient magnitudes between the reference image and the distorted image is defined as

$$gc = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{X}_r(i) + C_3}{\mathbf{X}_d(i) + C_3}, \tag{7}$$

where n is the number of pixels and C_3 is used for numerical stability when $\mathbf{X}_d(i)$ approaches zero. The gc between Fig. 9(a) and Fig. 9(b) is 0.6477, indicating the gradient increase induced by the distortion, while the one between Fig. 9(c) and Fig. 9(d) is 1.6167 indicating the gradient decrease induced by the distortion.

After obtaining gc, we further adjust the statistics to mitigate overestimation of the degradation degree. With the principle that the (overestimated) degradation represented by the (large) RD/SD and the (small) mean/median should be lessened if the gradient magnitudes increase ($gc < 1$), and recall that the local scores and their four statistics are all in $[0, 1]$, we simply modify the statistics as

$$\begin{gathered} SD' = SD^{1/gc},\;\; RD' = RD^{1/gc}, \\ mean' = (mean)^{gc},\;\; median' = (median)^{gc}. \end{gathered} \tag{8}$$

By using Equation (8), overestimation of the degradation is alleviated, because the SD or RD declines while the mean or median rises if gc is less than 1. Note that lower SD (or RD) and higher mean (or median) indicate better image quality.

2.4. Integration of the adjusted statistics

To obtain a single overall quality score, we need to integrate multiple sample statistics, as discussed in Section 2.2.2. Specifically, there are two principles that we apply here. First, the mean and median should only play a subsidiary role in the quantification of the degradation, and the SD or RD should dominate in the final quality score of the distorted image. Second, the excess kurtosis, which serves as a descriptor of the outlier ratio, is employed to determine the relative importance of robust statistics and non-robust statistics when combining them. Mean and median, as we discussed before, are complementary to RD and SD in the pooling step. However, both mean and median assume that all local regions contribute equally to the perception, which is at odds with the HVS [27, 28]. Therefore, we think mean and median should not dominate the overall score of the image quality. The proposed integration is divided into two steps. The first step is to integrate the measures of central tendency and dispersion, i.e. to combine the SD and mean (or the RD and median). The second step is to integrate the non-robust statistics (SD, mean) and robust statistics (RD, median).

2.4.1. Integration of the central tendency and the dispersion

A widely-used statistics for measuring data dispersion is the coefficient of variation (CV), defined as the ratio of SD to mean [57]. The CV has been used in the no-reference IQA as a feature in the statistical model [58, 59] and a pooling strategy in video quality assessment [60]. In this paper, the integration of central tendency and dispersion is inspired by the definition of CV, such that two distorted images with equal values of dispersion (SD or RD) but different central tendencies (mean or median), or nearly the same central tendency but entirely different dispersions, can be well discriminated.

However, directly using the definition of CV for integration, i.e. $f_1 = SD'/mean'$ and $f_2 = RD'/median'$, is inappropriate due to two undesirable properties of the CV when applied to IQA. Take f_1 as an example. First, the range of f_1 is from 0 to infinity. Second, the derivative of f_1 with respect to the mean is $-SD'/(mean')^2$, implying that f_1 is too sensitive to the mean when the mean value approaches 0. This sensitivity suggests that the mean would dominate the final quality score, which is undesirable.

Hence, to overcome the above two problems, we propose to use $(SD')^{mean'}$ and $(RD')^{median'}$ for the integration of central tendency and dispersion. It is clear that the ranges of the proposed $(SD')^{mean'}$ and $(RD')^{median'}$ are $[0,1]$. Moreover, they are less sensitive to the mean and median, since the derivative with respect to the mean or median is $\ln(SD')$ or $\ln(RD')$ when the mean or median approaches 0.

2.4.2. Integration of the robust and non-robust statistics

After robust statistics ($(RD')^{median'}$) and non-robust statistics ($(SD')^{mean'}$) are calculated, we integrate them using a weighted sum. As we mentioned earlier, the non-robust statistics need to be heavily weighted if the ratio of outliers is large. In this paper, we propose to use the excess kurtosis for determining the weights.

The kurtosis is the fourth standardized moment and it describes the tail of the distribution. Higher kurtosis results from the extreme deviations (outliers) and therefore we can use the kurtosis to indicate the ratio of outliers. The kurtosis for any univariate normal distribution is 3 and we compare the kurtosis of local quality scores to this value by subtracting 3 from the kurtosis, which is the excess kurtosis. If the excess kurtosis is less than 0, it means that the distribution of local quality scores produces fewer and less extreme outliers than does the normal distribution [61]. The excess kurtosis is written as

$$K = \frac{\frac{1}{n}\sum_i (S(i)-m)^4}{(\frac{1}{n}\sum_i (S(i)-m)^2)^2} - 3, \tag{9}$$

where m represents the mean of local quality scores.

Then the weights assigned to the robust statistics and non-robust statistics are determined by a function of excess kurtosis. We design the function that determines the weight for the robust statistics based on two generic principles: the range of function should be $[0,1]$ and is a decreasing function of K. Specifically, the weight w for robust statistics is designed as

$$w = \frac{1}{1+e^{\lambda K}}, \tag{10}$$

where λ is constant. (10) is a decreasing logistic function of K and constrains the weight to the range $[0,1]$, and λ is used to control the decreasing rate of the logistic function.

We integrate the robust statistics and non-robust statistics as

$$\begin{aligned} V_g &= (1-w)(std'_g)^{mean'_g} + w(RD'_g)^{median'_g}, \\ V_C &= (1-w)(std'_C)^{mean'_C} + w(RD'_C)^{\alpha\ median'_C}, \end{aligned} \tag{11}$$

where the subscript stands for the statistics computed from different channels: the subscript g represents the luminance channel and C stands for the I channel or Q channel. The parameter α in (11) adjusts the dynamic ranges of the two terms in the chromatic components.

Finally, a single overall quality score S_o can be obtained as

$$S_o = \gamma V_g + \frac{1-\gamma}{2}(V_I + V_Q), \tag{12}$$

where $0 \leq \gamma \leq 1$ representing relative importance of the distortion in different components.

3. Experimental results and analysis

3.1. Databases and performance measures

In all databases, the subjective scores are provided in the form of either the MOS or the differential mean opinion score (DMOS). The nonlinear quality rating compression in subjective testing generally results in the nonlinear relationship between the subjective scores and the objective scores predicted by IQA models [62]. Typically the nonlinearity in the predicted scores is removed before the comparison with the subjective scores by applying regression analysis to the original IQA scores. We use the logistic function to perform the regression:

$$p(S_o) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 - \exp(\beta_2(S_o - \beta_3))} \right) + \beta_4 S_o + \beta_5, \tag{13}$$

where β_1, β_2, β_3, β_4 and β_5 are regression parameters. We adopt the method in [2] to initialize the parameters in (13) to avoid local optimal solutions.

After regression, we compare mapped values with subjective scores to evaluate IQA methods. Prediction monotonicity and prediction accuracy are two main measures for such evaluation. Here we use the Spearman rank order correlation coefficient (SROCC) and the Kendall rank order correlation coefficient (KROCC) to assess the prediction monotonicity. The Pearson

Table 2: Basic information of benchmark databases

Database	#(reference images)	#(distorted images)	#(distortions)	Subjective score
LIVE	29	779	5	DMOS
LIVEMD	15	450	3	DMOS
CSIQ	30	866	6	DMOS
TID2008	25	1700	17	MOS
TID2013	25	3000	24	MOS

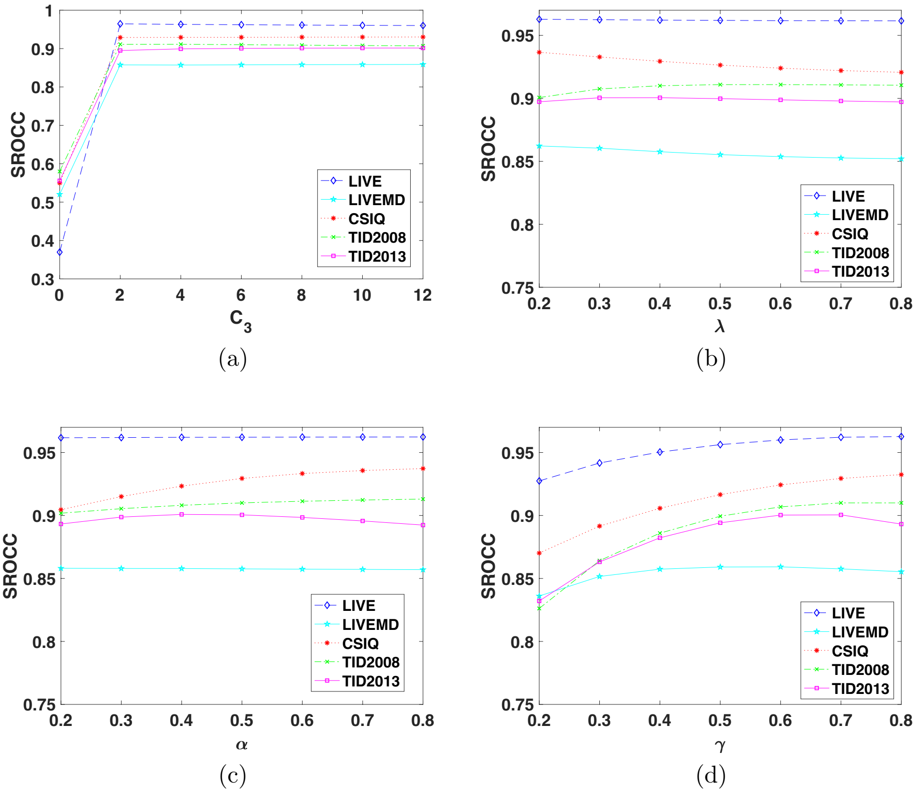
linear correlation coefficient (PLCC) and the root mean square error (RMSE) are used to evaluate the prediction accuracy.

We evaluate the proposed ASSP strategy on five publicly available image databases, namely LIVE [5], LIVEMD [63], CSIQ [6], TID2008 [7] and TID2013 [8], the basic information of which is given in Table 2. Each image in these databases has been assessed by human subjects under controlled conditions, and subjective scores (MOS or DMOS) are assigned to the distorted images.

In order to investigate the power of the proposed ASSP method, we compare it with two groups of state-of-the-art IQA metrics. The IQA models in the first group use handcrafted features for measuring the degradation, including the noise quality metric (NQM) [64], SSIM [10], MS-SSIM [11], FSIMc [15], IW-SSIM [12], VIF [25], IFC [24], GSM [14], GMSD [32], LLM [19] and SPSIM [21]. It is worthwhile mentioning that various IQA models in the first group use different pooling strategies: several methods adopt the weighted mean for pooling and use different features for calculating weights, including FSIMc with the phase congruency used for computing spatially-varying weights, SPSIM with the texture complexity, and IW-SSIM with the information content; GMSD uses standard deviation as a pooling strategy. The second group of models that are taken into account are deep learning approaches that require a training step for learning similarity maps and pooling weights, including WaDIQaM-FR [65] and RADN [66]. We compare ASSP with the deep learning models in terms of their cross-database performances. Note that the experimental results of LLM on LIVEMD are not shown since the code of LLM is not available.

Before using ASSP to quantify the degradation degree, we convert RGB images into the YIQ colour space, and follow the suggestion in [67] to determine a proper scale, which is to average local $F \times F$ pixels and downsample the image by a factor of F, with F given by

$$F = \max(1, \mathrm{round}(N/256)), \tag{14}$$



(a) (b) (c) (d)

Figure 10: The sensitivity to the parameters in terms of SROCC on four databases. (a) SROCC vs. C_3. (b) SROCC vs. λ. (c) SROCC vs. α. (d) SROCC vs. γ.

where N is the image height or width.

3.2. Investigation of parameters

There are several parameters in the proposed ASSP. In order to investigate the sensitivity of these parameters, a number of experiments are performed. In Section 2, we ignore the parameter subscripts that represent the channels for brevity except for Equations (11) and (12).

The sensitivity of parameters is shown in Fig. 10, where various values for a given parameter are used in the experiments, while the values of the other parameters are set as default values, which we shall discuss soon. As we mentioned earlier, C_3 is added to ensure the numerical stability particularly in low gradient regions. Fig. 10(a) shows that the performances of ASSP are consistent on all databases when $C_3 > 2$. Meanwhile, the performances achieved by various values of λ are also consistent across different databases as indicated in Fig. 10(b) except for CSIQ and LIVEMD, where slightly higher performance is achieved with lower λ. As shown in Fig. 10(c), there is no preference for the value of α for LIVE and LIVEMD. The curve for TID2008 in Fig. 10(c) shows a subtle growth with α increasing. For CSIQ, the corresponding curve also increases with a larger α, and the growth is larger when $\alpha > 0.5$. Meanwhile, the preferred value of α for TID2013 is around 0.5.

Regarding the performances obtained by setting different values for γ as shown in Fig. 10(d), the trends of performances for LIVE, LIVEMD, CSIQ and TID2008 are consistent in the sense that all three curves grow with γ increasing. The only difference between TID2013 and others lies in the range $[0.7, 0.8]$, where the performance on TID2013 decreases while the others increase. The different pattern observed in TID2013 results from different distortion types contained in these databases. Specifically, there is no colour-related distortion type in LIVE, LIVEMD, CSIQ and TID2008, which means that we only need to consider the distortion in the luminance channel. Thus the the proposed quality measure is only impacted by the distortion in the luminance channel as γ approaches 1 without being affected by other channels. For TID2013, there are colour-related distortion types and hence we need to consider the luminance and chromatic channels simultaneously for comprehensively measuring the distortion. When γ approaches 1, it leads to a measure tailored for the degree of degradation focusing on the distortion in the luminance channel and vice versa for γ close to 0. Hence the advantageous performance is achieved in the range $[0.5, 0.7]$ for TID2013. Regarding the choice of parameters in practice, we suggest that the weight γ for luminance channel should be higher than that for chromatic channels.

In summary, the proposed ASSP shows relatively good performance when $\gamma \in [0.5, 0.7]$, $C_3 > 2$, $\alpha \in [0.5, 0.8]$, and $\lambda \in [0.2, 0.6]$. In practice, the parameters required in the proposed method are set as $C_1 = 160$, $C_2 = 200$, $C_3 = 6$, $\gamma = 0.7$, $\lambda = 0.4$ and $\alpha = 0.5$. Since the local quality scores S_g, S_I and S_Q that we use have already been applied successfully in [15], we use the same values of C_1 and C_2 as those in [15] without tuning.

Table 3: SROCC between RD/IQR in different channels and the subjective scores provided by the selected datasets. Higher SROCC values for a channel are shown in bold.

Statistics	LIVE	LIVEMD	CSIQ	TID2013
RD(Y)	0.9503	0.8465	**0.9481**	**0.6437**
IQR(Y)	**0.9524**	**0.8621**	0.9181	0.6279
RD(I)	**0.6531**	**0.3320**	**0.7289**	0.3993
IQR(I)	0.6410	0.3078	0.7020	**0.4022**
RD(Q)	**0.6214**	**0.2277**	**0.6766**	**0.4068**
IQR(Q)	0.6030	0.1967	0.6353	0.4051

Table 4: Comparison of performances on benchmark databases. The top three models for each criterion are shown in boldface

Database	Criterion	NQM	SSIM	MS-SSIM	FSIMc	IW-SSIM	VIF	IFC	GSM	GMSD	LLM	SPSIM	ASSP
LIVE	SROCC	0.9086	0.9479	0.9513	**0.9645**	0.9567	**0.9636**	0.9259	0.9561	0.9603	0.9608	0.9620	**0.9621**
	KROCC	0.7413	0.7963	0.8045	**0.8363**	0.8175	**0.8282**	0.7579	0.8150	0.8254	0.8230	0.8271	**0.8275**
	PLCC	0.9122	0.9449	0.9489	**0.9613**	0.9522	**0.9604**	0.9268	0.9512	**0.9603**	0.9578	0.9599	0.9599
	RMSE	11.1926	8.9455	8.6188	**7.5296**	8.3473	**7.6137**	10.2643	8.4327	**7.6247**	7.7678	7.6288	7.6636
LIVEMD	SROCC	**0.8999**	0.8604	0.8363	0.8665	**0.8836**	0.8823	**0.8840**	0.8453	0.8448	-	0.8578	0.8576
	KROCC	**0.7208**	0.6692	0.6439	0.6765	**0.7014**	0.6966	**0.7034**	0.6547	0.6545	-	0.6689	0.6640
	PLCC	**0.9086**	0.8914	0.8747	0.8964	**0.9109**	0.9030	**0.9058**	0.8808	0.8809	-	0.8866	0.8852
	RMSE	**7.9007**	8.6025	9.2021	8.4760	**7.8173**	8.1235	**8.0129**	8.9559	8.9520	-	8.9213	8.8097
CSIQ	SROCC	0.7402	0.8756	0.9133	**0.9310**	0.9213	0.9195	0.7671	0.9108	**0.9572**	0.9050	**0.9440**	0.9294
	KROCC	0.5638	0.6907	0.7393	**0.7690**	0.7529	0.7537	0.5897	0.7374	**0.8134**	0.7238	**0.7880**	0.7663
	PLCC	0.7433	0.8613	0.8991	**0.9192**	0.9144	0.9277	0.8384	0.8964	**0.9542**	0.9000	**0.9344**	0.9137
	RMSE	0.1756	0.1334	0.1149	**0.1034**	0.1063	0.0980	0.1431	0.1164	**0.0786**	0.1232	**0.0934**	0.1067
TID2008	SROCC	0.6243	0.7749	0.8542	0.8840	0.8559	0.7491	0.5675	0.8504	0.8906	**0.9077**	**0.9104**	**0.9100**
	KROCC	0.4608	0.5768	0.6568	0.6991	0.6636	0.5860	0.4236	0.6596	0.7090	**0.7368**	**0.7303**	**0.7316**
	PLCC	0.6142	0.7732	0.8451	0.8762	0.8579	0.8084	0.7340	0.8422	0.8717	**0.8971**	**0.8927**	**0.9032**
	RMSE	1.0590	0.8511	0.7173	0.6468	0.6895	0.7899	0.9113	0.7235	0.6575	**0.5982**	**0.6046**	**0.5759**
TID2013	SROCC	0.6432	0.7417	0.7859	0.8510	0.7779	0.6769	0.5389	0.7946	0.8038	**0.9037**	**0.9044**	**0.9005**
	KROCC	0.4740	0.5588	0.6047	0.6665	0.5977	0.5147	0.3939	0.6255	0.6334	**0.7209**	**0.7251**	**0.7239**
	PLCC	0.6904	0.7895	0.8329	0.8769	0.8319	0.7720	0.5538	0.8464	0.8594	**0.9068**	**0.9091**	**0.9138**
	RMSE	0.8969	0.7608	0.6861	0.5959	0.6880	0.7880	1.0322	0.6603	0.6339	**0.5277**	**0.5165**	**0.5034**
Weighted Avg.	SROCC	0.6983	0.7986	0.8415	**0.8835**	0.8432	0.7724	0.6424	0.8475	0.8657	-	**0.9145**	**0.9108**
	KROCC	0.5291	0.6147	0.6604	**0.7078**	0.6660	0.6110	0.4885	0.6720	0.6987	-	**0.7424**	**0.7391**
	PLCC	0.7180	0.8191	0.8605	**0.8931**	0.8679	0.8312	0.7012	0.8660	0.8696	-	**0.9126**	**0.9145**
Direct Avg.	SROCC	0.7632	0.8401	0.8682	**0.8994**	0.8791	0.8383	0.7367	0.8732	0.8913	-	**0.9157**	**0.9119**
	KROCC	0.5921	0.6584	0.6898	**0.7295**	0.7066	0.6758	0.5737	0.6984	0.7271	-	**0.7479**	**0.7427**
	PLCC	0.7737	0.8521	0.8801	**0.9060**	0.8935	0.8743	0.7918	0.8834	0.8771	-	**0.9165**	**0.9152**

3.3. Comparing RD and IQR

As we suggested in Section 2.2.1, RD presented in adjusted boxplots is better suited for measuring dispersion compared to IQR in regular boxplots since the skewness of distributions is not taken into account in IQR. To further show the superiority of using RD rather than IQR, we compute the SROCC between RD/IQR in different channels and the subjective scores provided by the selected databases and provide the results in Table 3. As expected, RD achieves higher SROCC in most cases, indicating RD is a better measure than IQR.

3.4. Performance comparison on benchmark databases

We first examine the performances of IQA models in the first group. The results in terms of SROCC, KROCC, PLCC and RMSE on the five databases are given in Table 4 and the top three models for each criterion are shown in boldface. The number of times being among the top 3 models is: SPSIM (12 times), ASSP (10 times), LLM (8 times), FSIMc (8 times), GMSD (6 times) VIF (4 times), NQM (4 times), IW-SSIM (4 times), and IFC (4 times). TID2008 and TID2013 are two large-scale datasets and provide distortion images with various distortion types. ASSP performs the best in terms of PLCC and RMSE on both TID2008 and TID2013. ASSP is also in the top 3 models on LIVE in terms of SROCC and KROCC. Although the performance of ASSP on CSIQ is superior to LLM, it is less satisfying than SPSIM. Furthermore, the superiority of ASSP indicates that compared with GMSD demonstrates that our sample statistics-based pooling strategy performs much better than the simple SD pooling.

We also present the weighted average and the direct average of SROCC, KROCC and PLCC in Table 4. The weighted average and the direct average of RMSE are not provided because the range of RMSE for different databases varies. The weights in the weighted average are determined by the number of distorted images in the four databases. Although SPSIM achieves the largest number of being among the top 3 models, it does not imply that SPSIM is better than ASSP since the criteria based on the weighted average and direct average are roughly the same for both models. To sum up, we believe ASSP obtain comparable performance across different datasets if it is not superior to SPSIM.

3.5. Performance comparison on individual distortion types

Generally speaking, a good IQA model should not only perform well on databases, but also perform consistently across different distortion types. Therefore we evaluate IQA models in the first group on each distortion type individually for comprehensively assessing our IQA model. The results are listed in Table 5. To save the space, we use the abbreviation to denote distortion types and only provide SROCC for evaluation. We compare different IQA models on TID2013 since its number of distortion types is the largest compared with the other datasets. In Table 5, ASSP is among top 3 models 17 times, followed by LLM (15 times), GMSD (14 times), GSM (8 times), SPSIM (6 times), FSIMc (4 times), VIF (4 times), MS-SSIM(3 times), and IW-SSIM (1 time). It is worthwhile mentioning that, although SPSIM is

Table 5: SROCC performance comparison on each individual distortion type on TID2013 with the top 3 models highlighted in boldface for each distortion type

Type	NQM	SSIM	MS-SSIM	FSIMc	IW-SSIM	VIF	IFC	GSM	GMSD	LLM	SPSIM	ASSP
AGWN	0.8171	0.8671	0.8646	0.9101	0.8438	0.8994	0.6612	0.9064	**0.9462**	**0.9462**	0.9273	**0.9469**
ACN	0.7403	0.7726	0.7730	0.8537	0.7515	0.8299	0.5352	0.8175	**0.8684**	**0.8975**	0.8522	**0.8700**
SCN	0.7880	0.8515	0.8544	0.8900	0.8167	0.8835	0.6601	0.9158	**0.9350**	**0.9349**	0.9136	**0.9340**
MN	0.6852	0.7767	**0.8073**	**0.8094**	0.8020	**0.8450**	0.6932	0.7293	0.7075	0.7545	0.7543	0.7673
HFN	0.8700	0.8634	0.8604	0.9040	0.8553	0.8972	0.7406	0.8869	**0.9162**	**0.9524**	0.9062	**0.9220**
IN	0.7907	0.7503	0.7629	0.8251	0.7281	**0.8537**	0.6408	0.7965	0.7637	**0.8326**	0.8205	**0.8386**
QN	0.8261	0.8657	0.8706	0.8807	0.8468	0.7854	0.6282	0.8841	**0.9049**	**0.9055**	0.8897	**0.8957**
GB	0.9005	0.9668	**0.9673**	0.9551	**0.9701**	0.9650	0.8907	**0.9689**	0.9113	0.9451	0.9527	0.9363
ID	0.9195	0.9254	0.9268	0.9330	0.9152	0.8911	0.7779	0.9432	**0.9525**	**0.9478**	**0.9480**	0.9440
JPEG	0.8765	0.9200	0.9265	0.9339	0.9187	0.9192	0.8357	0.9284	**0.9507**	**0.9544**	0.9374	**0.9521**
JP2K	0.9269	0.9468	0.9504	0.9589	0.9506	0.9516	0.9078	0.9602	**0.9657**	**0.9702**	**0.9662**	0.9638
JGTE	0.7321	0.8493	0.8475	**0.8610**	0.8388	0.8409	0.7425	**0.8512**	0.8403	0.8459	0.8660	**0.8676**
J2TE	0.8068	0.8828	0.8889	0.8919	0.8656	0.8761	0.7769	**0.9182**	0.9136	**0.9176**	0.9119	**0.9157**
NEPN	0.7463	0.7821	0.7968	0.7937	0.8011	0.7720	0.5737	**0.8130**	**0.8140**	0.7967	0.8098	**0.8231**
LBWD	0.0064	0.5720	0.4801	0.5532	0.3717	0.5306	0.2414	**0.6418**	**0.6625**	0.6273	0.1887	**0.6932**
MS	0.6092	0.7752	**0.7906**	0.7487	0.7833	0.6276	0.5522	**0.7875**	0.7351	0.7586	**0.7846**	0.7156
CTC	0.4623	0.3775	0.4634	0.4679	0.4593	**0.8386**	0.1798	0.4857	0.3235	0.4634	**0.7148**	**0.7788**
CCS	0.1591	0.4141	0.4099	**0.8359**	0.4196	0.3099	0.4029	0.3578	0.2948	0.3117	**0.7913**	**0.8337**
MGN	0.7727	0.7803	0.7786	0.8569	0.7728	0.8468	0.6143	0.8348	**0.8886**	**0.9097**	0.8620	**0.8948**
CN	0.8755	0.8566	0.8528	0.9135	0.8762	0.8946	0.8160	0.9124	**0.9298**	**0.9455**	0.9144	**0.9296**
LCNI	0.9064	0.9057	0.9068	0.9485	0.9037	0.9204	0.8180	0.9563	**0.9629**	**0.9588**	0.9500	**0.9564**
ICQD	0.8635	0.8542	0.8555	0.8815	0.8401	0.8414	0.6006	0.8973	**0.9102**	**0.9155**	0.9063	**0.9180**
CA	0.8174	0.8775	0.8784	**0.8925**	0.8682	**0.8848**	0.8210	**0.8823**	0.8530	0.8682	0.8784	0.8821
SSR	0.9468	0.9461	0.9483	0.9576	0.9474	0.9353	0.8885	**0.9668**	0.9638	**0.9676**	**0.9642**	0.9608

Table 6: SROCC comparison in cross-database evaluation. The best model on each test database is highlighted in boldface

Trained/Tuned on	LIVE			LIVEMD			CSIQ			TID2013		
Tested on	LIVEMD	CSIQ	TID2013	LIVE	CSIQ	TID2013	LIVE	LIVEMD	TID2013	LIVE	LIVEMD	CSIQ
WaDIQaM-FR	0.8644	0.9090	0.7510	0.4432	0.4350	0.2672	**0.9661**	**0.8927**	0.7872	0.9360	0.7744	0.9310
RADN	**0.8764**	0.8233	0.6390	0.8202	0.7544	0.5280	0.9253	0.8006	0.5995	0.9443	0.8564	0.9179
ASSP	0.8611	**0.9435**	**0.8812**	**0.9624**	**0.9388**	**0.8900**	0.9637	0.8619	**0.8800**	**0.9599**	**0.8588**	**0.9440**

competitive with ASSP, its performance within each distortion type is inferior to that of ASSP.

3.6. Cross-database evaluation of deep learning models and ASSP

To examine the generalization ability of IQA models based on deep learning. We train WaDIQaM-FR/RADN on one database and evaluate the performance of the trained model on the other databases. Likewise, we assess the generalization ability of ASSP by first tuning its hyperparameters on one dataset and then examining its performance on the other datasets. Note that TID2008 is excluded for this comparison since TID2008 and TID2013 have the same references images, and performing cross-data evaluation between these two databases is inappropriate. Table 6 shows the cross-database performances of deep learning models and ASSP in terms of SROCC. It is observed that ASSP performs best in most combinations. When trained on a smaller database, such as LIVEMD, both WaDIQaM-FR and RADN exhibit rather poor generalization performance. Furthermore, using TID2013 which is the largest database as a training dataset for WaDIQaM-FR and RADN does not lead to the cross-database performance better than ASSP. The poor generalization ability shown by WaDIQaM-FR and RADN is expected as deep learning models often require quite large databases for training. Compared with the IQA models based on deep learning, ASSP shows good generalization ability and perform consistently well cross databases.

3.7. Proposed statistics-based pooling for other IQA models

To further investigate the effectiveness and applicability of ASSP, we incorporate it into several other IQA models: FSIMc, SSIM, GMSD, and MSE (which is equivalent to PSNR but can get local quality scores). For SSIM and MSE, we do not use the gradient information to adjust the statistics, since the gradient information is not included in the extracted features from them. For FSIMc and GMSD, we implement the adjustment.

Table 7: SROCC of the proposed pooling on other IQA models

	Original Pooling				Proposed Pooling				Weighted Average	
Database	LIVE	LIVEMD	CSIQ	TID2013	LIVE	LIVEMD	CSIQ	TID2013	Original Pooling	Proposed Pooling
SSIM	0.9479	0.8604	0.8756	0.7417	0.9496	0.8610	0.8476	0.7544	0.8065	0.8095
FSIMc	0.9645	0.8665	0.9310	0.8510	0.9626	0.8594	0.9354	0.9016	0.8833	0.9129
GMSD	0.9603	0.8448	0.9572	0.8038	0.9601	0.8538	0.9315	0.8136	0.8574	0.8596
MSE	0.9197	0.7459	0.8178	0.6340	0.9197	0.7435	0.8197	0.6273	0.7188	0.7150

Performance comparison between the original pooling of those IQA models and the proposed pooling is shown in Table 7. We can find that the performances of SSIM, FSIMc and GMSD are elevated by using our proposed pooling strategy in terms of the weighted average over the databases (see the rightmost block of Table 7). However, the results are worse MSE when using the proposed pooling strategy. This may be because the proposed strategy is designed based on the gradient similarity and the chromatic similarity, better suited for the local quality scores involving the gradient/chromatic features. It is also worth mentioning that FSIMc and SSIM adopt diverse types of features for calculating local quality scores. The improvement achieved by applying the proposed pooling strategy to FSIMc and SSIM may indicate that the proposed statistics-based pooling can account for the interaction of diverse features.

3.8. Computational complexity

The computational complexity of an implemented IQA model is instrumental in real-time applications. Table 8 presents the running time of several IQA models on a 512 × 512 image. All algorithms were run on a MacBook Pro with 2.8 GHz Intel Core i5 and 8 GB 1600 MHz DDR3. The software platform is Matlab R2017b. GMSD shows the fastest speed but its performance is inferior to that of the state-of-the-art methods. On the contrary, the proposed ASSP exhibits competitive performance and is faster than NQM, FSIMc, IW-SSIM, VIF, IFC, and SPSIM.

Table 8: Comparison of running time in seconds

Method	NQM	SSIM	MS-SSIM	FSIMc	IW-SSIM	VIF	IFC	GSM	GMSD	LLM	SPSIM	ASSP
Running time (s)	0.2539	0.0154	0.0600	0.4287	0.5095	1.0670	1.0780	0.0239	0.0090	-	0.1925	0.1380

4. Conclusion

This paper proposes a new sample statistics-based pooling strategy called ASSP for FR-IQA. We show that the median and RD are correlated with the HVS and can provide complementary information on the degree of distortion to that provided by the mean and SD. We integrate mean, SD, median and RD in an innovative way to obtain overall IQA score. The ASSP is mostly superior to the state-of-the-art IQA models compared, in terms of both mean and weighted mean of different evaluation criteria.

Table 7 shows that although the proposed pooling boosts the performance of several representative FR-IQA models (FSIMc, SSIM and GMSD), it fails to do so when applied to simple features which do not take the gradient/chromatic similarity into account. We hence think the proposed strategies are better suited for local quality scores that involve gradient magnitudes or chromatic information. Meanwhile, the failure of boosting MSE may result from the fact that the distributions of their local quality scores are quite different from that of the local quality scores adopted in this paper. Since the proposed approach is motivated by the distributional characteristics of local quality scores, such as the skewness and outliers. If the distributions of local quality scores are not skewed, using adjusted boxplots, as we did in the proposed pooling, to determine the robust dispersion and outliers may not be appropriate.

Our investigation so far focuses on one of the robust measures of the dispersion: RD. However, there are many alternatives that can also exclude the influence of outliers to some extent, such as the IQR and the interdecile range. The use of other robust measures of the dispersion and their correlation with the HVS will be one of our future work. In addition, in the current framework the values of parameters are determined empirically; hence investigating how to guide the learning of parameters by linking them to the HVS will be another of our future work.

5. Acknowledgement

We would like to thank Mr. Junhua Chen for his help on drawing Fig. 1 and Fig. 2.

References

[1] G. Zhai, X. Min, Perceptual image quality assessment: a survey, Science China Information Sciences 63 (2020) 1–52.

[2] W. Sun, F. Zhou, Q. Liao, MDID: A multiply distorted image database for image quality assessment, Pattern Recognition 61 (2017) 153 – 168.

[3] W. Liu, F. Zhou, T. Lu, J. Duan, G. Qiu, Image defogging quality assessment: Real-world database and method, IEEE Transactions on Image Processing 30 (2021) 176–190.

[4] F. Zhou, R. Yao, B. Liu, G. Qiu, Visual quality assessment for super-resolved images: Database and method, IEEE Transactions on Image Processing 28 (2019) 3528–3541.

[5] H. Sheikh, Z. Wang, L. Cormack, A. Bovik, LIVE image quality assessment database release 2, 2005. URL: http://live.ece.utexas.edu/research/quality.

[6] E. C. Larson, D. Chandler, Categorical image quality (CSIQ) database, 2010. URL: http://vision.okstate.edu/csiq.

[7] 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 (2009) 30–45.

[8] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. J. Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Processing: Image Communication 30 (2015) 57–77.

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

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[11] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: Asilomar Conference on Signals, Systems Computers, volume 2, 2003, pp. 1398–1402.

[12] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Transactions on Image Processing 20 (2011) 1185–1198.

[13] R. Zhu, F. Zhou, J.-H. Xue, MvSSIM: A quality assessment index for hyperspectral images, Neurocomputing 272 (2018) 250–257.

[14] A. Liu, W. Lin, M. Narwaria, Image quality assessment based on gradient similarity, IEEE Transactions on Image Processing 21 (2012) 1500–1512.

[15] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Transactions on Image Processing 20 (2011) 2378–2386.

[16] Z. Liu, R. Laganière, Phase congruence measurement for image similarity assessment, Pattern Recognition Letters 28 (2007) 166–172.

[17] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (2010) 011006.

[18] M. Narwaria, W. Lin, A. E. Cetin, Scalable image quality assessment with 2D mel-cepstrum and machine learning approach, Pattern Recognition 45 (2012) 299–313.

[19] H. Wang, J. Fu, W. Lin, S. Hu, C. C. J. Kuo, L. Zuo, Image quality assessment based on local linear information and distortion-specific compensation, IEEE Transactions on Image Processing 26 (2017) 915–926.

[20] A. Ahar, A. Barri, P. Schelkens, From sparse coding significance to perceptual quality: A new approach for image quality assessment, IEEE Transactions on Image Processing 27 (2018) 879–893.

[21] W. Sun, Q. Liao, J.-H. Xue, F. Zhou, SPSIM: A superpixel-based similarity index for full-reference image quality assessment, IEEE Transactions on Image Processing 27 (2018) 4232–4244.

[22] W. Kim, A.-D. Nguyen, S. Lee, A. C. Bovik, Dynamic receptive field generation for full-reference image quality assessment, IEEE Transactions on Image Processing 29 (2020) 4219–4231.

[23] M. Liu, L.-M. Po, X. Xu, K. W. Cheung, Y. Zhao, K. W. Lau, C. Zhou, Long-range dependencies and high-order spatial pooling for deep model-based full-reference image quality assessment, IEEE Access 8 (2020) 72007–72020.

[24] H. R. Sheikh, A. C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Transactions on Image Processing 14 (2005) 2117–2128.

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

[26] C.-Y. Wee, R. Paramesran, R. Mukundan, X. Jiang, Image quality assessment by discrete orthogonal moments, Pattern Recognition 43 (2010) 4055–4068.

[27] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.

[28] U. Rajashekar, I. van der Linde, A. C. Bovik, L. K. Cormack, GAFFE: A gaze-attentive fixation finding engine, IEEE Transactions on Image Processing 17 (2008) 564–573.

[29] L. Zhang, Y. Shen, H. Li, VSI: A visual saliency-induced index for perceptual image quality assessment, IEEE Transactions on Image Processing 23 (2014) 4270–4281.

[30] Z. Wang, X. Shang, Spatial pooling strategies for perceptual image quality assessment, in: International Conference on Image Processing, 2006, pp. 2945–2948.

[31] A. K. Moorthy, A. C. Bovik, Visual importance pooling for image quality assessment, IEEE Journal of Selected Topics in Signal Processing 3 (2009) 193–201.

[32] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Transactions on Image Processing 23 (2014) 684–695.

[33] F. Zhou, W. Yang, X. Gao, H. Liu, R. Zhu, J.-H. Xue, Special issue on advances in statistical methods-based visual quality assessment, Signal Processing: Image Communication 83 (2020) 115695.

[34] R. Zhu, F. Zhou, W. Yang, J.-H. Xue, On hypothesis testing for comparing image quality assessment metrics [tips & tricks], IEEE Signal Processing Magazine 35 (2018) 133–136.

[35] Z. Zeng, W. Yang, W. Sun, J.-H. Xue, Q. Liao, No-reference image quality assessment for photographic images based on robust statistics, Neurocomputing 313 (2018) 111–118.

[36] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, W. Zhang, Analysis of distortion distribution for pooling in image quality prediction, IEEE Transactions on Broadcasting 62 (2016) 446–456.

[37] L. Shen, X. Chen, Z. Pan, K. Fan, F. Li, J. Lei, No-reference stereoscopic image quality assessment based on global and local content characteristics, Neurocomputing 424 (2021) 132–142.

[38] X. Jiang, L. Shen, L. Yu, M. Jiang, G. Feng, No-reference screen content image quality assessment based on multi-region features, Neurocomputing 386 (2020) 30–41.

[39] D. Liang, X. Gao, W. Lu, J. Li, Deep blind image quality assessment based on multiple instance regression, Neurocomputing 431 (2021) 78–89.

[40] W. Xia, Y. Yang, J.-H. Xue, J. Xiao, Domain fingerprints for no-reference image quality assessment, IEEE Transactions on Circuits and Systems for Video Technology 31 (2021) 1332–1341.

[41] R. Maronna, R. D. Martin, V. Yohai, Robust statistics, John Wiley & Sons, Chichester. ISBN, 2006.

[42] I. E. Frank, R. Todeschini, The Data Analysis Handbook, volume 14, Elsevier, 1994.

[43] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Computational Statistics & Data Analysis 52 (2008) 5186–5201.

[44] Z. Wang, A. C. Bovik, Modern Image Quality Assessment, Morgan & Claypool, 2006.

[45] G. Cheng, J. Huang, C. Zhu, Z. Liu, L. Cheng, Perceptual image quality assessment using a geometric structural distortion model, in: IEEE International Conference on Image Processing, 2010, pp. 325–328.

[46] K. Gu, G. Zhai, X. Yang, W. Zhang, An efficient color image quality metric with local-tuned-global model, in: IEEE International Conference on Image Processing, 2014, pp. 506–510.

[47] S. H. K. Christopher C. Yang, Efficient gamut clipping for color image processing using LHS and YIQ, Optical Engineering 42 (2003) 701–711.

[48] J. Tukey, Exploratory Data Analysis, Addison-Wesley Publishing Company, CA, Menlo Park, 1977.

[49] G. Brys, M. Hubert, A. Struyf, A robust measure of skewness, Journal of Computational and Graphical Statistics 13 (2004) 996–1017.

[50] F. Röhrbein, P. Goddard, M. Schneider, G. James, K. Guo, How does image noise affect actual and predicted human gaze allocation in assessing image quality?, Vision Research 112 (2015) 11–25.

[51] M. S. Castelhano, J. M. Henderson, The influence of color on the perception of scene gist., Journal of Experimental Psychology: Human perception and performance 34 (2008) 660–675.

[52] A. B. Watson, A. J. Ahumada, Blur clarified: A review and synthesis of blur discrimination, Journal of Vision 11 (2011) 10–10.

[53] K. Guo, Y. Soornack, R. Settle, Expression-dependent susceptibility to face distortions in processing of facial expressions of emotion, Vision research (2018).

[54] J. Lu, D. Healy, Contrast enhancement via multiscale gradient transformation, in: IEEE International Conference on Image Processing, volume 2, IEEE, 1994, pp. 482–486.

[55] A. Polesel, G. Ramponi, V. J. Mathews, Image enhancement via adaptive unsharp masking, IEEE Transactions on Image Processing 9 (2000) 505–510.

[56] J. Tang, E. Peli, S. Acton, Image enhancement using a contrast measure in the compressed domain, IEEE Signal Processing Letters 10 (2003) 289–292.

[57] B. Everitt, A. Skrondal, The Cambridge dictionary of statistics, volume 106, Cambridge University Press Cambridge, 2002.

[58] M. A. Saad, A. C. Bovik, C. Charrier, A DCT statistics-based blind image quality index, IEEE Signal Processing Letters 17 (2010) 583–586.

[59] D. Kundu, D. Ghadiyaram, A. C. Bovik, B. L. Evans, No-reference image quality assessment for high dynamic range images, in: Asilomar Conference on Signals, Systems and Computers, IEEE, 2016, pp. 1847–1852.

[60] K. Seshadrinathan, A. C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Transactions on Image Processing 19 (2010) 335–350.

[61] P. H. Westfall, Kurtosis as peakedness, 1905–2014. R.I.P., The American Statistician 68 (2014) 191–195.

[62] Final report from the video quality experts group on the validation of objective models of video quality assessment VQEG, 2000. URL: `http://www.vqeg.org`.

[63] D. Jayaraman, A. Mittal, A. K. Moorthy, A. C. Bovik, Objective quality assessment of multiply distorted images, in: 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2012, pp. 1693–1697. doi:`10.1109/ACSSC.2012.6489321`.

[64] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, A. C. Bovik, Image quality assessment based on a degradation model, IEEE Transactions on Image Processing 9 (2000) 636–650.

[65] S. Bosse, D. Maniry, K. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on Image Processing 27 (2018) 206–219.

[66] S. Shi, Q. Bai, M. Cao, W. Xia, J. Wang, Y. Chen, Y. Yang, Region-adaptive deformable network for image quality assessment, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2021.

[67] Z. Wang, SSIM index for image quality assessment, 2003. URL: `http://www.ece.uwaterloo.ca/~z70wang/research/ssim/`.