Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator

Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet, Senior Member, IEEE

Abstract—Applications of perceptual image quality assessment (IQA) in image and video processing, such as image acquisition, image compression, image restoration and multimedia communication, have led to the development of many IQA metrics. In this paper, a reliable full reference IQA model is proposed that utilize gradient similarity (GS), chromaticity similarity (CS), and deviation pooling (DP). By considering the shortcomings of the commonly used GS to model human visual system (HVS), a new GS is proposed through a fusion technique that is more likely to follow HVS. We propose an efficient and effective formulation to calculate the joint similarity map of two chromatic channels for the purpose of measuring color changes. In comparison with a commonly used formulation in the literature, the proposed CS map is shown to be more efficient and provide comparable or better quality predictions. Motivated by a recent work that utilizes the standard deviation pooling, a general formulation of the DP is presented in this paper and used to compute a final score from the proposed GS and CS maps. This proposed formulation of DP benefits from the Minkowski pooling and a proposed power pooling as well. The experimental results on six datasets of natural images, a synthetic dataset, and a digitally retouched dataset show that the proposed index provides comparable or better quality predictions than the most recent and competing state-of-the-art IQA metrics in the literature, it is reliable and has low complexity. The MATLAB source code of the proposed metric is available at https://www.mathworks.com/matlabcentral/fileexchange/59809.

Index Terms—Image quality assessment, gradient similarity, chromaticity similarity, deviation pooling, synthetic image, Human visual system.

I. INTRODUCTION

AUTOMATIC image quality assessment (IQA) plays a significant role in numerous image and video processing applications. IQA is commonly used in image acquisition, image compression, image restoration, multimedia streaming, etc [1]. IQA models (IQAs) mimic the average quality predictions of human observers. Full reference IQAs (FR-IQAs), which fall within the scope of this paper, evaluate the perceptual quality of a distorted image with respect to its reference image. This quality prediction is an easy task for the human visual system (HVS) and the result of the evaluation is reliable. Automatic quality assessment, e.g. objective evaluation, is not an easy task because images may suffer from various types and degrees of distortions. FR-IQAs can be employed to compare two images of the same dynamic range (usually low dynamic range) [2] or different dynamic ranges [3, 4]. This paper is dedicated to the IQA for low dynamic range images.

Among IQAs, the conventional metric mean squared error (MSE) and its variations are widely used because of their simplicity. However, in many situations, MSE does not correlate with the human perception of image fidelity and quality [5]. Because of this limitation, a number of IQAs have been proposed to provide better correlation with the HVS [2], [6]–[24]. In general, these better performing metrics measure structural information, luminance information and contrast information in the spatial and frequency domains.

The most successful IQA models in the literature follow a top-down strategy [25]. They calculate a similarity map and use a pooling strategy that converts the values of this similarity map into a single quality score. For example, the luminance, contrast and structural information constitute a similarity map for the popular SSIM index [2]. SSIM then uses average pooling to compute the final similarity score. Different feature maps are used in the literature for calculation of this similarity map. Feature similarity index (FSIM) uses phase congruency and gradient magnitude features. GS [17] uses a combination of some designated gradient magnitudes and image contrast for this end, while the GMSD [19] uses only the gradient magnitude. SR_SIM [18] uses saliency features and gradient magnitude. VSI [21] likewise benefits from saliency-based features and gradient magnitude. SVD based features [26], features based on the Riesz transform [11], features in the wavelet domain [10], [27], [28] and sparse features [20] are used as well in the literature. Among these features, gradient magnitude is an efficient feature, as shown in [19]. In contrast, phase congruency and visual saliency features in general are not fast enough features to be used. Therefore, the features being used play a significant role in the efficiency of IQAs.

As we mentioned earlier, the computation of the similarity map is followed by a pooling strategy. The state-of-the-art pooling strategies for perceptual image quality assessment (IQA) are based on the mean and the weighted mean [2], [7], [11], [15]–[17]. They are robust pooling strategies that usually provide a moderate to high performance for different IQAs. Minkowski pooling [29], local distortion pooling [12], [29], [30], percentile pooling [31] and saliency-based pooling [18], [21] are other possibilities. Standard deviation (SD) pooling was also proposed and successfully used by GMSD [19]. The image gradients are sensitive to image distortions. Different local structures in a distorted image suffer from different degrees of degradations. This is the motivation that the authors in [19] used to explore the standard variation of the gradient-
based local similarity map for overall image quality prediction. In general, features that constitute the similarity map and the pooling strategy are very important factors in designing high performance IQA models.

Here, we propose an IQA model called the mean deviation similarity index (MDSI) that shows very good compromise between prediction accuracy and model complexity. The proposed index is efficient, effective and reliable at the same time. It also shows consistent performance for both natural and synthetic images. The proposed metric follows a top-down strategy. It uses gradient magnitude to measure structural distortions and use chrominance features to measure color distortions. These two similarity maps are then combined to form a gradient-chromaticity similarity map. We then propose a novel deviation pooling strategy and use it to compute the final quality score. Both image gradient [16], [17], [19], [21], [32], [33] and chrominance features [16], [21] have been already used in the literature. The proposed MDSI uses a new gradient similarity which is more likely to follow HVS. Also, MDSI uses a new chromaticity similarity map which is efficient and shows good performance when used with the proposed metric. The proposed index uses the summation over similarity maps to give independent weights to them. Also, less attention has been paid to the deviation pooling strategy, except for a special case of this type of pooling, namely, standard deviation pooling [19]. We therefore provide a general formulation for the deviation pooling strategy and show its power in the case of the proposed IQA model. In the following, the main contributions of the paper as well as its differences with respect to the previous works are briefly explained.

Unlike previous researches [16], [17], [19], [21], [32], [33] that use a similar gradient similarity map, a new gradient similarity map is proposed in this paper which is more likely to follow the human visual system (HVS). This statement is supported by visual examples and experimental results.

This paper proposes a new chromaticity similarity map with the following advantages over the previously used chromaticity similarity maps [16], [21]. Its complexity is lower and it provides slightly better quality predictions when used with the proposed metric.

Motivated by a previous study that proposed to use standard deviation pooling [19], we propose a systematic and general formulation of the deviation pooling which has a comprehensive scope.

The rest of the paper is organized as follows. The proposed mean deviation similarity index is presented in section II. Extensive experimental results and discussion on six natural datasets, a synthetic dataset, and a digitally retouched dataset are provided in section III. Section IV presents our conclusions.

II. MEAN DEVIATION SIMILARITY INDEX

The proposed IQA model uses two similarity maps. Image gradient, which is sensitive to structural distortions, is used as the main feature to calculate the first similarity map. Then, color distortions are measured by a chromaticity similarity map. These similarity maps are combined and pooled by a proposed deviation pooling strategy. In this paper, conversion to luminance is done through the following formula: $L = 0.2989R + 0.5870G + 0.1140B$. In addition, two chromaticity channels of a Gaussian color model [34] are used:

$$\begin{bmatrix} H \\ M \end{bmatrix} = \begin{bmatrix} 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

(1)

A. Gradient Similarity

It is very common that gradient magnitude in the discrete domain is calculated on the basis of some operators that approximate derivatives of the image function using differences. These operators approximate vertical $G_y(x)$ and horizontal $G_x(x)$ gradients of an image $f(x)$ using convolution: $G_x(x) = h_x \ast f(x)$ and $G_y(x) = h_y \ast f(x)$, where $h_x$ and $h_y$ are horizontal and vertical gradient operators and $\ast$ denotes the convolution. The first derivative magnitude is defined as $G(x) = \sqrt{G_x^2(x) + G_y^2(x)}$. The Sobel operator [35], the Scharr operator, and the Prewitt operator are common gradient operators that approximate first derivatives. Within the proposed IQA model, these operators perform almost the same.

Through this paper, Prewitt operator is used to compute gradient magnitudes of luminance $L$ channels of reference and distorted images, $R$ and $D$. From which, gradient similarity (GS) is computed by the following SSIM induced equation:

$$GS(x) = \frac{2G_R(x)G_D(x) + C_1}{G_R^2(x) + G_D^2(x) + C_1}$$

(2)

where, parameter $C_1$ is a constant to control numerical stability. The gradient similarity (GS) is widely used in the literature [16], [17], [19], [21], [32], [33] and its usefulness to measure image distortions was extensively investigated in [19].

In many scenarios, human visual system (HVS) disagrees with the judgments provided by the GS for structural distortions. In fact, in such a formulation, there is no difference between an added edge to or a removed edge from the distorted image with respect to the reference image. An extra edge in $D$ bring less attention of HVS if its color is close to the relative pixels of that edge in $R$. Likewise, HVS pays less attention to a removed edge from $R$ that is replaced with pixels of the same or nearly the same color. In another scenario, suppose that edges are preserved in $D$ but with different colors than in $R$. In this case, GS is likely to fail at providing a good judgment “on the edges”. These shortcomings of the GS motivated us to propose a new GS map.

B. The Proposed Gradient Similarity

The aforementioned shortcomings of the conventional gradient similarity map (equation 2) are mainly because $G_R$ and $G_D$ are computed independent of each other. In the following, we propose a fusion technique to include the correlation between $R$ and $D$ images into computation of the gradient similarity map.
We fuse the luminance $L$ channels of the $R$ and $D$ by a simple averaging: $F = 0.5 \times (R + D)$. Two extra GS maps are computed as follows:

$$\text{GS}_{RF}(x) = \frac{2G_R(x)G_F(x) + C_2}{G_R^2(x) + G_F^2(x) + C_2}$$  \hspace{1cm} (3)$$

$$\text{GS}_{DF}(x) = \frac{2G_D(x)G_F(x) + C_2}{G_D^2(x) + G_F^2(x) + C_2}$$  \hspace{1cm} (4)$$

where, $G_F$ is the gradient magnitude of the fused image $F$, and $C_2$ is used for numerical stability. Note that $G_F \neq (G_R + G_D)/2$, and that $\text{GS}_{RF}(x)$ and $\text{GS}_{DF}(x)$ can or cannot not be equal. The proposed gradient similarity ($\hat{\text{GS}}$) is computed by:

$$\hat{\text{GS}}(x) = \text{GS}(x) + [\text{GS}_{DF}(x) - \text{GS}_{RF}(x)].$$  \hspace{1cm} (5)$$

The added term $[\text{GS}_{DF}(x) - \text{GS}_{RF}(x)]$, will put more emphasize on removed edges from $R$ than added edges to the $D$. For weak added/removed edges, it is likely that weak edges smooth out in $F$. Therefore, $[\text{GS}_{DF}(x) - \text{GS}_{RF}(x)]$ always put less emphasize on weak edges.

Comparing visually some outputs of the GS and $\hat{\text{GS}}$ at this step might not be fair because they have different numerical scales. GS is bounded between 0 and 1, while $\hat{\text{GS}}$ might have negative values greater than -1, and/or positive values smaller than +2. Therefore, this comparison is performed on the final similarity map and is presented in subsection III-E as well as more explanation on how the proposed GS works.

C. Chromaticity Similarity

For the case of color changes and especially when the structure of the distorted image remains unchanged, the gradient similarity (GS) and the proposed $\hat{\text{GS}}$ may lead to inaccurate quality predictions. Therefore, previous researches such as [16], [21] used a color similarity map to measure color differences. Let $H$ and $M$ denote two chromaticity channels regardless of the type of the color space. In [16], [21], for each channel a color similarity is computed and their result is combined as:

$$\text{CS}(x) = \frac{2H_R(x)H_D(x) + C_4}{H_R^2(x) + H_D^2(x) + C_3} \times \frac{2M_R(x)M_D(x) + C_3}{M_R^2(x) + M_D^2(x) + C_3}$$  \hspace{1cm} (6)$$

where $C_4$ is a constant to control numerical stability. In this paper, we propose a new formulation to calculate color similarity. The proposed formulation calculates a color similarity map using both chromaticity channels at once:

$$\hat{\text{CS}}(x) = \frac{2(H_R(x)H_D(x) + M_R(x)M_D(x))^2 + C_3}{H_R^2(x) + H_D^2(x) + M_R^2(x) + M_D^2(x) + C_3}$$  \hspace{1cm} (7)$$

Similar to the CS in equation (6), the above joint color similarity ($\hat{\text{CS}}$) formulation gives equal weight to both chromaticity channels $H$ and $M$. It is clear that $\text{CS}$ is more efficient than $\text{CS}$. $\text{CS}$ needs 7 multiplications, 6 summations, 2 divisions, and 2 shift operations (multiplications by 2), while $\hat{\text{CS}}$ needs 6 multiplications, 6 summations, 1 division, and 1 shift operation. Note that $\text{CS}$ can also be computed through 8 multiplications, 6 summations, 1 division, and 2 shift operations. In experimental results section, an experiment is conducted to compare usefulness of the CS and $\hat{\text{CS}}$ along with the proposed metric.

The gradient similarity maps ($\text{GS}$ or $\hat{\text{GS}}$) can be combined with the joint color similarity map $\hat{\text{CS}}$ through the following summation (weighted average) scheme:

$$\text{GCS}(x) = \alpha \text{GS}(x) + (1 - \alpha)\hat{\text{CS}}(x)$$  \hspace{1cm} (8)$$

$$\hat{\text{GCS}}(x) = \alpha \hat{\text{GS}}(x) + (1 - \alpha)\hat{\text{CS}}(x)$$  \hspace{1cm} (9)$$

where the parameter $0 \leq \alpha \leq 1$ adjusts the relative importance of the gradient and chromaticity similarity maps. The proposed metric MDSI uses equation (8). Equation (5) is included to be compared with equation (9). An alternative combination scheme which is very popular in state-of-the-art is through multiplication in the form of $[\hat{\text{GS}}(x)]^\gamma[\hat{\text{CS}}(x)]^\beta$, where the parameters $\gamma$ and $\beta$ are used to adjust the relative importance of the two similarity maps. For several reasons, the proposed index uses the summation scheme (refer to subsection III-E).

In Fig 1 two examples are provided to show that these two similarity maps, e.g. GS and $\hat{\text{CS}}$, are complementary. In the first example, there is a considerable difference between the gradient maps of the reference and the distorted images. Hence, the GS map is enough for a good judgment. However, this difference in the second example (second row) is trivial,
which leads to a wrong prediction by using GS as the only similarity map. The examples in Fig 1 show that the gradient similarity and chromaticity similarity are complementary.

D. Deviation Pooling

The motivation of using the deviation pooling is that HVS is sensitive to both magnitude and the spread of the distortions across the image. Other pooling strategies such as Minkowski pooling and percentile pooling adjust the magnitude of distortions or discard the less/non distorted pixels. These pooling strategies and the mean pooling do not take into account the spread of the distortions. It is shown in [19] by case examples and experimental results that a common wrong prediction by mean pooling is where it calculates the same quality scores for two distorted images of different type. In such cases, deviation pooling is likely to provide good judgments over their quality through spread of the distortions. This is the reason why mean pooling have good inter-class (one distortion type) quality prediction but its performance might be degraded for intra-class (whole dataset) quality prediction. While this statement can be verified from the experimental results provided in [19], an example is also provided in subsection III-D to support this statement. Human visual system penalizes severer distortions much more than the distortion-free regions, and these pixels may constitute different fractions of distorted images. Mean pooling, however, depending on this fraction, is likely to nullify the impact of the severer distortions by inclusion of distortion-free regions into the average computation. Fig. 2 shows overlapped histograms of two similarity maps corresponding to two distorted images. While mean pooling indicate that image #1 is of better quality than image #2 ($\mu_1 > \mu_2$), deviation pooling provides an opposite assessment ($\sigma_1 > \sigma_2$).

Given that $\mu_1 > \mu_2$, and that image #1 has more severe distortions compared to image #2 with their values farther from $\mu_1$ than $\mu_2$, there are larger deviations in similarity map of image #1 than that of image #2. Therefore, deviation pooling is an alternative to the mean pooling that can also measure different levels of distortions. In the following, we propose the deviation pooling (DP) strategy and provide a general formulation of this pooling. DP for IQAs is rarely used in as the Minkowski distance of order $\rho$ between vector $x$ and its MCT (Measure of Central Tendency):

$$D^{(\rho)}_p = \left(\frac{1}{N} \sum_{i=1}^{N} |x_i - \text{MCT}|^{\rho}\right)^{1/\rho}. \tag{10}$$

where $\rho \geq 1$ indicates the type of deviation. The only MCT that is used in this paper is mean. Though other MCTs such as median and mode can be used, we found that these MCTs do not provide satisfactory quality predictions.

Several researches shown that more emphasis on the severer distortions can lead to more accurate predictions [29], [31].

The Minkowski pooling [29] and the percentile pooling [31] are two examples. As mentioned before, these pooling strategies follow a property of HVS that penalize severer distortions much more than the less distorted ones even though they constitute a small portion of total distortions. Hence, they try to moderate the weakness of the mean pooling through adjusting magnitudes of distortions [29] or discarding the less/non distorted regions [31]. The deviation pooling can be generalized to consider the aforementioned property of HVS:

$$D^{(\rho,q)}_p = \left(\frac{1}{N} \sum_{i=1}^{N} |x_i^q - \text{MCT}|^{\rho}\right)^{1/\rho}. \tag{11}$$

where, $q$ adjusts the emphasis of the values in vector $x$, and MCT is calculated through $x_i^q$ values. Furthermore, we propose to use power pooling in conjunction with the deviation pooling to control numerical behavior of the final quality scores:

$$D^{(\rho,q,o)}_p = \left[\left(\frac{1}{N} \sum_{i=1}^{N} |x_i^q - \text{MCT}|^{\rho}\right)^{1/\rho}\right]^o. \tag{12}$$

where, $o$ is the power pooling applied on the final value of the deviation. The power pooling can be used to make an IQA model more linear versus the subjective scores or might be used for better visualization of the scores. Linearity might not be a significant advantage of an IQA, but it is pointed to be of interest in [19]. Also, according to [36], linearity against subjective data is one of the measures for validation of IQAs that should be examined [1]. The power pooling can also have small impact on the values of Pearson linear Correlation Coefficient (PCC) and Root Mean Square Error (RMSE). Note that the above deviation pooling is equal to the Minkowski pooling [29] when $MCT = 0$, $\rho = 1$ and $o = 1$. It is equal to the mean absolute deviation (MAD) to the power of $o$ for $\rho = 1$, and equal to the standard deviation (SD) to the power of $o$ for $\rho = 2$. The three parameters should be set according to the IQA model. More analysis on these three parameters can be found in experimental results section. For the proposed index MDSI, we set $\rho = 1$, $q = \frac{1}{4}$ and $o = \frac{1}{4}$. Therefore, the proposed IQA model can be written as:

$$\text{MDSI} = \left[\frac{1}{N} \sum_{i=1}^{N} \hat{GCS}^{1/4} - \left(\frac{1}{N} \sum_{i=1}^{N} \hat{GCS}^{1/4}\right)\right]^{1/4}. \tag{13}$$

Note that possible interval for $\hat{GCS}$ is $[0 - \delta_1, 1 + \delta_2]$, where $\delta_1 < 1$ and $\delta_2 < 1$. It is worth to mention that values of GCS

1 Though linearity is measured after a nonlinear analysis.
mostly remain in $[0, 1]$. Also, $\hat{\text{GCS}} < 0$ are highly distorted pixels, while $\hat{\text{GCS}} > (1 - \epsilon)$ refer to less/non-distorted pixels, where \( \epsilon < 1 \) is a very small number. The global variations of $\hat{\text{GCS}}^{1/4}$ is computed by mean absolute deviation, which is followed by power pooling. Note that since absolute of deviations is computed, the quality scores are positive. Larger values of the quality predictions provided by the proposed index indicate to the more severe distorted images, while an image with perfect quality is assessed by a quality score of zero since there is no variation in its similarity map. The important point on the use of the Minkowski pooling on final similarity maps is that terms like “more emphasize” and “less emphasize”, regardless of the \( q \) values have been used, depends also on the pooling strategy and underlying similarity map. For example, placing more emphasize on highly distorted regions by Minkowski pooling will decrease the quality score computed by the mean pooling, but the quality score provided by the deviation pooling might become larger or smaller depending on the spread of the distortions which is directly related to the underlying similarity map.

\[ E. \text{ Analysis and Examples of GCS Maps} \]

In this section, final similarity maps after applying the Minkowski pooling, e.g. $\text{GCS}^{1/4}$ and $\hat{\text{GCS}}^{1/4}$, are compared along with sufficient explanations. The difference between these two similarity map is their use of gradient similarity. GCS uses conventional GS, while GCS uses the proposed gradient similarity $\hat{\text{GS}}$. The best way to analyze the effect of the proposed gradient similarity is through step by step explanation and visualization of different examples. In subsection [II-A] several disadvantages of the traditional GS was mentioned. Here, each of them are explained and examples are provided.

\[ a) \text{ Case 1 (Removed edge):} \text{ Missing edges in distorted image with respect to its original image means that structural information are removed, hence this disappearance bring attention of the HVS. These regions has to be strongly highlighted in the similarity map.} \]

\[ b) \text{ Case 2 (A weak added/removed edge):} \text{ An extra edge in } D \text{ or a removed edge from } R \text{ bring less attention of HVS if its color is close to the relative pixels of that edge in } R (D), \text{ or simply it is a weak edge.} \]

Fig. 3 shows how the proposed gradient similarity map $\hat{\text{GS}}$ performs for case 1 and case 2 as a part of the GCS compared to the GS for GCS. We can see that $\hat{\text{GS}}_{RD}$ (GS) highlighted differences with details. The edges corresponding to the location of ropes in original image are mainly replaced with pixels of another color (dark replaced with green), but many other edges with smaller strengths in $R$ are replaced with pixels having the same color (green). This latter holds for added edges to the distorted image. In fused image ($L_D$), some of these weaker edges are smooth. This can be seen by comparing $\hat{\text{GS}}_{RD}$ and $\hat{\text{GS}}_{DF}$. Both $\hat{\text{GS}}_{RD}$ and $\hat{\text{GS}}_{DF}$ indicate to high differences at the location of the ropes. $\hat{\text{GS}}_{RD} + \hat{\text{GS}}_{DF}$ will also put high emphasize on this location, but less emphasize on the weaker edges. The results is then subtracted by $\hat{\text{GS}}_{RF}$ which in turn again less emphasize is placed on the weak edges (relevant to the darker pixels in $\hat{\text{GS}}_{RF}$). Note that GCS and $\hat{\text{GCS}}$ have different numerical behavior, so it is fair to compare them by looking at the $\text{GCS}^{1/4}$ and $\hat{\text{GCS}}^{1/4}$. Compared to the $\text{GCS}^{1/4}$, $\hat{\text{GCS}}^{1/4}$ indicate to larger differences at the location of ropes, but smaller differences elsewhere.

\[ c) \text{ Case 3 (Preserved edge but with different color):} \text{ Although a color similarity map should measure color differences at the location of the inverted edges, edges constitute a small fraction of total pixels in images, and it is common to give smaller weights to a color similarity map than structural sim-} \]
ilarities such as gradient similarity. While traditional gradient similarity does not work well in this situation, the proposed gradient similarity can partially solve this problem. Fig. [4] provides an example in which most of the edges are inverted in the distorted image. We can see that $\hat{GCS}_{1/4}$ highlighted much more differences than $GCS_{1/4}$ at these locations, thanks to the added term $(GS_{DF} - GS_{RF})$ to the traditional gradient similarity. In fact, $GS_{DF}$ is likely to be different than $GS_{RF}$ in this case because these edges in $F$ are likely to become closer to their surrounding pixels in either $R$ or $D$ images.

### III. Experimental Results and Discussion

In the experiments, eight datasets were used. The LIVE dataset [37] contains 29 reference images and 779 distorted images of five categories. The TID2008 [38] dataset contains 25 reference images and 1700 distorted images. For each reference image, 17 types of distortions of 4 degrees are available. CSIQ [12] is another dataset that consists of 30 reference images; each is distorted using six different types of distortions at four to five levels of distortion. The large TID2013 [39] dataset contains 25 reference images and 3000 distorted images. For each reference image, 24 types of distortions of 5 degrees are available. VCL@FER database [40] consists of 23 reference images and 552 distorted images, with four degradation types and six degrees of degradation. In addition to these five datasets, contrast distorted images of the CCID2014 dataset [41] are used in the experiments. This dataset contains 655 contrast distorted images of five types. Gamma transfer, convex and concave arcs, cubic and logistic functions, mean shifting, and a compound function are used to generate these five types of distortions. We also used the ESPL synthetic image database [42] which contains 25 synthetic images of video games and animated movies. It contains 500 distorted images of 5 categories. Fig. [5] shows an example of a reference and a distorted synthetic image. Finally, the digitally retouched image quality (DRIQ) dataset [43] was used in the experiments. It contains 26 reference images and 3 enhanced images for each reference image.

For objective evaluation, four popular evaluation metrics were used in the experiments: the Spearman Rank-order Correlation coefficient (SRC), the Pearson linear Correlation Coefficient (PCC) after a nonlinear regression analysis (equation [14]), the Kendall Rank Correlation coefficient (KRC) and the Root Mean Square Error (RMSE). The SRC, PCC, and RMSE metrics measure prediction monotonicity, prediction linearity, and prediction accuracy, respectively. The KRC was used to evaluate the degree of similarity between quality scores and MOS. In addition, Pearson linear Correlation Coefficient without nonlinear analysis is used and denoted by LPCC.

Twelve state-of-the-art IQA models were chosen for comparison [7], [9], [12], [15], [16], [18]–[23], [44] including the most recent indices in literature [19]–[23], [44]. It should be noted that the five indices SFF [20], GMSD [19], VSI [21], [22], and SCQI [23] have shown superior performance over state-of-the-art indices.

#### A. Performance comparison

In Table I, the overall performance of thirteen IQA models on eight benchmark datasets, e.g. TID2008, CSIQ, LIVE, TID2013, VCL@FER, CCID2014, ESPL, and DRIQ, is listed. For each dataset and evaluation metric, the top three IQA models are highlighted. On eight datasets, MDSI is 31 times among the top indices, followed by ADD-GSIM (16 times), SCQI (12 times), SFF/FSIM/VIF (6 times), VSI (5 times), GMSD/SR_SIM/MAD (4 times), DSCSI (2 times), and MSSIM/IWSSIM (0 times). To provide a conclusion on the overall

\[ \text{Note the conflict between ‘MAD’ [12] as an IQA model, and ‘MAD’ as a pooling strategy.} \]
performance of these indices, direct and weighted\footnote{The dataset size-weighted average is commonly used in the literature\cite{19-21}.} overall performances on the eight datasets (8150 images) are also listed in Table I. It can be seen that MSDI has the best overall performance on the eight datasets, while metrics VSI and SCQI are the second, and third best, respectively.

B. Visualization and statistical evaluation

For the purpose of visualizing quality scores of the proposed index, the scatter plots of the proposed IQA model MSDI with and without using power pooling are shown in Fig. 6. The logistic function suggested in\cite{45} was used to fit a curve on each plot:

\[
f(x) = \beta_1 \left( \frac{1}{1 + e^{-(x - \beta_3)}} \right) + \beta_4 x + \beta_5
\]

where $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\beta_5$ are fitting parameters computed by minimizing the mean square error between quality predictions $x$ and subjective scores MOS. It should be noted that reported PCC and RMSE values in this paper are computed after mapping quality scores to MOS based on above function.

The reported results in Table I show the difference between different IQA models. As suggested in\cite{45, 46}, we use F-test to decide whether a metric is statistically superior to another index. The F-test is based on the residuals between the quality scores given by an IQA model after applying nonlinear mapping of equation\cite{(14)}, and the mean subjective scores MOS. The ratio of variance between residual errors of an IQA model to another model at 95% significance level is used by F-test. The result of the test is equal to 1 if we can reject the null hypothesis and 0 otherwise. The results of F-test on eight datasets are listed in Table II. In this Table, +1/-1 indicate that corresponding index is statistically superior/inferior to the other index being compared to. If the difference between two indices is not significant, the result is shown by 0. We note that

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline
IQA & SRC & TID & KRC & RMSE & SRC & TID & KRC & RMSE & SRC & TID & KRC & RMSE \\
\hline
PCC & 0.8542 & 0.8451 & 0.8658 & 0.7173 & 0.9133 & 0.8991 & 0.8559 & 0.1149 & 0.9513 & 0.9489 & 0.8044 & 0.8618 \\
\hline
RMSE & 0.8401 & 0.8290 & 0.6645 & 0.7537 & 0.9467 & 0.9500 & 0.7970 & 0.0980 & 0.9669 & 0.9604 & 0.8282 & 7.6137 \\
\hline
CSIQ & 0.8789 & 0.7698 & 0.7087 & 0.8073 & 0.8073 & 0.8119 & 0.8299 & 0.8073 & 0.7698 & 0.8073 & 0.8299 & 7.6137 \\
\hline
LIVE & 0.8239 & 0.7720 & 0.8267 & 0.8063 & 0.8239 & 0.7720 & 0.8267 & 0.8063 & 0.8239 & 0.7720 & 0.8267 & 7.6137 \\
\hline
VCL & 0.8044 & 0.6380 & 0.5877 & 0.6771 & 0.8044 & 0.6380 & 0.5877 & 0.6771 & 0.8044 & 0.6380 & 0.5877 & 7.6137 \\
\hline
FER & 0.8559 & 0.8661 & 0.6976 & 0.6880 & 0.8559 & 0.8661 & 0.6976 & 0.6880 & 0.8559 & 0.8661 & 0.6976 & 7.6137 \\
\hline
CCID & 0.7770 & 0.8349 & 0.7451 & 0.7811 & 0.7770 & 0.8349 & 0.7451 & 0.7811 & 0.7770 & 0.8349 & 0.7451 & 7.6137 \\
\hline
ESPI & 0.7247 & 0.7488 & 0.8264 & 0.8270 & 0.7247 & 0.7488 & 0.8264 & 0.8270 & 0.7247 & 0.7488 & 0.8264 & 7.6137 \\
\hline
DRIQ & 0.7058 & 0.8496 & 0.6967 & 0.7155 & 0.7058 & 0.8496 & 0.6967 & 0.7155 & 0.7058 & 0.8496 & 0.6967 & 7.6137 \\
\hline
Direct Avg & 0.8284 & 0.8234 & 0.8411 & 0.8408 & 0.8284 & 0.8234 & 0.8411 & 0.8408 & 0.8284 & 0.8234 & 0.8411 & 7.6137 \\
\hline
Weighted Avg & 0.8575 & 0.8570 & 0.8573 & 0.8575 & 0.8575 & 0.8570 & 0.8573 & 0.8575 & 0.8575 & 0.8570 & 0.8573 & 7.6137 \\
\hline
\end{tabular}
\caption{Performance comparison of the proposed IQA model, MSDI, and twelve popular/competing indices on eight benchmark datasets. Note that top three IQA models are highlighted.}
\end{table}
The results of Table II show that MDSI is significantly better than the other indices on TID2008, TID2013, ESPL, and DRIQ datasets. Its sum value in the last column is +9 for these four datasets. SCQI is statistically superior to the other indices on the TID2013 dataset except for MDSI. On the LIVE dataset, indices MAD, MDSI, and ADD-GSIM are significantly better than the other indices. On the CSIQ dataset, only SFM performs significantly better than MDSI. On the CCID2014 dataset, ADD-GSIM is significantly better than the other indices, while the statistically indistinguishable indices VIF and MDSI show promising results. Considering all eight datasets used in this experiment, with a minimum sum value of +6, the proposed index MDSI performs very well in comparison with the other indices. We can simply add the eight cumulative sum values of each metric for the eight datasets to have an overall comparison based on the statistical significance test. This score indicates how many times a metric is statistically superior to the other metrics. The results show that MDSI is the best performing index by a score of +62 (out of maximum +72), followed by ADD-GSIM (+25), VSI (+1), SCQI (-2), FSIMc (-4), GMSD (-5), SFF (-6), SR_SIM (-19), MAD (-24), and VIF (-26). The results based on the statistical significance test verify that unlike other IQA models, the proposed metric MDSI is among the best performing indices on different datasets.

C. Performance comparison on individual distortions

A good IQA model should perform not only accurate quality predictions for a whole dataset; it should provide good judgments over individual distortion types. We list in Table III the average SRC, and PCC values of thirteen IQA models for 61 sets of distortions available in the six datasets of TID2008, CSIQ, LIVE, TID2013, VCL@FER, and ESPL. The minimum value for each evaluation metric and standard deviation of these 61 values are also listed. These two evaluations indicate the reliability of an IQA model. An IQA model should provide good prediction accuracy for all of the distortion types. If a metric fails at assessing one or more types of distortions, that index cannot be reliable.

The proposed index MDSI, has the best SRC, and PCC average on distortion types. MDSI, SCQI and FSIMc, in the worst case perform better than other IQA models, as can be seen in the min column for each evaluation metric. This shows the reliability of the proposed index. The negative min values and close to zero min values in Table III indicate the unreliability of related models when dealing with some
distortion types. The standard deviation of 61 values for each evaluation metric is another reliability factor. According to Table III, MDSI, SCQI and FSIM have the lowest variation. Therefore, we can conclude that indices MDSI, SCQI and FSIM are more reliable than the other indices.

### Table III

| Model | SRC (Distortions) | PCC (Distortions) |
|-------|-------------------|-------------------|
| MESSIM | 0.8343 -0.4099 0.1989 | 0.8560 -0.4448 0.1944 |
| VIF | 0.8537 -0.3099 0.1811 | 0.8760 -0.3443 0.1812 |
| MAD | 0.8111 -0.0575 0.2315 | 0.8296 0.0417 0.2108 |
| IWSSIM | 0.8329 -0.4196 0.2019 | 0.8586 -0.4503 0.1962 |
| SR_SIM | 0.8609 -0.2053 0.1806 | 0.8785 -0.3162 0.1839 |
| FSIM | 0.8775 0.4679 0.1041 | 0.8967 0.5488 0.0880 |
| GMSD | 0.8542 -0.2948 0.1954 | 0.8785 -0.3625 0.1851 |
| SFF | 0.8538 0.1786 0.1472 | 0.8721 0.0786 0.1441 |
| VSI | 0.8779 0.1713 0.1360 | 0.8969 0.4875 0.1044 |
| DSCSI | 0.8722 0.3534 0.1242 | 0.8908 0.5166 0.1093 |
| ADD-GSIM | 0.8650 -0.2053 0.1686 | 0.8799 -0.2190 0.1691 |
| SCQI | 0.8826 0.4479 0.1057 | 0.9010 0.6493 0.0841 |
| MDSI | 0.8903 0.4787 0.1030 | 0.9005 0.6899 0.0805 |

### Table IV

| Pooling | Weighted avg. SRC (8 datasets) | Avg. SRC (61 Distortions) |
|---------|-------------------------------|---------------------------|
|        | Mean | MAD | SD  | Mean | MAD | SD  |
| q = 1/4 | 0.8864 | 0.9066 | 0.8776 | 0.8919 | 0.8903 | 0.8828 |
| q = 1/2 | 0.8833 | 0.9067 | 0.8820 | 0.8912 | 0.8898 | 0.8826 |
| q = 1   | 0.8730 | 0.9041 | 0.8899 | 0.8899 | 0.8890 | 0.8820 |
| q = 2   | 0.8519 | 0.8928 | 0.8972 | 0.8888 | 0.8866 | 0.8820 |
| q = 4   | 0.8301 | 0.8766 | 0.8922 | 0.8869 | 0.8780 | 0.8753 |

### D. Parameters of deviation pooling ($\rho$, $q$, $o$)

Considering the formulation of deviation pooling in equation (12), we used the mean absolute deviation (MAD), e.g. $\rho = 1$, for the proposed metric. Standard deviation (SD), e.g. $\rho = 2$, is another option that can be used for deviation pooling. In addition, the Minkowski power ($q$) of the deviation pooling can have significant impact on the proposed index. In Table IV, the SRC performance of the proposed index is analyzed for different values of $q$ and $\rho = \{1, 2\}$. Mean pooling is also used in this experiment. The results show that MAD pooling with $q = 1$ is a better choice for the proposed index. Also, the performance of the mean pooling on 61 distortion set confirms our statement that mean pooling has a good performance for inter-class quality prediction.

The impact of the proposed power pooling of the deviation pooling on the proposed metric was shown in Fig. 6. Power pooling can be also used to increase linearity of other indices as well. For example, LPCC and PCC values of VSI [21] for TID2013 dataset, by setting $o = 18$, can be increased from 0.8373 to 0.8928, and 0.9000 to 0.9011, respectively.

### E. Summation vs. Multiplication

Two options for combination of the two similarity maps GS/GS and CS are summation and multiplication as explained in subsection [L-C]. Deciding whether one approach is superior to another for an index depends on many factors. These factors might be the pooling strategy being used, overall performance, performance on individual distortions, reliability, efficiency, simplicity, etc. In an experiment, the performance of the MDSI using the multiplication approach was examined. Based on the many set of parameters were tested, we found that $\gamma = 0.2$ and $\beta = 0.1$ are good parameters to combine GS and CS via the multiplication scheme. The observation was that summation is a better choice for TID2008, TID2013, VCL@FER, and DRIQ datasets, while multiplication is a better choice for ESPL dataset, and that both approaches show almost the same performance on other datasets. Overall, the summation approach provides better performance on individual distortions. This experiment also shown that MDSI is more reliable through summation than multiplication based on the reliability measures introduced in this paper. Based on this experiment, the simplicity of the summation combination approach and its efficiency over multiplication, the former was used along with MDSI. Table V justifies our choice.

### F. Parameters of model

The proposed IQA model MDSI has four parameters to be set. The four parameters of MDSI are $C_1$, $C_2$, $C_3$ and $\alpha$. To further simplify the MDSI, we set $C_3 = 4C_1 = 10C_2$. Therefore, MDSI has only two parameters to set, e.g. $C_3$ and $\alpha$. For an example, we refer to the SSIM index [2] that also uses such a simplification. Note that gradient similarities and chromaticity similarity have different dynamic ranges, therefore, these parameters should be set such that the relation between these maps also be taken into account.

In Fig. 7, the impact of these two parameters on the performance of the MDSI is shown. Even though the parameters $C_1$, $C_2$ and $C_3$ are set approximately, it can be seen that MDSI is very robust under different setup of parameters. MDSI has greater weighted average SRC than 0.90 for any $\alpha \in [0.5, 0.7]$ and $C_3 \in [300, 600]$. Note that many other possible setup of parameters are not included in this plot. In the experiments, we set $\alpha = 0.6$, $C_1 = 140$, $C_2 = 55$, and $C_3 = 550$.

### G. Effect of chromaticity similarity maps CS and $\hat{C}$S

In this section, the impact of using $S_3$ [21] and proposed $\hat{C}$S on the performance of the proposed index is studied through the following experiment. Contrast distorted images of the CCID2014 dataset [41] were chosen. The reason of choosing this dataset is to evaluate the ability of measuring color changes by CS and $\hat{C}$S. We analyzed the SRC performance of the CS and $\hat{C}$S as a part of the proposed index for wide

---

**TABLE V**

DIFERENT CRITERIA USED TO CHOOSE THE COMBINATION SCHEME.

| Property | Summation | Multiplication |
|----------|-----------|---------------|
| Statistically superior over more considered datasets | ✔️ | ✔️ |
| Better dataset-weighted average | ✔️ | ✔️ |
| Better performance on individual distortions | ✔️ | ✔️ |
| Reliability | ✔️ | ✔️ |
| Simplicity | ✔️ | ✔️ |
| Efficiency | ✔️ | ✔️ |

---

**TABLE VI**

PERFORMANCE OF THE PROPOSED INDEX MDSI WITH DIFFERENT POOLING STRATEGIES AND VALUES OF PARAMETER $q$.

| Pooling | Mean | MAD | SD |
|---------|------|-----|----|
| Weighted SRC (8 datasets) | 0.8919 | 0.8903 | 0.8828 |
| Avg. SRC (61 Distortions) | 0.8919 | 0.8903 | 0.8828 |
| q = 1/4 | 0.8864 | 0.9066 | 0.8776 |
| q = 1/2 | 0.8833 | 0.9067 | 0.8820 |
| q = 1   | 0.8730 | 0.9041 | 0.8899 |
| q = 2   | 0.8519 | 0.8928 | 0.8972 |
| q = 4   | 0.8301 | 0.8766 | 0.8922 |
range of $C_3$ values. Three pooling strategies were used in this experiment, e.g., mean pooling, mean absolute deviation (MAD) pooling and the standard deviation (SD) pooling. Fig. 8 shows the SRC performance of the proposed index for different scenarios. From the plot in Fig. 8 the following conclusions can be drawn. MAD pooling and both CS and $\hat{\text{CS}}$ are good choices for MDSI. For almost every pooling strategy and parameter of $C_3$, the proposed $\hat{\text{CS}}$ performs better than CS. This advantage is at the same time that the proposed $\hat{\text{CS}}$ is more efficient than the existing CS.

Fig. 7. The weighted SRC performance of MDSI for different values of $C_3$ and $\alpha$ on eight datasets (TID2008, CSIQ, LIVE, TID2013, VCL@FER, CCID2014, ESPL, and DRIQ).

Fig. 8. The SRC performance of the proposed index MDSI with two chromaticity similarity maps CS and $\hat{\text{CS}}$ (proposed) for different values of $C_3$ and three pooling strategies on CCID2014 dataset [41].

H. Implementation and efficiency

Another very important factor of a good IQA model is its efficiency. The proposed index has a very low complexity. It first applies average filtering of size $M \times M$ on each channel of the $R$ and $D$ images, downsample them by a factor of $M$ and convert the results to a luminance and two chromaticity channels [47]. The value of $M$ is set to $\lfloor \min(h, w)/256 \rfloor$ [38], where $h$ and $w$ are image height and width, and $\lfloor . \rfloor$ is the round operator. Then, the proposed index calculates the gradient magnitudes of luminance channel, the chromaticity similarity map, and apply deviation pooling. All these steps are computationally efficient. Table VI lists the run times of fifteen IQA models when applied on images of size $384 \times 512$ and $1080 \times 1920$. The experiments were performed on a Corei7 3.40GHz CPU with 16 GB of RAM. The IQA models were implemented in MATLAB 2013b running on Windows 7. It can be seen that MDSI is among top five fastest indices. The proposed index is less than 2 times slower than the competing GMSD index. The reason for this is that GMSD only uses the luminance feature. Compared to the other competing indices, SCQI, VSI, ADD-GSIM, SFF, and FSIM, the proposed index MDSI is about 3 to 6 times, 3 to 9 times, 4 to 5 times, and 4 to 11 times faster, respectively. Another observation from the Table VI is that the ranking of indices might not be the same when they are tested on images of different size. For example, SSIM performs slower than the proposed index on smaller images, but faster on larger images.

IV. Conclusion

We propose an effective, efficient, and reliable full reference IQA model based on the new gradient and chromaticity similarities. The gradient similarity was used to measure local structural distortions. In a complementary way, a chromaticity similarity was proposed to measure color distortions. The proposed metric, called MDSI, use a novel deviation pooling to compute the quality score from the two similarity maps. Extensive experimental results on natural and synthetic benchmark datasets prove that the proposed index is effective and reliable, has low complexity, and is fast enough to be used in real-time FR-IQA applications.

ACKNOWLEDGMENTS

The authors thank the NSERC of Canada for their financial support under Grants RGPDD 451272-13 and RGPIN 138344-14.

REFERENCES

[1] Z. Wang, “Applications of objective image quality assessment methods [applications corner],” IEEE Signal Processing Magazine, vol. 28, no. 6, pp. 137–142, Nov 2011.
[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004.
[3] H. Yeganeh and Z. Wang, “Objective quality assessment of tone-mapped images,” IEEE Transactions on Image Processing, vol. 22, no. 2, pp. 657–667, Feb 2013.
