TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

C2G	Color to gray
CS	Chromaticity similarity
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DP	Deviation pooling
DWT	Discrete wavelet transform
FR-IQA	Full-reference image quality assessment
GS	Gradient similarity
HDR	High dynamic range
HVS	Human visual system
IQA	Image quality assessment
KRC/KRCC	Kendall rank correlation coefficient
LDR	Low dynamic range
LWMPA	Locally weighted mean phase angle
MAD	Mean absolute deviation
MATLAB	Matrix laboratory
MDM	Minkowski distance based metric
MUG	Median of unique gradients
MSE	Mean square error

NUG	Number of unique gradients
NR-IQA	No-reference image quality assessment
PCC/PLCC	Pearson linear correlation coefficient
RMSE	Root mean square error
RR	Reduced-reference
RR-IQA	Reduced-reference image quality assessment
SD	Standard deviation
SRC/SRCC	Spearman rank-order correlation coefficient
SSIM	Structural similarity
SVR	Support vector regression
TMO	Tone-mapping operator
TMQI	Tone mapping quality index

INTRODUCTION

Recent years have seen a tremendous growth in acquisition, transmission, and storage of digital media data. It is predicted that every second, nearly a million minutes of video content will cross the network by 2019 (Cisco Systems, 2015). Consequently, there has been an increasing demand for accurate and efficient image and video quality assessment metrics in order to monitor, maintain and store visual media data.

Human visual system is the ultimate viewer of the visual media data. A typical visual media is subject to one or more processing stages such as acquisition/rendering, transmission, compression, among other processes. Each of these stages may change information and affect the quality of the visual data. It is necessary to maintain a satisfactory quality of experience for viewers of visual data.

The most accurate way to judge the quality of images and videos is to conduct subjective experiments. However, given the huge amount of visual data, such experiments are very time-consuming. Therefore, automatic objective image/video quality assessment metrics that can mimic subjective evaluation of visual data are of great interest. These computational models take into account changes in visual data information only if these changes cause annoyance for viewers. The non-visible information changes in visual data are ignored by these metrics.

Objective image quality assessment (IQA) models can be categorized into full-reference (FR), reduced-reference (RR), and blind/no-reference (NR) depending on their access to the reference image with pristine quality. Figure 0.1 provides an illustration of this categorization. Both reference image and possibly distorted image are available for a FR-IQA metric. RR-IQA models have full access to the distorted image and perform assessments with respect to some certain statistical properties of the reference image. To perform quality assessment of a possibly distorted image, NR-IQA models have no access to the reference image.

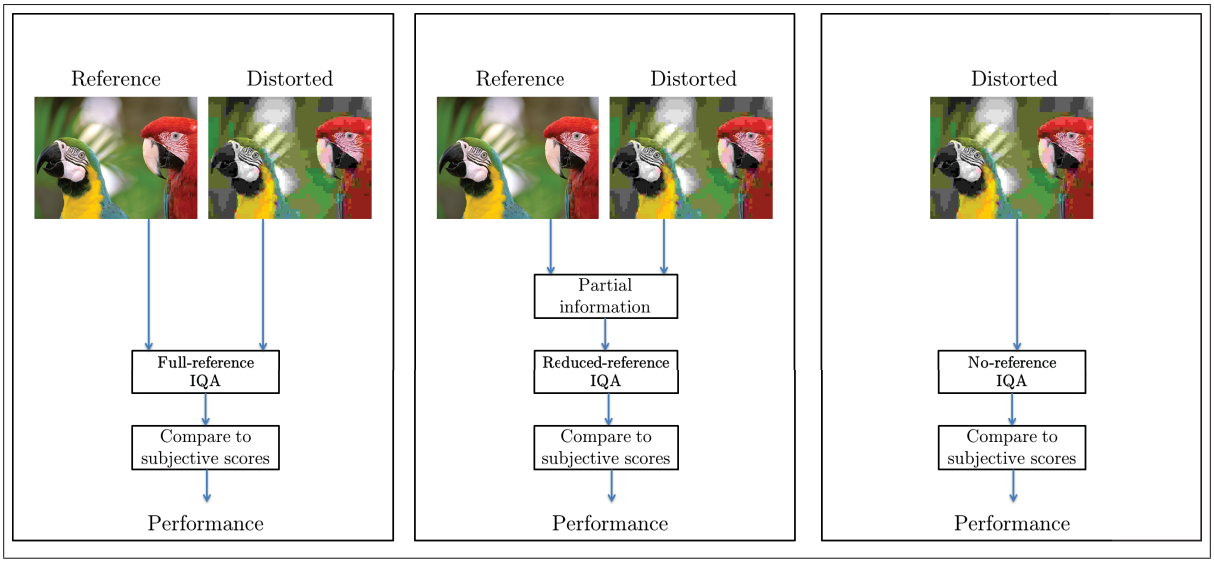

Figure 0.1 Category of image quality assessment models. From left to right, full-reference IQA model, reduced-reference IQA model, and no-reference IQA model.

According to (Wang & Bovik, 2009), objective image quality assessment is a multidisciplinary topic with a wide range of research directions at the intersection of image and signal processing, computer vision, visual psychophysics, neural physiology, information theory, machine learning, design of image acquisition, communication, and display systems.

IQA models can be used in parallel with an image processing system to provide feedback for the system or can be directly embedded into the system. Performance of a recognition system for an application can be greatly affected by image distortions. Objective IQA can help to estimate performance expectation of a recognition system or can provide information to preprocess the input image first and run the recognition system on the processed image.

While IQA is useful in many image processing scenarios, for real-time image processing systems it can add a significant computational complexity to the system. Even for applications where speed is not a major factor, more efficient algorithms that do not sacrifice performance are obviously preferred. This opens up the main question of this thesis, **what are efficient and effective features for designing perceptually consistent image quality assessment metrics?**

According to (Chandler, 2013), run-time performance is one of the major challenges in image quality assessment. For better understanding of what is supposed to be assessed by IQA models, we detail in the next section some of the main challenges in IQA, with a focus on distortion types and image content.

0.1 Problem statement

Despite recent advancements in the field of image quality assessment, because of its large scope, it is a challenging problem which is yet to be solved. In the following, major challenges of full and no reference IQA models are explained.

a. Distortion types: Different distortion types change different properties of images such as structure and color. The human visual system has separate processing mechanism for achromatic and chromatic signals (Lin & Kuo, 2011). Many IQA models do not consider chrominance information, while others give more weight to luminance which means current IQA models may not deliver high accuracy predictions for chromatic distortions. For an example to visualize the variety of distortions in images, we mention TID2013 dataset (Ponomarenko *et al.*, 2013) which contains 24 types of distortion with different levels of distortion. 24 distortion types seem to be a relatively large number, yet this dataset does not contain multiply-distorted images nor contains some of the chromatic distortions such as distortions caused by gamut mapping. Assessment of arbitrary distortion types become much more challenging in no-reference IQA. Figure 0.2 shows a reference image and its distorted versions with different artifacts.

b. Image content: Pristine-quality images have different edge and color information. An image that is rich in edge information is called to have high spatial information, while an image that is rich in color is referred to have high colorfulness. The human visual system does not equally evaluate the quality of two images with different spatial information and colorfulness.

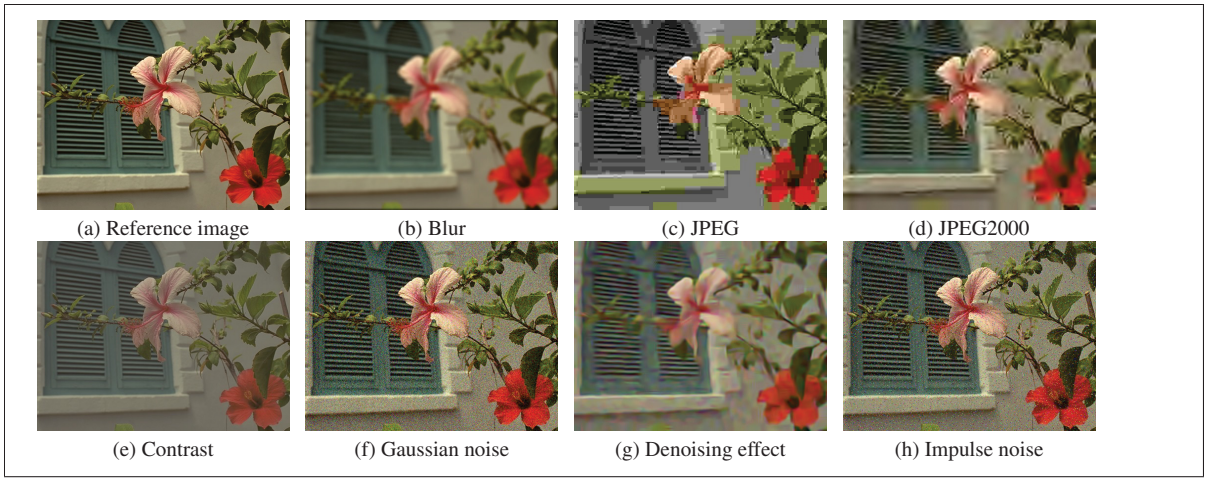

Figure 0.2 An example of a reference image and its seven distorted versions.

In practice given large variability in size and nature of edge and color features in images, it is difficult to consider different image contents by a single IQA model. NR-IQA becomes much more difficult when unknown image contents exist in test images. Figure 0.3 shows sample images with different edge and color information.

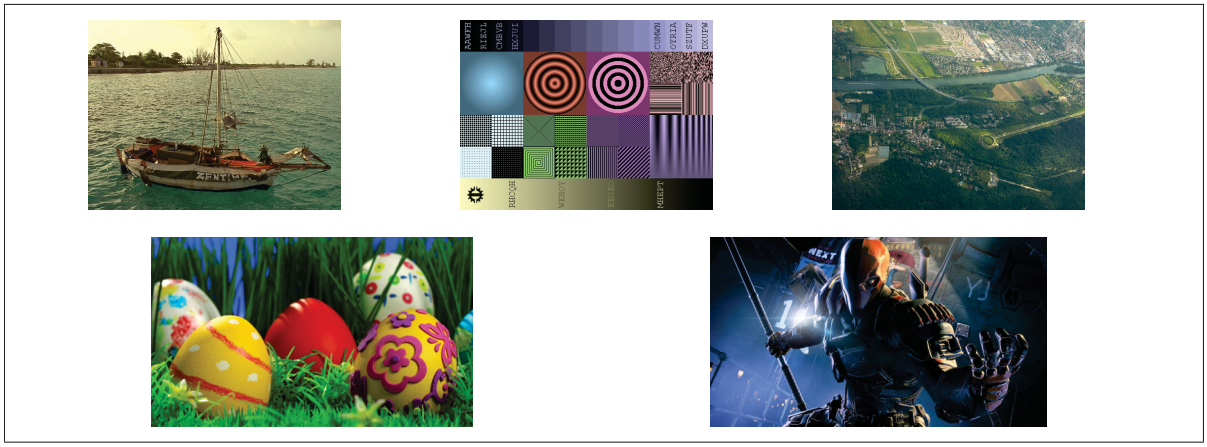

Figure 0.3 Example of images with different edge and color information. Images are taken from (Ponomarenko *et al.*, 2013; Zaric *et al.*, 2012; Du *et al.*, 2015; Kundu & Evans, 2015).

c. Dynamic range: When reference image and its processed image (for example tone-mapped or color to gray converted) do not have the same dynamic range or do not belong to the same

color space, it introduces new challenges for quality assessment of the processed image. Traditional IQA models cannot be used in such scenarios.

d. Run-time performance: In order to account for different distortion types that were mentioned above, IQA models may use different domains, image transformations, and several color spaces. As a result, IQA models become more complex and it is difficult to properly tune several introduced parameters by this consideration. This problem becomes crucial for NR-IQA models that usually extract several statistical features from images to account for different distortion types and contents. Such NR-IQA models are not suitable to be used in real-time image processing systems. For example, a recently proposed NR-IQA metric called FRIQUEE (Ghadiyaram & Bovik, 2017) performs assessment by extracting a large number of features (564 features).

Considering the aforementioned challenges, the choice of efficient and effective similarity maps or statistical features for image quality assessment remains an open question. This is demonstrated by a large number of features and IQA models proposed in literature without being effective and efficient simultaneously.

0.2 Contributions

Past research introduced several features and image quality assessment models. As mentioned in the previous section, it is difficult to find effective and efficient features and models. The search for salient, more relevant yet efficient features is still an active field of research. **Therefore, the purpose of this thesis is to introduce new similarity maps and features and use them to design efficient perceptually consistent image quality assessment models.** Our research focuses on two categories of image quality assessment: full-reference models and no-reference models.

First, we focus on a full-reference IQA model that is able to assess a wide range of image distortions and image types/contents with a consistent performance. Contribution will be made on the proposition of a new gradient similarity map to better measure the structural distortions and a new color similarity map that is more efficient and effective than the existing one. The proposed gradient similarity measures structural distortion by introducing an intermediate image. In addition, a general formulation for deviation pooling is proposed in this thesis which is novel and can be used along with the proposed similarity maps and existing similarity maps in the literature. It is shown in the experimental results that the proposed FR-IQA model performs significantly better than the existing metrics.

A second contribution on FR-IQA category is the proposition of a feature similarity index for tone-mapped images. The proposed metric is one of the first and few available metrics. It is based on the high agreement of human visual system to the phase information of images. Instead of previous work, the proposed metric directly compares high dynamic range and its low dynamic range version without transforming them into a specific color space.

Secondly, we focus on no-reference IQA models for JPEG compressed and contrast distorted images. The proposed NR-IQA metric for JPEG compressed images utilizes statistics of edges with a new approach. Unlike previous blockiness metrics that use prior information on block size and their position, the proposed blockiness metric is parameterless and almost invariant to misalignment and block size. Compared to a competing blockiness metric, the proposed index is hundreds of times faster.

Another contribution on NR-IQA is the proposition of highly efficient features for quality assessment and classification of contrast distorted images. This metric uses three features to train a model that provides more accurate quality predictions and have much lower computational complexity than the existing metrics. Moreover, the proposed features have high discriminative power for classification of contrast distortion types.

Thirdly, a real-time perceptually consistent color to gray methodology is proposed which is based on the correlation. Through subjective and objective evaluation, the performance of the proposed method is validated. Finally, we recommend using an objective quality assessment model for the color to gray image conversion that shows a higher correlation with the subjective evaluations.

0.3 Outline of the thesis

In this thesis, we focus on the image quality assessment topic, its challenges, and solutions that we bring to tackle these challenges.

- In **Literature review** (Chapter 1), we discuss basic and recent state of the art features and metrics that were proposed for full-reference and no-reference image quality assessment. However, these features and metrics do not solve the efficiency issues that we have posed in the Introduction, leading to the Methodology section. Each feature or metric is first described and its weakness is explained. Complementary literature review can be found in the chapters 3 to 7 concerning the journal publications.

- **General Methodology** (Chapter 2) explains the methodologies in our work and gives a brief overview of used techniques. This chapter also defines our objectives more precisely and explains our motivation behind developing each methodology.

- **Journal publications** are five chapters dedicated to our journal publications (Chapters 3 to 7). In these chapters, two proposed full-reference metrics for quality assessment of low dynamic range images and tone-mapped images are described. This is followed by two proposed no-reference metrics for quality assessment of JPEG compressed and contrast distorted images. Then, the proposed perceptually consistent color to gray methodology is explained.

- Chapter **General Discussion** (Chapter 8) provides discussion on the strengths and weaknesses of the proposed methods.
- Finally, **General Conclusion** summarizes the work accomplished in this thesis and provides our recommendations and perspectives.

CHAPTER 1

LITERATURE REVIEW

In this chapter, we review the relevant literature related to features and similarity maps used in FR and NR-IQA models. We first start with features and design strategies of FR-IQA models. We also review FR-IQA models where reference and test images are not in the same color space or dynamic range. Then, features and design strategies of NR-IQA models are reviewed. We especially focus on distortion-specific NR-IQA models for JPEG compressed and contrast distorted images. This is followed by a brief overview of the color to gray (C2G) image conversion methods.

1.1 Full-reference image quality assessment

As illustrated in section Introduction, FR-IQA models evaluate the perceptual quality of a distorted image with respect to its reference pristine-quality image. The following factors are indicators of a good FR-IQA model. FR-IQA model should provide high correlation with subjective ratings, have low complexity, provide accurate local quality map, and have mathematical properties like convexity and differentiability. Existing FR-IQA models barely satisfy aforementioned factors altogether (Bae & Kim, 2016a).

The mean square error (MSE) and peak-signal to noise ratio (PSNR) are the most widely used FR-IQA metrics because of their simplicity and efficiency. For a reference image R and its distorted version D, MSE is computed by averaging squared intensity differences of R and D:

$$\mathrm{MSE}(R,D) = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} \big(R(i,j) - D(i,j)\big)^2 \tag{1.1}$$

where w and h are image width and height, respectively. However, in many situations, MSE does not correlate with the human perception of image fidelity and quality (Wang & Bovik, 2009). MSE treats all pixels of an image equally and does not consider structural distortions.

The introduction of Structural SIMilarity index (SSIM) (Wang *et al.*, 2004) has led to the development of several SSIM-induced FR-IQA metrics. SSIM assumes that human visual system (HVS) is highly sensitive to structural distortions. SSIM is a combination of three components: luminance (mean) distortion, contrast (variance) distortion, and structure (correlation) distortion:

$$\mathrm{SSIM}(R,D) = l(R,D) \cdot c(R,D) \cdot s(R,D) \tag{1.2}$$

where:

$$l(R,D) = \frac{2\overline{R}\,\overline{D} + C_1}{\overline{R}^2 + \overline{D}^2 + C_1} \tag{1.3}$$

$$c(R,D) = \frac{2s_R s_D + C_2}{s_R^2 + s_D^2 + C_2} \tag{1.4}$$

$$s(R,D) = \frac{s_{R,D} + C_3}{s_R s_D + C_3} \tag{1.5}$$

where $\overline{R}$ and $\overline{D}$ denote the local mean of R and D, s_R^2 and s_D^2 denote the local variance of R and D, $s_{R,D}$ represents the local covariance between R and D. C_1, C_2 and C_3 are constants used for numerical stability. By substitution and setting $C_3 = C_2/2$, SSIM index can be written as:

$$\mathrm{SSIM}(R,D) = \frac{(2\overline{R}\,\overline{D} + C_1)(2s_{R,D} + C_2)}{(\overline{R}^2 + \overline{D}^2 + C_1)(s_R^2 + s_D^2 + C_2)} \tag{1.6}$$

In comparison with MSE and PSNR, SSIM has shown a better correlation with subjective quality assessment results (Wang & Bovik, 2011). Figure 1.1 shows similarity map of SSIM and absolute error map for a JPEG compressed image. It can be seen that SSIM better highlights the blocking artifacts produced by JPEG compression.

(a) Reference image (b) JPEG compressed image (c) Absolute error (d) SSIM map

Figure 1.1 Quality map of SSIM versus absolute error map for a JPEG compressed image (b) with respect to its reference image (a).

SSIM is further extended to multi-scale SSIM to account for the viewing distance of human subject:

$$\mathrm{MSSSIM}(R,D) = \prod_{r=1}^{R} [l(R_r,D_r)]^{\alpha_r} \cdot [c(R_r,D_r)]^{\beta_r} \cdot [s(R_r,D_r)]^{\gamma_r} \tag{1.7}$$

where r denotes the resolution, and α_r, β_r and γ_r are relative weights for each SSIM component at each resolution. The MSSSIM index generally outperforms the SSIM index. SSIM and MSSSIM use mean pooling strategy. IWSSIM (Wang & Li, 2011) uses weighted mean with better performance than the MSSSIM. SSIM-induced metrics in the literature follow a top-down strategy (Lin & Kuo, 2011).

In fact, the most successful IQA models in the literature follow a top-down strategy. They calculate a similarity map and use a pooling strategy that converts the values of this similarity map into a single quality 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. The pooling stage is also done based on phase congruency. FSIM_c is an extension of FSIM with an added chromatic term to measure color distortions. GS (Liu *et al.*, 2012) uses a combination of some designated gradient magnitudes and image contrast for this end, while the GMSD (Xue *et al.*, 2014b) uses only the gradient magnitude. SR_SIM (Zhang & Li, 2012) uses saliency features and gradient magnitude. VSI (Zhang *et al.*, 2014) likewise benefits from saliency-based features and gradient magnitude. SVD based features

(Shnayderman *et al.*, 2006), features based on the Riesz transform (Zhang *et al.*, 2010), features in the wavelet domain (Chandler & Hemami, 2007; Li *et al.*, 2011; Sampat *et al.*, 2009; Rezazadeh & Coulombe, 2013) and sparse features (Chang *et al.*, 2013) are used as well in the literature.

Among these features, gradient magnitude is an efficient feature, as shown in (Xue *et al.*, 2014b). 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 (Wang *et al.*, 2004, 2003; Wang & Li, 2011; Liu *et al.*, 2012; Zhang *et al.*, 2011, 2010). They are robust pooling strategies that usually provide a moderate to high performance for different IQAs. Minkowski pooling (Wang & Shang, 2006), local distortion pooling (Wang & Shang, 2006; Moorthy & Bovik, 2009a; Larson & Chandler, 2010), percentile pooling (Moorthy & Bovik, 2009b) and saliency-based pooling (Zhang & Li, 2012; Zhang *et al.*, 2014) are other possibilities. Standard deviation (SD) pooling was also proposed and successfully used by GMSD (Xue *et al.*, 2014b). 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 (Xue *et al.*, 2014b) 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.

1.1.1 FR quality assessment of tone-mapped images

Tone-mapping operators have been used to convert HDR images into their LDR associated images for visibility purposes on non-HDR displays. Unfortunately, TMO methods perform differently, depending on the HDR image to be converted, which means that the best TMO

method must be found for each individual case. A survey of various TMOs for HDR images and videos is provided in (Yeganeh & Wang, 2013a) and (Eilertsen *et al.*, 2013). Traditionally, TMO performance has been evaluated subjectively. In (Ledda *et al.*, 2005), a subjective assessment was carried out using an HDR monitor. Mantiuk et al. (Mantiuk *et al.*, 2005) propose an HDR visible difference predictor (HDR-VDP) to estimate the visibility differences of two HDR images, and this tool has also been extended to a dynamic range independent image quality assessment (Aydin *et al.*, 2008). However, the authors did not arrive at an objective score, but instead evaluated the performance of the assessment tool on HDR displays. Although subjective assessment provides true and useful references, it is an expensive and time-consuming process. In contrast, the objective quality assessment of tone mapping images enables an automatic selection and parameter tuning of TMOs (Yeganeh & Wang, 2010; Ma *et al.*, 2014). Consequently, objective assessment of tone-mapping images, which is proportional to the subjective assessment of the images, is currently of great interest.

Recently, an objective index, called the tone mapping quality index (TMQI) was proposed by (Yeganeh & Wang, 2013a) to objectively assess the quality of the individual LDR images produced by a TMO. The TMQI is based on combining an SSIM-motivated structural fidelity measure with a statistical naturalness:

$$\mathrm{TMQI}(H,L) = a[S(H,L)]^{\alpha} + (1-a)[N(L)]^{\beta}. \tag{1.8}$$

where S and N denote the structural fidelity and statistical naturalness, respectively. H and L denote the HDR and LDR images. The parameters α and β determine the sensitivities of the two factors, and a ($0 \leq a \leq 1$) adjusts their relative importance. Both S and N are upper bounded by 1, and so the TMQI is also upper bounded by 1 (Ma *et al.*, 2014). Although the TMQI clearly provides better assessment for tone-mapped images than the popular image quality assessment metrics, like SSIM (Wang *et al.*, 2004), MS-SSIM (Wang *et al.*, 2003), and FSIM (Zhang *et al.*, 2011), its performance is not perfect. Liu et al. (Liu *et al.*, 2014b) replaced the pooling strategy of the structural fidelity map in the TMQI with various visual saliency-based strategies for

better quality assessment of tone mapped images. They examined a number of visual saliency models and conclude that integrating saliency detection by combining simple priors (SDSP) into the TMQI provides better assessment capability than other saliency detection models.

1.2 No-reference image quality assessment

As illustrated in section Introduction, NR-IQA models evaluate the perceptual quality of a distorted image without any access to a reference pristine-quality image. NR-IQAs are of high interest because in the most present and emerging practical real-world applications, the reference signals are not available (Wang & Bovik, 2011). NR-IQA metrics perform according to the statistical regularities of natural images in spatial and transformed domains. The deviation between statistical regularities of distortion-free and distorted images is considered in the design of the NR-IQA models. First, general-purpose NR-IQA metrics that are not restricted to the distortion type are briefly explained. These metrics might be inefficient and perform inaccurate predictions for some of the distortion types. Therefore, distortion-specific NR-IQA models have been proposed to accurately predict specific distortions. We will have an overview of these metrics after explaining the general-purpose NR metrics.

The so-called NR-IQA metric DIVIINE (Moorthy & Bovik, 2011a), first classifies distortion types. Then, subband coefficients of discrete wavelet transform (DWT) are fitted by generalized Gaussian distribution (GGD). The statistics of GGD determine the severity of distortions and quality scores are thus estimated by regression. BLIINDS-II (Saad *et al.*, 2012a) is a non-distortion specific NR-IQA metric based on the statistics of the discrete cosine transform (DCT) coefficients. The popular NR-IQA metric BRISQUE (Mittal *et al.*, 2012) uses the statistics of natural images in the spatial domain. The distribution of mean subtracted contrast normalized (MSCN) coefficients in two image scales is fitted by symmetric GGD and asymmetric GGD. MSCN coefficients are widely used by NR-IQA models. Similar to the metric DIIVINE, CurveletQA (Liu *et al.*, 2014a) is also a two stage distortion classification and distortion severity estimation NR-IQA model. It performs according to the statistics of the curvelet coefficients extracted from the images after applying the curvelet transform. BQMS (Gu *et al.*,

2016b) is a NR-IQA metric specifically proposed for the quality assessment of screen content images. It performs by using the principles of free energy theory. Recently, an effective NR-IQA metric called FRIQUEE (Ghadiyaram & Bovik, 2017) was proposed. It is based on a bag of features approach. The metric extracts a large number of features in spatial and frequency domains, and considers color features extracted from different color spaces.

In the following, more details on how the popular NR-IQA metric BRISQUE works is provided. At each image scale, BRISQUE extracts 18 features by analyzing the distribution of mean subtracted contrast normalized (MSCN) coefficients. MSCN coefficients (Daniel L Ruderman, 1994) provides deceleration between image features. The distribution of MSCN coefficients follows the Gaussian distribution. Image distortions change the shape of the MSCN distribution which can be measured and used to predict image quality. Figure 1.2 shows how the shape of MSCN distributions changes for a pristine-quality reference image and four types of distortions. Totally, BRISQUE extracts 36 features and uses support vector regression for model training. Given that BRISQUE performs calculations in the spatial domain, it has an acceptable run-time.

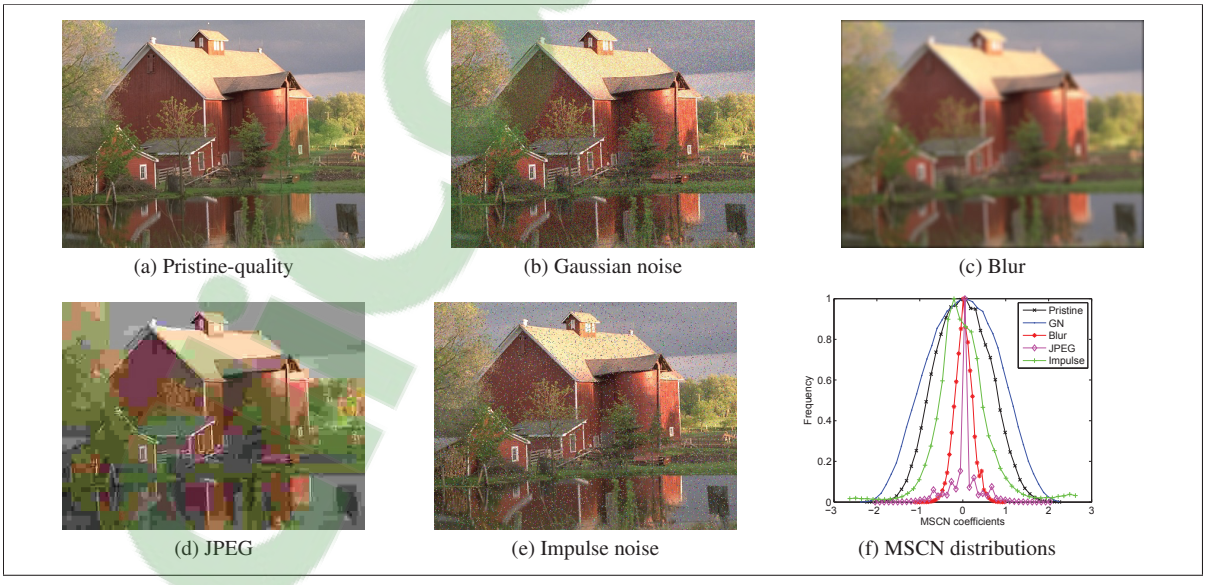

Figure 1.2 MSCN coefficients distributions of a reference image and four distorted versions of that image. Note the difference between the shape of distributions.

1.2.1 NR-IQA of JPEG compressed images

JPEG lossy compression is one of the most common coding techniques to store images. It uses a block-based coding scheme in frequency domain, e.g. discrete cosine transform (DCT), for compression. Since $B \times B$ blocks are coded independent of each other, blocking artifacts are visible in JPEG compressed images specially under low bit rate compression. We have shown two examples of JPEG compressed images in Figure 1.1(b) and Figure 1.2(d).

Several no-reference image quality assessment models (NR-IQAs) have been proposed to objectively assess the quality of the JPEG compressed images (Wu & Yuen, 1997; Tan & Ghanbari, 2000a,b; Wang *et al.*, 2000; Bovik & Liu, 2001; Wang *et al.*, 2002; Pan *et al.*, 2004; Perra *et al.*, 2005; Park *et al.*, 2007; Zhai *et al.*, 2008; Liu & Heynderickx, 2009; Chen & Bloom, 2010; Lee & Park, 2012; Golestaneh & Chandler, 2014; Li *et al.*, 2014a,b, 2015). In (Wu & Yuen, 1997) for each block, horizontal and vertical difference at block boundaries are used to measure horizontal (M_h) and vertical (M_v) blockiness, respectively. The authors in (Tan & Ghanbari, 2000a) proposed a blockiness metric via analysis of harmonics. They used both the amplitude and the phase information of harmonics to compute a quality score. Harmonic analysis was also used to model another blockiness metric in (Tan & Ghanbari, 2000b).

Wang et. al. (Wang *et al.*, 2000) modeled the blocky image as a non-blocky image interfered with a pure blocky signal. Energy of the blocky signal is then used to calculate a quality score. In DCT domain, a metric was proposed in (Bovik & Liu, 2001) that models the blocking artifacts by a 2-D step function. The quality score is calculated following the human vision measurement of block impairments. The metric proposed in (Park *et al.*, 2007) measures blockiness artifact in both the pixel and the DCT domains. In (Golestaneh & Chandler, 2014), zero values DCT coefficients within each block are counted and a relevance map is estimated that distinguishes between naturally uniform blocks and compressed uniform blocks. For this end, an analysis in both DFT and DCT domains is conducted.

Wang et. al. (Wang *et al.*, 2002) proposed an efficient metric that measures blockiness via horizontally and vertically computed features. These features are average differences across block

boundaries, average absolute difference between in-block image samples, and zero crossing rate. Using a set of subjective scores, five parameters of this model are estimated via nonlinear regression analysis. In (Pan *et al.*, 2004), the edge orientation changes of blocks were used to measure severity of blockiness artifacts. Perra et. al. (Perra *et al.*, 2005) analyzed the horizontal, vertical and intra-block sets of 8×8 blocks after applying the Sobel operator to the JPEG compressed images.

The difference of block boundaries plus luminance adaptation and texture masking were used in (Zhai *et al.*, 2008) to form a noticeable blockiness map (NBM) from which, the quality score is calculated by a Minkowski summation pooling. In (Liu & Heynderickx, 2009), 1-D signal profile of gradient image is used to extract block sizes and then priodic peaks in DCT domain are analyzed to calculate a quality score. Chen et. al. (Chen & Bloom, 2010) proposed a very similar metric.

In (Li *et al.*, 2014a), three features including the corners, block boundaries (horizontal, vertical and intra-block), and color changes, together with the subjective scores are used to train a support vector regression (SVR) model. Li et. al. (Li *et al.*, 2014b) measured the blocking artifacts through weighting a set of blockiness scores calculated by Tchebichef moments of different orders.

Lee and Park (Lee & Park, 2012) proposed a blockiness metric that first identifies candidates of having blockiness artifacts. The degree of blockiness of these candidates is then used to compute a quality score. Recently a blockiness metric is proposed that performs in three steps (Li *et al.*, 2015). Block grids are extracted in the spatial domain and their strength and regularity is measured. Afterwards, a masking function is used that gives different weights to the smooth and textured regions.

The aforementioned indices have at least one of the following drawbacks. They might not be robust to block size and block misalignment (examples are (Wang *et al.*, 2002; Pan *et al.*, 2004; Perra *et al.*, 2005; Golestaneh & Chandler, 2014; Li *et al.*, 2014b; Zhai *et al.*, 2008)). They are complex (examples are (Bovik & Liu, 2001; Golestaneh & Chandler, 2014; Li *et al.*, 2015,

2014b,a; Liu & Heynderickx, 2009)), or have many parameters to set ((Wang *et al.*, 2002; Golestaneh & Chandler, 2014; Li *et al.*, 2015; Liu & Heynderickx, 2009; Li *et al.*, 2014b,a)). Indices like NJQA (Golestaneh & Chandler, 2014) and GridSAR (Li *et al.*, 2015) are too slow to compute. Some indices need training ((Wang *et al.*, 2002; Li *et al.*, 2014a)). Also, the range of quality scores provided by some of the indices like (Wang *et al.*, 2002) is not well defined, or they show other numerical issues (Li *et al.*, 2015).

1.2.2 RR and NR-IQA of contrast distorted images

Contrast distortion is commonly produced in image acquisition setup. Poor and varying illumination conditions and poor camera's quality can drastically change image contrast and visibility.

With introduction of quality aware images (Wang *et al.*, 2006), RR-IQAs have shown their usefulness at assessment of image distortions caused by transmission in particular. Prior information about reference image is embedded inside the image to be transmitted, and receiver decodes this information and uses it for quality assessment and even correction of distortions. The resulting metrics that eventually doesn't need training are good examples to illustrate RR-IQAs. In (Gu *et al.*, 2013), a RR-IQA called the RIQMC was proposed to assess the quality of contrast distorted images. RIMQC is a two-step model that uses entropy and four order statistics, e.g. mean, standard deviation, skewness and kurtosis. These features are directly computed from the raw intensity values of contrast distorted images. These are then linearly combined and a quality score is calculated. Seven parameters of the RIQMC are trained based on the 322 images of the CID2013 dataset that also introduced in (Gu *et al.*, 2013). The performance of the RIQMC is very high and at the level of the leading FR-IQA models. The RIQMC was further modified in (Gu *et al.*, 2015a) by computing the phase congruency of the reference and distorted images. In (Gu *et al.*, 2014a), a more efficient RR-IQA called QMC was proposed that uses entropy and saliency features of the reference and distorted images for quality prediction. RCIQM is a more recent RR-IQA model that benefits from a bottom-up and top-down strategy (Liu *et al.*, 2017). It is based on bottom-up analysis of the free energy principle

and top-down analysis of histograms of the reference and distorted images. RCIQM delivers a high performance for quality assessment of the contrast distorted images. The problem with these RR-IQAs is that they necessarily need reference or original image to be available.

There are limited methods in order to assess quality of the contrast distorted images (Fang *et al.*, 2015; Gu *et al.*, 2017). The authors in (Fang *et al.*, 2015) use a natural scene statistics (NSS) induced model to blindly predict the quality of contrast distorted images. They also use five features based on the NSS models of mean, standard deviation, skewness, kurtosis and entropy. Then, support vector regression is utilized to find a mapping function between these five feature set and subjective quality scores. They used 16873 images to train their NSS model. The NR-IQA model in (Gu *et al.*, 2017) called NIQMC takes into account both local and global aspects of the contrast distorted images. In the local part, entropy of salient regions is computed. For the global part, a histogram analysis is proposed. NIQMC provides accurate quality predictions for contrast distorted images. The problem with this method is its high computational time.

1.3 Overview on color to gray conversion methodologies

In many real-world image/video processing and computer vision applications, the 3D color image needs to be transformed into a 1D grayscale image. This is a lossy but a necessary conversion for several applications (Kanan & Cottrell, 2012). Recent years have seen several efforts in developing novel decolorization methods that are more likely to follow human perception of brightness and contrast (Gooch *et al.*, 2005; Neumann *et al.*, 2007; Grundland & Dodgson, 2007; Smith *et al.*, 2008; Kim *et al.*, 2009; Lu *et al.*, 2012; Song *et al.*, 2013, 2014; Du *et al.*, 2015; Liu *et al.*, 2015, 2016; Tao *et al.*, 2017). Color to gray (C2G) conversion methods can be categorized into global, local, and hybrid. The global mapping approach has the potential to produce natural looking grayscale outputs. In contrast, local mapping techniques (Neumann *et al.*, 2007; Smith *et al.*, 2008) that better preserve the local contrast may produce unnatural outputs. In local mapping methods, the same color pixel within an image might be mapped into different grayscale values, which is generally not desired. Therefore, several methods consider

global and local contrast or features for conversion (Kuk *et al.*, 2011; Jin *et al.*, 2014; Du *et al.*, 2015). Besides, video decolorization methods such as (Song *et al.*, 2014; Tao *et al.*, 2017) are specifically developed in order to maintain temporal coherence of videos.

Since the proposed method belongs to the category of global mapping, we focus on these methods. Gooch et al. (Gooch *et al.*, 2005) proposed a method to maintain color contrast between pixel pairs by optimizing an objective contrast function. Kim et al. (Kim *et al.*, 2009) proposed a non-linear parametric model in which the parameters are estimated by minimizing an objective function that preserves color differences. In several recent global mapping methods, the input color image I is converted into a grayscale output g by linear weighting of the R, G, and B channels, i.e. $g(i,j) = \sum_{c=R,G,B} \lambda_c \mathrm{I}_c(i,j)$, where $\sum_{c=R,G,B} \lambda_c = 1$. Here, the three linear weighting parameters λ, should be estimated on the basis of some models. In (Lu *et al.*, 2012), a gradient error energy function is minimized to compute the three linear weighting parameters. This interesting approach was given notable consideration and some variations of this method has been proposed (Liu *et al.*, 2015, 2016). While the method of (Liu *et al.*, 2015) objectively preserves the contrast and run in real-time, it may produce grayscale outputs with an unnatural appearance. In contrast, the method proposed in (Liu *et al.*, 2016) produces mostly natural outputs but at the cost of being several times slower. As mentioned above, several methods in the literature estimate the weighting parameters λ by optimizing an objective function (Gooch *et al.*, 2005; Lu *et al.*, 2012; Liu *et al.*, 2015, 2016). The problem with such methods is that the defined objective function does not necessarily follow human perception of brightness and contrast.

CHAPTER 2

GENERAL METHODOLOGY

In this chapter, we expose our general methodology as well as the rationale. It is in accordance with the main purpose of this thesis, which is to introduce new similarity maps and features and use them to design efficient perceptually consistent image quality assessment models. These features will help to improve image quality assessment performance and to better model human visual perception of image quality. Some of the features that we propose are used in new applications such as quality assessment of tone-mapped images that no or few features already exist. We especially focus on efficient features that can be used in real-time applications. First, to address particular problems posed earlier in this thesis, five objectives are defined to be tackled in this thesis. Then, the general approach of this thesis is explained.

2.1 Research objectives

As stated in the introduction, **the main purpose of this thesis is to introduce new similarity maps and features and use them to design efficient and reliable perceptually consistent image quality assessment models.** It will be achieved with five specific objectives, all of them related to the human visual perception of image quality as in the following:

2.1.1 Objective 1: Develop an effective, efficient and reliable full-reference IQA model with new features and pooling strategy

Existing FR-IQA metrics that measure structural distortion with a gradient similarity, do not mark specific structural distortions caused by color changes. Also, metrics that focus on color distortion measurement are either complex or not reliable to assess different image distortions. Therefore, the first objective is to propose a new gradient similarity which is more likely to follow HVS. This gradient similarity is computed by introducing an intermediate image from the reference and distorted images. In addition, a complementary color similarity map is proposed which is maximally efficient. Common existing metrics use mean pooling to compute

a final quality score from a similarity map. Mean pooling only takes into account the magnitude of distortions. We propose a deviation pooling that considers both magnitude and spread of distortions. The proposed approach is inspired by the human visual perception of image quality, and it will be described in Chapter 3. It provides the first reliable FR-IQA metric with consistent performance for natural, synthetic, and photo-retouched images.

2.1.2 Objective 2: Develop a full-reference IQA model for tone-mapped images

Traditional FR-IQA models cannot be used to assess quality of a tone-mapped low dynamic range (LDR) image with respect to a high dynamic range (HDR) image since they have different dynamic ranges. Previous research converts both images to a color space and calculates their similarity which means that the aforementioned dynamic range problem still exists. Our proposed method is to compare local phase of LDR to local phase of HDR. The rationale is that local phase is mostly based on directional information that we suppose it should have remained fixed in tone-mapped image. Previous studies have shown that HVS has high agreement with phase-derived features. Therefore, the proposed approach is again inspired by concepts of the human visual system. We describe the proposed metric in Chapter 4. The proposed metric provides good performance for two available datasets and has a moderate complexity.

2.1.3 Objective 3: Develop a parameterless no-reference IQA model for JPEG compressed images which is robust to block size and misalignment

Current NR-IQA metrics for JPEG compressed images have major difficulty in dealing with misaligned blocks of JPEG compressed images. Some of the existing metrics are very complex and numerically unstable. Our approach to solving this problem is to introduce a parameterless metric which is robust to misalignment and block size, and it is numerically stable. For this end, simple statistics of edges in JPEG compressed images are considered by the proposed metric. Our approach to assessing blockiness is detailed in Chapter 5. The experimental results on seven natural and synthetic datasets verify high efficiency and performance of the proposed metric.

2.1.4 Objective 4: Propose highly efficient features and develop efficient NR-IQA metric for assessment and classification of contrast distorted images

Existing metrics for NR-IQA of contrast distorted images are either low in performance or complex, and are not able to classify contrast distortion type. Our proposed metric has a very low complexity, it delivers high performance, and it is able to classify contrast distortion type. It benefits from higher orders of Minkowski distance as well as power transformation. The proposed method purposely increases the severity of contrast distortion and perform distortion measurements accordingly. Assessment and classification at the same time are very useful for real-time contrast enhancement. Chapter 6 details the proposed method. Experimental results verify the efficiency of the proposed NR metric.

2.1.5 Objective 5: Propose a perceptually consistent highly efficient color to gray image conversion method

Gray-scale outputs of previous methods of color to gray (C2G) image conversion may not be consistent with the human perception of color and brightness. Some of the C2G methods are computationally inefficient. We propose a highly efficient C2G image conversion that produces perceptually consistent outputs. It is based on the correlation between channels of a color image with a contrast map. A channel with a higher correlation with this contrast map will take larger value for its weighting parameter. At the same time a channel with an inverse correlation with the contrast map takes lower weighting parameter. We conducted a subjective evaluation on the gray-scale outputs of different methods and compared the results with those given by objective metrics. The proposed method and detailed experiments are presented in Chapter 7.

2.2 General approach

New features, metrics, and methods have been proposed in this thesis, all of them are consistent with the human visual system. They can be split into three main themes: full-reference IQA models, no-reference IQA models, and perceptually consistent color to gray conversion.

2.2.1 New full-reference image quality assessment metrics

Two new full-reference IQA metrics have been proposed for low dynamic range images and tone-mapped images. The first metric is a reliable full reference IQA model that utilizes gradient similarity (GS), chromaticity similarity (CS), and deviation pooling (DP). By considering the shortcomings of the commonly used GS to model the 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 thesis 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 (Chapter 3).

For the second full-reference metric, based on the local phase information of images, an objective index, called the feature similarity index for tone-mapped images (FSITM), is proposed. To evaluate a tone mapping operator (TMO), the proposed index compares the locally weighted mean phase angle map of an original high dynamic range (HDR) to that of its associated tone-mapped image calculated using the output of the TMO method. In experiments on two standard databases, it is shown that the proposed FSITM method outperforms the state-of-the-art index, the tone mapped quality index (TMQI) (Yeganeh & Wang, 2013a). In addition, a higher performance is obtained by combining the FSITM and TMQI indices (Chapter 4).

2.2.2 Efficient no-reference image quality assessment metrics

Two new no-reference IQA metrics have been proposed for JPEG compressed and contrast distorted images. We propose a quality assessment model for JPEG compressed images that overcomes several drawbacks of the current blockiness metrics. The proposed index is very simple and efficient, it is parameterless, and robust to block size and misalignment. The proposed metric called MUG is based on two simple facts about blockiness artifact. As a result of more JPEG compression, the number of unique gradient magnitude values decreases, and the median value of unique gradient magnitude values increases. The proposed blockiness metric (MUG) uses these two simple facts to provide accurate quality predictions for JPEG compressed images. Unlike other metrics that presume the position of blocks beforehand or localize the position of blocks, MUG is not a local model and hence does not need any information on the position of blocks. The experimental results on six benchmark datasets of natural images and a benchmark dataset of synthetic images show that MUG is comparable to the state-of-the-art indices in literature. In addition, its performance remains unchanged for the case of the cropped images in which block boundaries are not known (Chapter 5).

The second no-reference metric is an efficient Minkowski Distance based Metric (MDM) for NR quality assessment of contrast distorted images. It is shown that higher orders of Minkowski distance provide accurate quality predictions for the contrast distorted images. The proposed metric performs predictions by extracting only three features from the distorted images followed by a regression analysis. Furthermore, the proposed features are able to classify the type of the contrast distorted images with a high accuracy. Experimental results on the three datasets of CSIQ, TID2013, and CCID2014 show that the proposed metric with a very low complexity provides better quality predictions than the state-of-the-art NR metrics (Chapter 6).

2.2.3 Efficient perceptually consistent color to gray image conversion

A novel decolorization method is proposed in this thesis to convert color images into grayscale. The proposed method, called CorrC2G, estimates the three global linear weighting parameters of the color to gray conversion by correlation. These parameters are estimated directly from the correlations between each channel of the RGB image and a contrast image. The proposed method works directly on the RGB channels; it does not use any edge information nor any optimization or training. The objective and subjective experimental results on three available benchmark datasets of color to gray conversion, e.g. Cadik, CSDD and Color250, show that the proposed decolorization method is highly efficient and comparable to recent state-of-the-art decolorization methods (Chapter 7).

CHAPTER 3

MEAN DEVIATION SIMILARITY INDEX: EFFICIENT AND RELIABLE FULL-REFERENCE IMAGE QUALITY EVALUATOR

Hossein Ziaei Nafchi[1], Atena Shahkolaei[1], Rachid Hedjam[2], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Geography, McGill University,
805 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

Published in IEEE Access
Volume 4, August 2016, Pages 5579-5590

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.

Keywords
Image quality assessment, gradient similarity, chromaticity similarity, deviation pooling, synthetic image, Human visual system.

3.1 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 (Wang, 2011). 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) (Wang *et al.*, 2004) or different dynamic ranges (Yeganeh & Wang, 2013a; Ziaei Nafchi *et al.*, 2015). 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 (Wang & Bovik, 2009). Because of this limitation, a number of IQAs have been proposed to provide better correlation with the HVS (Damera-Venkata *et al.*, 2000; Wang *et al.*, 2004, 2003; Sheikh *et al.*, 2005; Sheikh & Bovik, 2006; Chandler & Hemami, 2007; Zhang *et al.*, 2010; Larson & Chandler, 2010; Narwaria & Lin, 2010; Li & Bovik, 2010; Wang & Li, 2011; Zhang *et al.*, 2011; Liu *et al.*, 2012; Zhang & Li, 2012; Xue *et al.*, 2014b; Chang *et al.*, 2013; Zhang *et al.*, 2014; Gu *et al.*, 2016a; Bae & Kim,

2016b). 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 (Lin & Kuo, 2011). 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 (Wang *et al.*, 2004). 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 (Liu *et al.*, 2012) uses a combination of some designated gradient magnitudes and image contrast for this end, while the GMSD (Xue *et al.*, 2014b) uses only the gradient magnitude. SR_SIM (Zhang & Li, 2012) uses saliency features and gradient magnitude. VSI (Zhang *et al.*, 2014) likewise benefits from saliency-based features and gradient magnitude. SVD based features (Shnayderman *et al.*, 2006), features based on the Riesz transform (Zhang *et al.*, 2010), features in the wavelet domain (Chandler & Hemami, 2007; Li *et al.*, 2011; Sampat *et al.*, 2009) and sparse features (Chang *et al.*, 2013) are used as well in the literature. Among these features, gradient magnitude is an efficient feature, as shown in (Xue *et al.*, 2014b). 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 (Wang *et al.*, 2004, 2003; Wang & Li, 2011; Liu *et al.*, 2012; Zhang *et al.*, 2011, 2010). They are robust pooling strategies that usually provide a moderate to high performance for different IQAs. Minkowski pooling (Wang & Shang, 2006), local distortion pooling (Wang & Shang, 2006; Moorthy & Bovik, 2009a; Larson & Chandler, 2010), percentile pooling (Moorthy & Bovik, 2009b) and saliency-based pooling (Zhang & Li, 2012; Zhang *et al.*, 2014) are other possibilities. Standard deviation (SD) pooling was also proposed and successfully used by GMSD (Xue *et al.*, 2014b). The image gradients are sen-

sitive to image distortions. Different local structures in a distorted image suffer from different degrees of degradations. This is the motivation that the authors in (Xue *et al.*, 2014b) 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 (Chen *et al.*, 2006; Kim *et al.*, 2010; Liu *et al.*, 2012; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Zhang *et al.*, 2014) and chrominance features (Zhang *et al.*, 2011, 2014) 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 (Xue *et al.*, 2014b). 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 (Chen *et al.*, 2006; Kim *et al.*, 2010; Liu *et al.*, 2012; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Zhang *et al.*, 2014) 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 chormaticity similarity map with the following advantages over the previously used chromaticity similarity maps (Zhang *et al.*, 2011, 2014). 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 (Xue *et al.*, 2014b), 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 3.2. Extensive experimental results and discussion on six natural datasets, a synthetic dataset, and a digitally retouched dataset are provided in section 3.3. Section 3.4 presents our conclusions.

3.2 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 (Geusebroek *et al.*, 2001) are used:

$$\begin{bmatrix} H \\ M \end{bmatrix} = \begin{pmatrix} 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.1}$$

3.2.1 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 $G_y(\mathbf{x})$ and horizontal $G_x(\mathbf{x})$ gradients of an image $f(\mathbf{x})$ using convolution: $G_x(\mathbf{x}) = h_x * f(\mathbf{x})$ and $G_y(\mathbf{x}) = h_y * f(\mathbf{x})$, where h_x and h_y are horizontal and vertical gradient operators and $*$ denotes the convolution. The first derivative magnitude is defined as $G(\mathbf{x}) = \sqrt{G_x^2(\mathbf{x}) + G_y^2(\mathbf{x})}$. The Sobel operator (I. Sobel, 1968), 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:

$$\mathrm{GS}(\mathbf{x}) = \frac{2G_R(\mathbf{x})G_D(\mathbf{x}) + C_1}{G_R^2(\mathbf{x}) + G_D^2(\mathbf{x}) + C_1} \tag{3.2}$$

where, parameter C_1 is a constant to control numerical stability. The gradient similarity (GS) is widely used in the literature (Chen *et al.*, 2006; Kim *et al.*, 2010; Liu *et al.*, 2012; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Zhang *et al.*, 2014) and its usefulness to measure image distortions was extensively investigated in (Xue *et al.*, 2014b).

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.

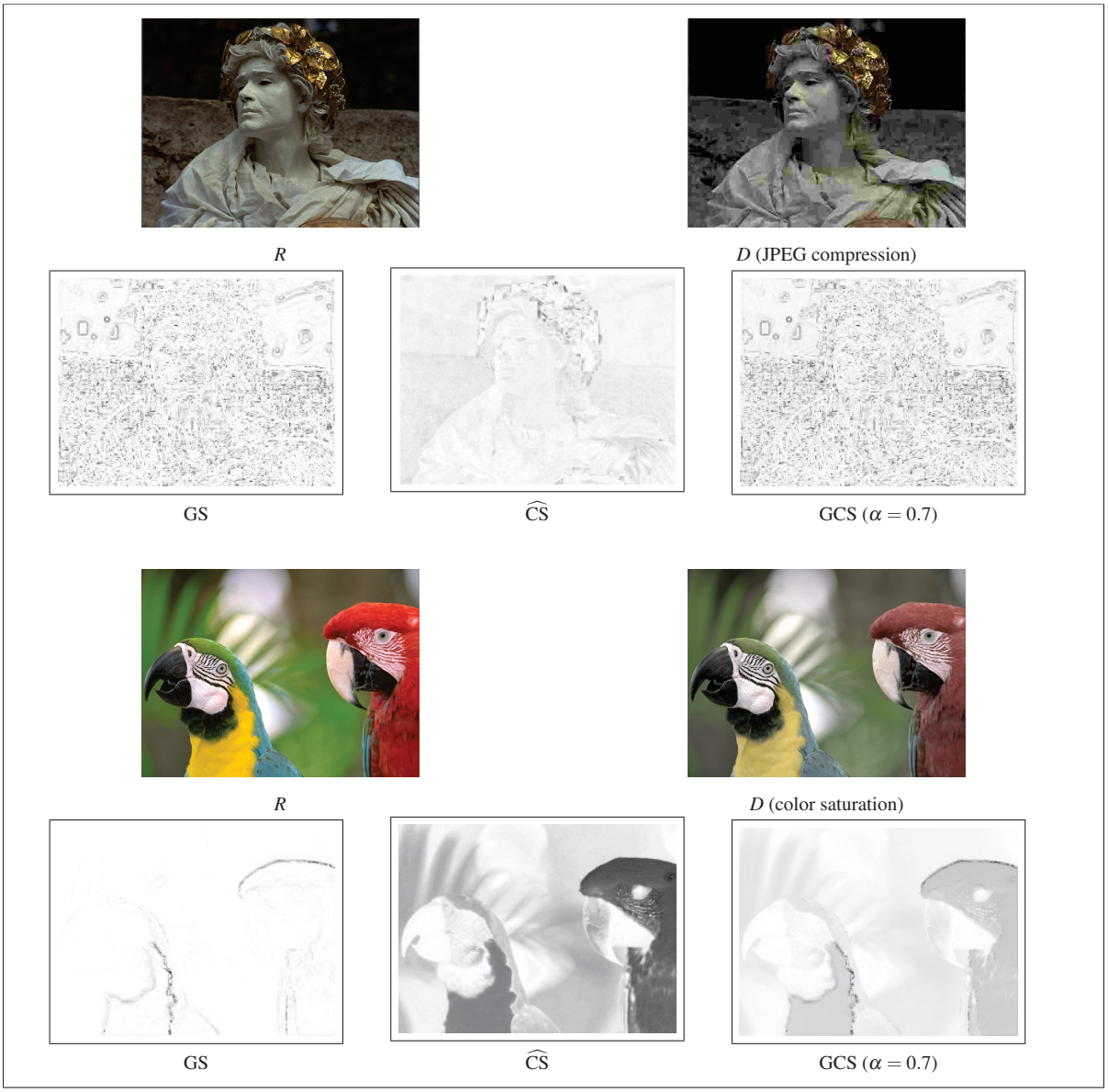

Figure 3.1 Complementary behavior of the gradient similarity (GS) and chromaticity similarity ($\widehat{\text{CS}}$) maps.

3.2.2 The Proposed Gradient Similarity

The aforementioned shortcomings of the conventional gradient similarity map (equation 3.2) 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:

$$\mathrm{GS}_{RF}(\mathbf{x}) = \frac{2G_R(\mathbf{x})G_F(\mathbf{x}) + C_2}{G_R^2(\mathbf{x}) + G_F^2(\mathbf{x}) + C_2} \tag{3.3}$$

$$\mathrm{GS}_{DF}(\mathbf{x}) = \frac{2G_D(\mathbf{x})G_F(\mathbf{x}) + C_2}{G_D^2(\mathbf{x}) + G_F^2(\mathbf{x}) + C_2} \tag{3.4}$$

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 $\mathrm{GS}_{RF}(\mathbf{x})$ and $\mathrm{GS}_{DF}(\mathbf{x})$ can or can not be equal. The proposed gradient similarity ($\widehat{\mathrm{GS}}$) is computed by:

$$\widehat{\mathrm{GS}}(\mathbf{x}) = \mathrm{GS}(\mathbf{x}) + \left[\mathrm{GS}_{DF}(\mathbf{x}) - \mathrm{GS}_{RF}(\mathbf{x})\right]. \tag{3.5}$$

The added term $\left[\mathrm{GS}_{DF}(\mathbf{x}) - \mathrm{GS}_{RF}(\mathbf{x})\right]$, will put more emphasis 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, $\left[\mathrm{GS}_{DF}(\mathbf{x}) - \mathrm{GS}_{RF}(\mathbf{x})\right]$ always put less emphasis on weak edges.

Comparing visually some outputs of the GS and $\widehat{\mathrm{GS}}$ at this step might not be fair because they have different numerical scales. GS is bounded between 0 and 1, while $\widehat{\mathrm{GS}}$ 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 3.2.5 as well as more explanation on how the proposed $\widehat{\text{GS}}$ works.

3.2.3 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 $\widehat{\text{GS}}$ may lead to inaccurate quality predictions. Therefore, previous researches such as (Zhang *et al.*, 2011, 2014) 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 (Zhang *et al.*, 2011, 2014), for each channel a color similarity is computed and their result is combined as:

$$\text{CS}(\mathbf{x}) = \frac{2H_R(\mathbf{x})H_D(\mathbf{x}) + C_3}{H_R^2(\mathbf{x}) + H_D^2(\mathbf{x}) + C_3} \times \frac{2M_R(\mathbf{x})M_D(\mathbf{x}) + C_3}{M_R^2(\mathbf{x}) + M_D^2(\mathbf{x}) + C_3} \tag{3.6}$$

where C_3 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:

$$\widehat{\text{CS}}(\mathbf{x}) = \frac{2\Big(H_R(\mathbf{x})H_D(\mathbf{x}) + M_R(\mathbf{x})M_D(\mathbf{x})\Big) + C_3}{H_R^2(\mathbf{x}) + H_D^2(\mathbf{x}) + M_R^2(\mathbf{x}) + M_D^2(\mathbf{x}) + C_3} \tag{3.7}$$

Similar to the CS in equation (3.6), the above joint color similarity ($\widehat{\text{CS}}$) formulation gives equal weight to both chromaticity channels H and M. It is clear that $\widehat{\text{CS}}$ is more computationally efficient than CS. CS needs 7 multiplications, 6 summations, 2 divisions, and 2 shift operations (multiplications by 2), while $\widehat{\text{CS}}$ needs 6 multiplications, 6 summations, 1 division, and 1 shift operation. Note that 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 $\widehat{\text{CS}}$ along with the proposed metric.

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

$$\text{GCS}(\mathbf{x}) = \alpha \text{GS}(\mathbf{x}) + (1-\alpha)\widehat{\text{CS}}(\mathbf{x}) \tag{3.8}$$

$$\widehat{\text{GCS}}(\mathbf{x}) = \alpha \widehat{\text{GS}}(\mathbf{x}) + (1-\alpha)\widehat{\text{CS}}(\mathbf{x}) \tag{3.9}$$

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 (3.9). Equation (3.8) is included to be compared with equation (3.9). An alternative combination scheme which is very popular in state-of-the-art is through multiplication in the form of $[\widehat{\text{GS}}(\mathbf{x})]^{\gamma}[\widehat{\text{CS}}(\mathbf{x})]^{\beta}$, where the parameters γ and β 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 3.3.5).

In Figure 3.1, two examples are provided to show that these two similarity maps, e.g. GS and $\widehat{\text{CS}}$, 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 Figure 3.1 show that the gradient similarity and chromaticity similarity are complementary.

3.2.4 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 (Xue *et al.*, 2014b) 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 (Xue *et al.*, 2014b), an example is also provided in subsection 3.3.4 to support this statement. Human visual system penalizes more severe 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. Figure 3.2 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 μ_1 than μ_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 the literature, except the standard deviation used in GMSD (Xue *et al.*, 2014b), which is a special case of DP. A deviation can be seen as the Minkowski distance of order ρ between vector $\mathbf{x}$ and its MCT (Measure of Central Tendency):

$$\mathrm{DP}^{(\rho)} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{x}_i - \mathrm{MCT}\right|^{\rho}\right)^{1/\rho}. \tag{3.10}$$

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.

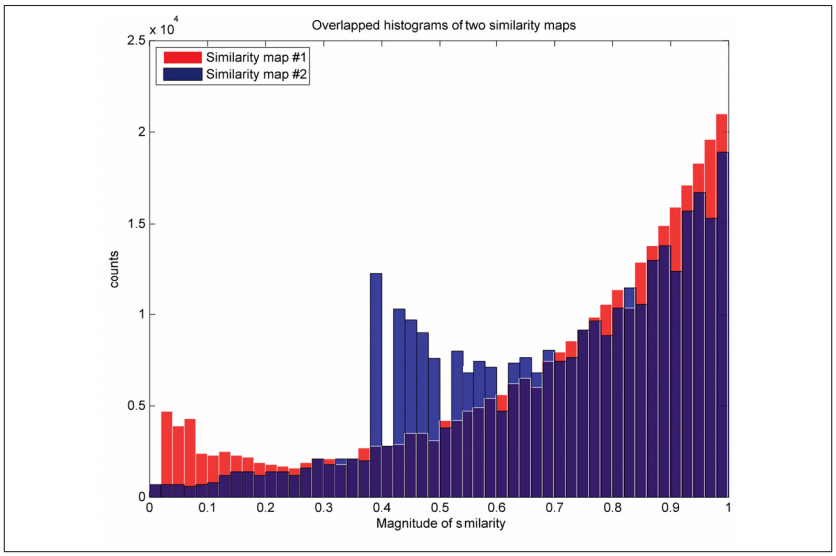

Figure 3.2 Overlapped histograms of two similarity maps corresponding to two distorted images. Lower values of similarity maps indicate to more severe distortions, while higher values refer to less/non distorted pixels.

Several researches have shown that more emphasis on the severer distortions can lead to more accurate predictions (Wang & Shang, 2006; Moorthy & Bovik, 2009b). The Minkowski pooling (Wang & Shang, 2006) and the percentile pooling (Moorthy & Bovik, 2009b) 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 (Wang & Shang, 2006) or discarding the less/non distorted regions (Moorthy & Bovik, 2009b). The deviation pooling can be generalized to consider the aforementioned property of HVS:

$$\mathrm{DP}^{(\rho,q)} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{x}_i^q - \mathrm{MCT}\right|^{\rho}\right)^{1/\rho}. \tag{3.11}$$

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

$$\mathrm{DP}^{(\rho,q,o)} = \left[\left(\frac{1}{N} \sum_{i=1}^{N} \left| \mathbf{x}_i^q - \mathrm{MCT} \right|^{\rho} \right)^{1/\rho} \right]^{o}. \tag{3.12}$$

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 (Xue *et al.*, 2014b). Also, according to (ITU-T P. 1401, 2012), 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 (Wang & Shang, 2006) when $\mathrm{MCT} = 0$, $\rho = 1$ 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:

$$\mathrm{MDSI} = \left[\frac{1}{N} \sum_{i=1}^{N} \left| \widehat{\mathrm{GCS}}_i^{1/4} - \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\mathrm{GCS}}_i^{1/4} \right) \right| \right]^{1/4}. \tag{3.13}$$

Note that possible interval for $\widehat{\mathrm{GCS}}$ is $[0-\delta_1 \; 1+\delta_2]$, where $\delta_1 < 1$ and $\delta_2 < 1$. It is worth to mention that values of $\widehat{\mathrm{GCS}}$ mostly remain in $[0 \; 1]$. Also, $\widehat{\mathrm{GCS}} < 0$ are highly distorted pixels, while $\widehat{\mathrm{GCS}} > (1-\varepsilon)$ refer to less/non-distorted pixels, where $\varepsilon < 1$ is a very small number. The global variations of $\widehat{\mathrm{GCS}}^{1/4}$ 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

[1] Though linearity is measured after a nonlinear analysis.

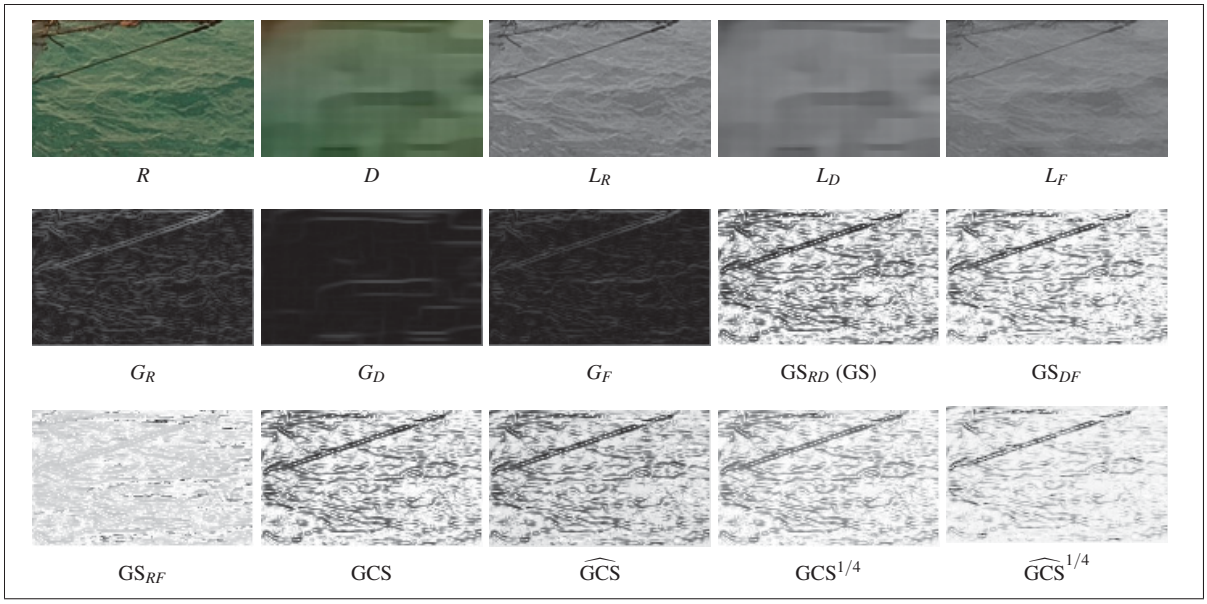

Figure 3.3 The difference between similarity maps GCS and $\widehat{\text{GCS}}$ that use conventional gradient similarity and the proposed gradient similarity, respectively.

the Minkowski pooling on final similarity maps is that terms like “more emphasis" and “less emphasis", regardless of the q values have been used, depends also on the pooling strategy and underlying similarity map. For example, placing more emphasis 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.

3.2.5 Analysis and Examples of GCS Maps

In this section, final similarity maps after applying the Minkowski pooling, e.g. $\text{GCS}^{1/4}$ and $\widehat{\text{GCS}}^{1/4}$, are compared along with sufficient explanations. The difference between these two similarity map is their use of gradient similarity. GCS uses conventional GS, while $\widehat{\text{GCS}}$ uses the proposed gradient similarity $\widehat{\text{GS}}$. 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 3.2.1, several disadvantages of the traditional GS was mentioned. Here, each of them are explained and examples are provided.

Case 1 (Removed edge): Missing edges in distorted image with respect to its original image means that structural information are removed, hence this disappearance brings attention of the HVS. These regions have to be strongly highlighted in the similarity map.

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

Figure 3.3 shows how the proposed gradient similarity map $\widehat{\mathrm{GS}}$ performs for *case 1* and *case 2* as a part of the $\widehat{\mathrm{GCS}}$ compared to the GS for GCS. We can see that 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_F), some of these weaker edges are smoothed. This can be seen by comparing GS_{RD} and GS_{DF}. Both GS_{RD} and GS_{DF} indicate high differences at the location of the ropes. $\mathrm{GS}_{RD} + \mathrm{GS}_{DF}$ will also put high emphasis on this location, but less emphasis on the weaker edges. The results is then subtracted by GS_{RF} which in turn again less emphasize is placed on the weak edges (relevant to the darker pixels in GS_{RF}). Note that GCS and $\widehat{\mathrm{GCS}}$ have different numerical behavior, so it is fair to compare them by looking at the $\mathrm{GCS}^{1/4}$ and $\widehat{\mathrm{GCS}}^{1/4}$. Compared to the $\mathrm{GCS}^{1/4}$, $\widehat{\mathrm{GCS}}^{1/4}$ indicate to larger differences at the location of ropes, but smaller differences elsewhere.

Case 3 (Preserved edge but with different color): Although a color similarity map should measure color differences at the location of the inverted edges, edges constitute a small fraction of the total pixels in images, and it is common to give smaller weights to a color similarity map than structural similarities such as gradient similarity. While traditional gradient similarity does not work well in this situation, the proposed gradient similarity can partially solve this problem. Figure 3.4 provides an example in which most of the edges are inverted in the distorted image. We can see that $\widehat{\mathrm{GCS}}^{1/4}$ highlighted much more differences than $\mathrm{GCS}^{1/4}$ at these locations, thanks to the added term $(\mathrm{GS}_{DF} - \mathrm{GS}_{RF})$ to the traditional gradient similarity. In fact, GS_{DF}

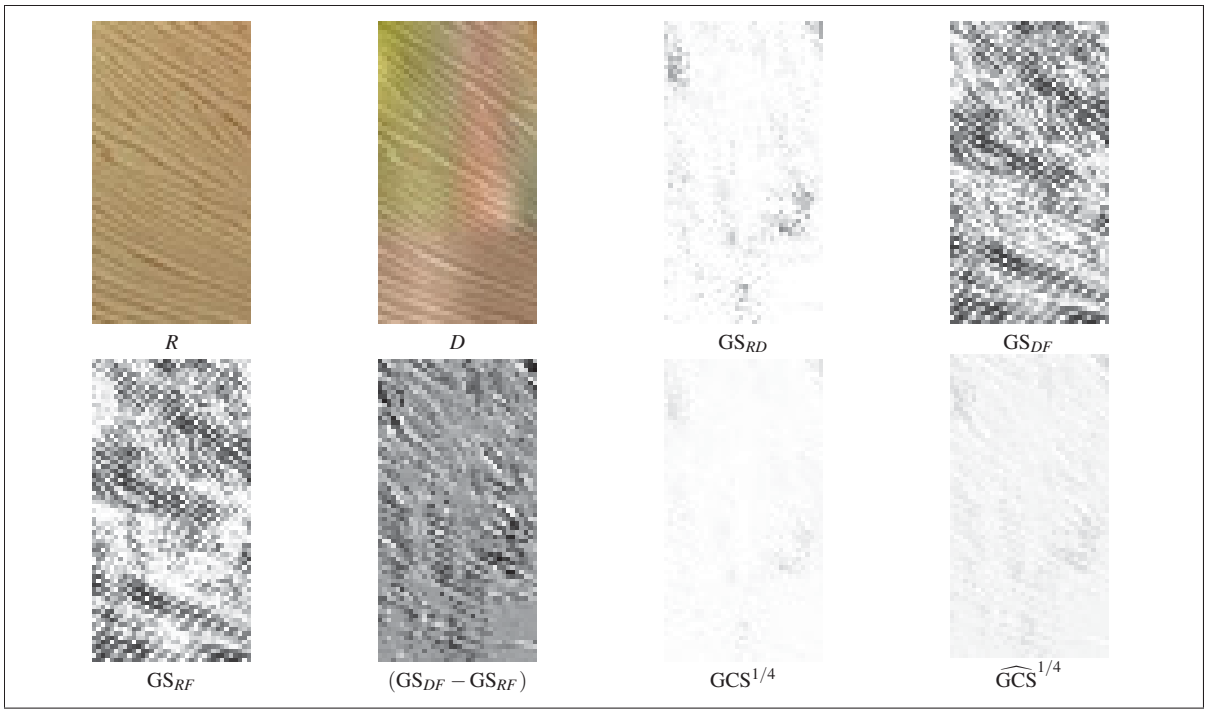

Figure 3.4 The difference between similarity maps $\mathrm{GCS}^{1/4}$ and $\widehat{\mathrm{GCS}}^{1/4}$ for the case of the inverted edges. Note that some intermediate outputs are not shown.

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.

3.3 Experimental results and discussion

In the experiments, eight datasets were used. The LIVE dataset (Sheikh *et al.*) contains 29 reference images and 779 distorted images of five categories. The TID2008 (Ponomarenko *et al.*, 2009) dataset contains 25 reference images and 1700 distorted images. For each reference image, 17 types of distortions of 4 degrees are available. CSIQ (Larson & Chandler, 2010) 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 (Ponomarenko *et al.*, 2013) 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 (Zaric *et al.*, 2012) 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 (Gu *et al.*, 2015a) 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 (Kundu & Evans, 2015) which contains 25 synthetic images of video games and animated movies. It contains 500 distorted images of 5 categories. Figure 3.5 shows an example of a reference and a distorted synthetic image. Finally, the digitally retouched image quality (DRIQ) dataset (Vu *et al.*, 2012) was used in the experiments. It contains 26 reference images and 3 enhanced images for each reference image.

R *D* (Gaussian noise)

Figure 3.5 An example of reference *R* and distorted *D* image in the ESPL synthetic images database (Kundu & Evans, 2015).

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 3.14), 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.

Table 3.1 Performance comparison of the proposed IQA model, MDSI, and twelve popular/competing indices on eight benchmark datasets. Note that top three IQA models are highlighted.

		MSSSIM	VIF	MAD	IWSSIM	SR_SIM	$FSIM_c$	GMSD	SFF	VSI	DSCSI	ADD-GSIM	SCQI	MDSI
TID 2008	SRC	0.8542	0.7491	0.8340	0.8559	0.8913	0.8840	0.8907	0.8767	0.8979	0.8634	**0.9094**	**0.9051**	**0.9208**
	PCC	0.8451	0.8084	0.8290	0.8579	0.8866	0.8762	0.8788	0.8817	0.8762	0.8445	**0.9120**	**0.8899**	**0.9160**
	KRC	0.6568	0.5861	0.6445	0.6636	0.7149	0.6991	0.7092	0.6882	0.7123	0.6651	**0.7389**	**0.7294**	**0.7515**
	RMSE	0.7173	0.7899	0.7505	0.6895	0.6206	0.6468	0.6404	0.6333	0.6466	0.7187	**0.5504**	**0.6120**	**0.5383**
CSIQ	SRC	0.9133	0.9195	0.9467	0.9213	0.9319	0.9310	**0.9570**	**0.9627**	0.9423	0.9417	0.9422	0.9434	**0.9569**
	PCC	0.8991	0.9277	0.9500	0.9144	0.9250	0.9192	**0.9541**	**0.9643**	0.9279	0.9313	0.9342	0.9268	**0.9531**
	KRC	0.7393	0.7537	0.7970	0.7529	0.7725	0.7690	**0.8129**	**0.8288**	0.7857	0.7787	0.7894	0.7870	**0.8130**
	RMSE	0.1149	0.0980	0.0820	0.1063	0.0997	0.1034	**0.0786**	**0.0695**	0.0979	0.0956	0.0937	0.0986	**0.0795**
LIVE	SRC	0.9513	0.9636	**0.9669**	0.9567	0.9618	0.9645	0.9603	0.9649	0.9524	0.9487	**0.9681**	0.9406	**0.9667**
	PCC	0.9489	0.9604	**0.9675**	0.9522	0.9553	0.9613	0.9603	0.9632	0.9482	0.9434	**0.9657**	0.9344	**0.9659**
	KRC	0.8044	0.8282	**0.8421**	0.8175	0.8299	0.8363	0.8268	0.8365	0.8058	0.7982	**0.8474**	0.7835	**0.8395**
	RMSE	8.6188	7.6137	**6.9072**	8.3472	8.0812	7.5296	7.6214	7.3460	8.6817	9.0635	**7.0925**	9.7355	**7.0790**
TID 2013	SRC	0.7859	0.6769	0.7807	0.7779	0.8073	0.8510	0.8044	0.8513	**0.8965**	0.8744	0.8285	**0.9052**	**0.8899**
	PCC	0.8329	0.7720	0.8267	0.8319	0.8663	0.8769	0.8590	0.8706	**0.9000**	0.8782	0.8807	**0.9071**	**0.9085**
	KRC	0.6047	0.5147	0.6035	0.5977	0.6406	0.6665	0.6339	0.6581	**0.7183**	0.6862	0.6646	**0.7327**	**0.7123**
	RMSE	0.6861	0.7880	0.6976	0.6880	0.6193	0.5959	0.6346	0.6099	**0.5404**	0.5930	0.5871	**0.5219**	**0.5181**
VCL@ FER	SRC	0.9227	0.8866	0.9061	0.9163	0.9021	**0.9323**	0.9177	0.7738	0.9317	0.9289	**0.9366**	0.9083	**0.9318**
	PCC	0.9232	0.8938	0.9053	0.9191	0.9023	0.9329	0.9176	0.7761	0.9320	**0.9338**	**0.9339**	0.9107	**0.9349**
	KRC	0.7497	0.6924	0.7213	0.7372	0.7183	**0.7643**	0.7406	0.5779	0.7633	0.7588	**0.7731**	0.7316	**0.7629**
	RMSE	9.4398	11.014	10.433	9.6788	10.589	8.8480	9.7643	15.488	8.9051	**8.7902**	**8.7819**	10.147	**8.7136**
CCID 2014	SRC	0.7770	**0.8349**	0.7451	0.7811	0.7363	0.7657	0.8077	0.6859	0.7734	0.7400	**0.8698**	0.7811	**0.8128**
	PCC	0.8278	**0.8588**	0.7516	0.8342	0.7834	0.8204	0.8521	0.7575	0.8209	0.7586	**0.8935**	0.8200	**0.8576**
	KRC	0.5845	**0.6419**	0.5490	0.5898	0.5372	0.5707	0.6100	0.5012	0.5735	0.5468	**0.6840**	0.5812	**0.6181**
	RMSE	0.3668	**0.3350**	0.4313	0.3606	0.4064	0.3739	0.3422	0.4269	0.3734	0.4260	**0.2936**	0.3734	**0.3363**
ESPL	SRC	0.7247	0.7488	0.8624	0.8270	**0.8802**	**0.8766**	0.8209	0.8127	0.8717	0.7263	0.7828	0.8292	**0.8806**
	PCC	0.7322	0.7423	0.8677	0.8300	**0.8732**	**0.8738**	0.8234	0.8179	0.8726	0.7302	0.7902	0.8356	**0.8802**
	KRC	0.5208	0.5565	0.6720	0.6221	**0.6932**	**0.6853**	0.6178	0.6127	0.6765	0.5222	0.5814	0.6243	**0.6895**
	RMSE	9.4519	9.2985	6.8985	7.7404	**6.7646**	**6.7482**	7.8753	7.9844	6.7791	9.4815	8.5053	7.6241	**6.5862**
DRIQ	SRC	0.6692	0.8078	0.6867	0.6903	0.7551	0.7751	0.7762	**0.8342**	0.8222	0.8167	0.7661	**0.8482**	**0.8508**
	PCC	0.7058	**0.8496**	0.6967	0.7155	0.8027	0.7989	0.8001	0.8420	0.8477	0.8463	0.8053	**0.8638**	**0.8702**
	KRC	0.4739	0.5997	0.4898	0.4952	0.5604	0.5771	0.5758	**0.6477**	0.6177	0.6104	0.5618	**0.6490**	**0.6557**
	RMSE	1.4450	**1.0759**	1.4631	1.4249	1.2165	1.2268	1.2235	1.1004	1.0820	1.0864	1.2092	**1.0277**	**1.0050**
Direct Avg.	SRC	0.8248	0.8234	0.8411	0.8408	0.8583	0.8725	0.8669	0.8453	**0.8860**	0.8550	0.8754	**0.8826**	**0.9013**
	PCC	0.8394	0.8516	0.8493	0.8569	0.8743	0.8824	0.8807	0.8591	**0.8907**	0.8583	**0.8894**	0.8860	**0.9108**
	KRC	0.6418	0.6466	0.6649	0.6595	0.6834	0.6960	0.6909	0.6689	**0.7067**	0.6708	**0.7051**	0.7023	**0.7303**
Weighted Avg.	SRC	0.8335	0.7783	0.8374	0.8387	0.8578	0.8769	0.8626	0.8585	**0.8974**	0.8698	0.8783	**0.8977**	**0.9066**
	PCC	0.8521	0.8287	0.8546	0.8626	0.8810	0.8877	0.8838	0.8729	0.8963	0.8679	**0.8995**	**0.8967**	**0.9160**
	KRC	0.6511	0.6112	0.6626	0.6587	0.6880	0.6999	0.6913	0.6791	**0.7206**	0.6855	0.7140	**0.7231**	**0.7375**

Twelve state-of-the-art IQA models were chosen for comparison (Wang *et al.*, 2003; Sheikh & Bovik, 2006; Larson & Chandler, 2010; Wang & Li, 2011; Zhang & Li, 2012; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Chang *et al.*, 2013; Zhang *et al.*, 2014; Lee & Plataniotis, 2015; Gu *et al.*, 2016a; Bae & Kim, 2016b) including the most recent indices in literature (Chang *et al.*, 2013; Xue *et al.*, 2014b; Zhang *et al.*, 2014; Lee & Plataniotis, 2015; Gu *et al.*, 2016a; Bae & Kim, 2016b). It should be noted that the five indices SFF (Chang *et al.*, 2013), GMSD (Xue *et al.*, 2014b), VSI (Zhang *et al.*, 2014), (Gu *et al.*, 2016a), and SCQI (Bae & Kim, 2016b) have shown superior performance over state-of-the-art indices.

3.3.1 Performance comparison

In Table 3.1, 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 32 times among the top indices (everywhere), followed by ADD-GSIM (16 times), SCQI (12 times), SFF/FSIM$_c$/VIF (6 times), VSI/GMSD/SR_SIM/MAD[2] (4 times), DSCSI (2 times), and MSSSIM/IWSSIM (0 times). To provide a conclusion on the overall performance of these indices, direct and weighted[3] overall performances on the eight datasets (8150 images) are also listed in Table 3.1. It can be seen that MDSI has the best overall performance on the eight datasets, while metrics VSI and SCQI are the second, and third best, respectively.

3.3.2 Visualization and statistical evaluation

For the purpose of visualizing quality scores of the proposed index, the scatter plots of the proposed IQA model MDSI with and without using power pooling are shown in Figure 3.6. The logistic function suggested in (Sheikh *et al.*, 2006) was used to fit a curve on each plot:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5 \tag{3.14}$$

where β_1, β_2, β_3, β_4 and β_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 3.1 show the difference between different IQA models. As suggested in (Video Quality Experts Group, 2003; Sheikh *et al.*, 2006), 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

[2] Note the conflict between 'MAD' (Larson & Chandler, 2010) as an IQA model, and 'MAD' as a pooling strategy.

[3] The dataset size-weighted average is commonly used in the literature (Wang & Li, 2011; Chang *et al.*, 2013; Xue *et al.*, 2014b; Zhang *et al.*, 2014).

Table 3.2 The results of statistical significance test for ten IQA models on eight datasets. The result of the F-test is equal to +1 if a metric is significantly better than another metric, it is equal to -1 if that metric is statistically inferior to another metric, and the result is equal to 0 if two metrics are statistically indistinguishable. The cumulative sum of individual tests for each metric is listed in the last column with top three IQA models being highlighted in the same column.

▷TID2008		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	-	-1	-1	-1	-1	-1	-1	-1	-1	-1	-9
2	MAD	**+1**	-	-1	-1	-1	-1	-1	-1	-1	-1	-7
3	SR_SIM	**+1**	**+1**	-	**+1**	**+1**	**+1**	**+1**	-1	-1	-1	+3
4	FSIM$_c$	**+1**	**+1**	-1	-	-1	-1	**+1**	-1	-1	-1	-3
5	GMSD	**+1**	**+1**	-1	**+1**	-	-1	**+1**	-1	-1	-1	-1
6	SFF	**+1**	**+1**	-1	**+1**	**+1**	-	**+1**	-1	-1	-1	+1
7	VSI	**+1**	**+1**	-1	0	-1	-1	-	-1	-1	-1	-4
8	ADD-GSIM	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+1**	-1	**+7**
9	SCQI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-1	-	-1	**+5**
10	MDSI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+9**

▷CSIQ		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	**-**	-1	**+1**	**+1**	-1	-1	0	-1	0	-1	-3
2	MAD	+1	**-**	**+1**	**+1**	-1	-1	**+1**	**+1**	**+1**	-1	+3
3	SR_SIM	-1	-1	**-**	**+1**	-1	-1	-1	-1	-1	-1	-7
4	FSIM$_c$	-1	-1	-1	**-**	-1	-1	-1	-1	-1	-1	-9
5	GMSD	**+1**	**+1**	**+1**	**+1**	**-**	-1	**+1**	**+1**	**+1**	0	**+6**
6	SFF	**+1**	**+1**	**+1**	**+1**	**+1**	**-**	**+1**	**+1**	**+1**	**+1**	**+9**
7	VSI	0	-1	**+1**	**+1**	-1	-1	**-**	-1	0	-1	-3
8	ADD-GSIM	**+1**	-1	**+1**	**+1**	-1	-1	**+1**	**-**	**+1**	-1	+1
9	SCQI	0	-1	**+1**	**+1**	-1	-1	0	-1	**-**	-1	-3
10	MDSI	**+1**	**+1**	**+1**	**+1**	0	-1	**+1**	**+1**	**+1**	**-**	**+6**

▷LIVE		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	**-**	-1	**+1**	0	0	-1	**+1**	-1	**+1**	-1	-1
2	MAD	**+1**	**-**	**+1**	**+1**	**+1**	**+1**	**+1**	0	**+1**	0	**+7**
3	SR_SIM	-1	-1	**-**	-1	-1	-1	**+1**	-1	**+1**	-1	-5
4	FSIM$_c$	0	-1	**+1**	**-**	0	0	**+1**	-1	**+1**	-1	0
5	GMSD	0	-1	**+1**	0	**-**	-1	**+1**	-1	**+1**	-1	-1
6	SFF	**+1**	-1	**+1**	0	**+1**	**-**	**+1**	-1	**+1**	-1	+2
7	VSI	-1	-1	-1	-1	-1	-1	**-**	-1	**+1**	-1	-7
8	ADD-GSIM	**+1**	0	**+1**	**+1**	**+1**	**+1**	**+1**	**-**	**+1**	0	**+7**
9	SCQI	-1	-1	-1	-1	-1	-1	-1	-1	**-**	-1	-9
10	MDSI	**+1**	0	**+1**	**+1**	**+1**	**+1**	**+1**	0	**+1**	**-**	**+7**

▷TID2013		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	**-**	-1	-1	-1	-1	-1	-1	-1	-1	-1	-9
2	MAD	**+1**	**-**	-1	-1	-1	-1	-1	-1	-1	-1	-7
3	SR_SIM	**+1**	**+1**	**-**	-1	**+1**	-1	-1	-1	-1	-1	-3
4	FSIM$_c$	**+1**	**+1**	**+1**	**-**	**+1**	**+1**	-1	-1	-1	-1	+1
5	GMSD	**+1**	**+1**	-1	-1	**-**	-1	-1	-1	-1	-1	-5
6	SFF	**+1**	**+1**	**+1**	-1	**+1**	**-**	-1	0	-1	-1	0
7	VSI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**-**	**+1**	-1	-1	**+5**
8	ADD-GSIM	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-1	-	-1	-1	+3
9	SCQI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	-1	**+7**
10	MDSI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+9**

Table 3.2 The results of statistical significance test for ten IQA models on eight datasets. The result of the F-test is equal to +1 if a metric is significantly better than another metric, it is equal to -1 if that metric is statistically inferior to another metric, and the result is equal to 0 if two metrics are statistically indistinguishable. The cumulative sum of individual tests for each metric is listed in the last column with top three IQA models being highlighted in the same column (continued).

▷VCL @FER		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	-	-1	-1	-1	-1	**+1**	-1	-1	-1	-1	-7
2	MAD	**+1**	-	**+1**	-1	-1	**+1**	-1	-1	-1	-1	-3
3	SR_SIM	**+1**	-1	-	-1	-1	**+1**	-1	-1	-1	-1	-5
4	$FSIM_c$	**+1**	**+1**	**+1**	-	**+1**	**+1**	0	0	**+1**	0	**+6**
5	GMSD	**+1**	**+1**	**+1**	-1	-	**+1**	-1	-1	**+1**	-1	+1
6	SFF	-1	-1	-1	-1	-1	-	-1	-1	-1	-1	-9
7	VSI	**+1**	**+1**	**+1**	0	**+1**	**+1**	-	0	**+1**	-1	+5
8	ADD-GSIM	**+1**	**+1**	**+1**	0	**+1**	**+1**	0	-	**+1**	0	**+6**
9	SCQI	**+1**	**+1**	**+1**	-1	-1	**+1**	-1	-1	-	-1	-1
10	MDSI	**+1**	**+1**	**+1**	0	**+1**	**+1**	**+1**	0	**+1**	-	**+7**

▷CCID2014		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	-	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-1	**+1**	0	**+6**
2	MAD	-1	-	-1	-1	-1	-1	-1	-1	-1	-1	-9
3	SR_SIM	-1	**+1**	-	-1	-1	**+1**	-1	-1	**+1**	-1	-3
4	$FSIM_c$	-1	**+1**	**+1**	-	-1	**+1**	0	-1	**+1**	-1	0
5	GMSD	-1	**+1**	**+1**	**+1**	-	**+1**	**+1**	-1	**+1**	-1	+3
6	SFF	-1	**+1**	-1	-1	-1	-	-1	-1	0	-1	-6
7	VSI	-1	**+1**	**+1**	0	-1	**+1**	-	-1	**+1**	-1	0
8	ADD-GSIM	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+1**	**+1**	**+9**
9	SCQI	-1	**+1**	-1	-1	-1	0	-1	-1	-	-1	-6
10	MDSI	0	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-1	**+1**	-	**+6**

▷ESPL		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	-	-1	-1	-1	-1	-1	-1	-1	-1	-1	-9
2	MAD	**+1**	-	-1	-1	**+1**	**+1**	-1	**+1**	**+1**	-1	+1
3	SR_SIM	**+1**	**+1**	-	0	**+1**	**+1**	**+1**	**+1**	**+1**	-1	**+6**
4	$FSIM_c$	**+1**	**+1**	0	-	**+1**	**+1**	**+1**	**+1**	**+1**	-1	**+6**
5	GMSD	+1	-1	-1	-1	-	**+1**	-1	**+1**	-1	-1	-3
6	SFF	**+1**	-1	-1	-1	-1	-	-1	**+1**	-1	-1	-5
7	VSI	**+1**	**+1**	-1	-1	**+1**	**+1**	-	**+1**	**+1**	-1	+3
8	ADD-GSIM	**+1**	-1	-1	-1	-1	-1	-1	-	-1	-1	-7
9	SCQI	**+1**	-1	-1	-1	**+1**	**+1**	-1	**+1**	-	-1	-1
10	MDSI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+9**

▷DRIQ		1	2	3	4	5	6	7	8	9	10	sum
1	VIF	-	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	0	-1	**+5**
2	MAD	-1	-	-1	-1	-1	-1	-1	-1	-1	-1	-9
3	SR_SIM	-1	**+1**	-	0	0	-1	-1	-1	-1	-1	-5
4	$FSIM_c$	-1	**+1**	0	-	0	-1	-1	-1	-1	-1	-5
5	GMSD	-1	**+1**	0	0	-	-1	-1	-1	-1	-1	-5
6	SFF	-1	**+1**	**+1**	**+1**	**+1**	-	0	**+1**	-1	-1	+2
7	VSI	-1	**+1**	**+1**	**+1**	**+1**	0	-	**+1**	-1	-1	+2
8	ADD-GSIM	-1	**+1**	**+1**	**+1**	**+1**	-1	-1	-	-1	-1	-1
9	SCQI	0	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	-1	**+6**
10	MDSI	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	-	**+9**

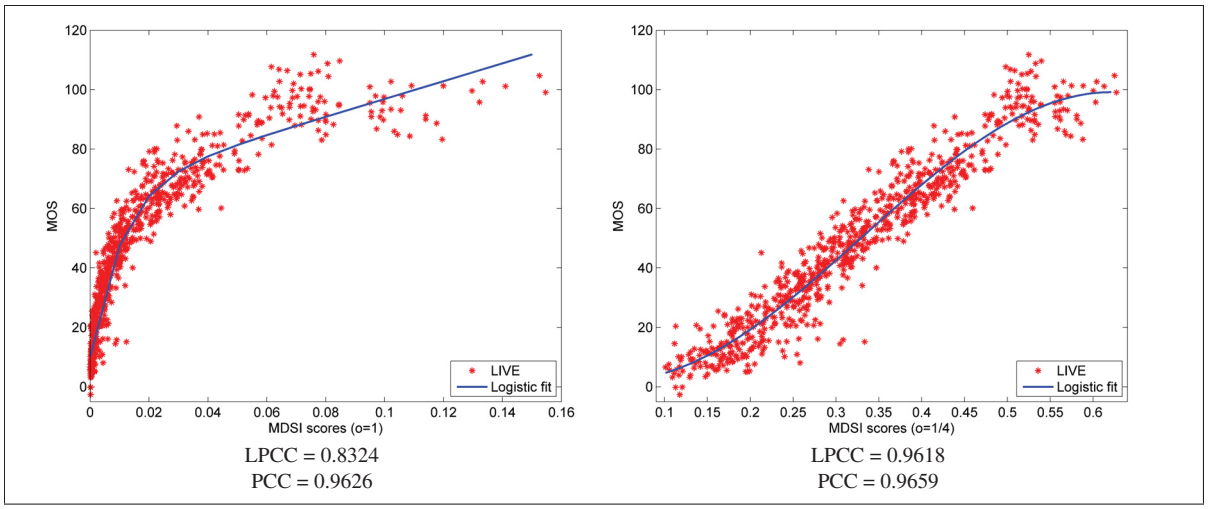

Figure 3.6 Scatter plots of quality scores against the subjective MOS on the LIVE dataset for the proposed model MDSI with and without using the power pooling. Comparison of LPCC and PCC values indicate that MDSI becomes more linear with respect to MOS (the right plot) by using the power pooling.

equation (3.14), 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 3.2. 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 type I error might be occurred, specially when quality scores of IQA models are not Gaussian. However, even existence of possible errors is very unlikely to result in another conclusion about the superiority of the proposed index because there is a considerable gap between the proposed index and the other metrics as discussed in the following.

From the results of Table 3.2, we can see that MDSI is significantly better than the other indices on TID2008, TID2013, ESPL, and DRIQ datasets. Therefore, 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 SFF 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), $FSIM_c$ (-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.

3.3.3 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 3.3 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 to 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 can not be reliable.

The proposed index MDSI, has the best SRC, and PCC average on distortion types. MDSI, SCQI and $FSIM_c$ in the worst case perform better than the 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 3.3 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 3.3, MDSI, SCQI

and FSIM$_c$ have the lowest variation. Therefore, we can conclude that indices MDSI, SCQI and FSIM$_c$ are more reliable than the other indices.

Table 3.3 Overall performance comparison of the proposed IQA model MDSI and twelve popular/competing indices on individual distortion types of six datasets (TID2008, CSIQ, LIVE, TID2013, VCL@FER, and ESPL). The six datasets contain 61 distortion set, therefore results on distortion types are reported based on average of 61 correlation values. Top three IQA models are highlighted.

IQA model	SRC (Distortions)			PCC (Distortions)		
	avg	min	std	avg	min	std
MSSSIM	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	0.8568	-0.4503	0.1962
SR_SIM	0.8609	-0.2053	0.1806	0.8785	-0.3162	0.1839
FSIM$_c$	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**	**0.4378**	**0.1030**	**0.9095**	**0.6899**	**0.0805**

Table 3.4 Performance of the proposed index MDSI with different pooling strategies and values of parameter q.

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

3.3.4 Parameters of deviation pooling (ρ, q, o)

Considering the formulation of deviation pooling in equation (3.12), 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 3.4, 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 $q \leq 1$ 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 Figure 3.6. Power pooling can be also used to increase linearity of other indices as well. For example, LPCC and PCC values of VSI (Zhang *et al.*, 2014) for TID2013 dataset, by setting $o = 18$, can be increased from 0.8373 to 0.8928, and 0.9000 to 0.9011, respectively.

3.3.5 Summation vs. Multiplication

Two options for combination of the two similarity maps GS/$\widehat{\text{GS}}$ and $\widehat{\text{CS}}$ are summation and multiplication as explained in subsection 3.2.3. 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 $\widehat{\text{GS}}$ and $\widehat{\text{CS}}$ 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 3.5 justifies our choice.

Table 3.5 Different 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	✓	

3.3.6 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 α. 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 α. For an example, we refer to the SSIM index (Wang *et al.*, 2004) 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 Figure 3.7, 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 $C_3 = 550$.

3.3.7 Effect of chromaticity similarity maps CS and $\widehat{\text{CS}}$

In this section, the impact of using CS (Zhang *et al.*, 2014) and proposed $\widehat{\text{CS}}$ on the performance of the proposed index is studied through the following experiment. Contrast distorted images of the CCID2014 dataset (Gu *et al.*, 2015a) were chosen. The reason of choosing this dataset is to evaluate the ability of measuring color changes by CS and $\widehat{\text{CS}}$. We analyzed the SRC performance of the CS and $\widehat{\text{CS}}$ as a part of the proposed index for wide range of C_3 val-

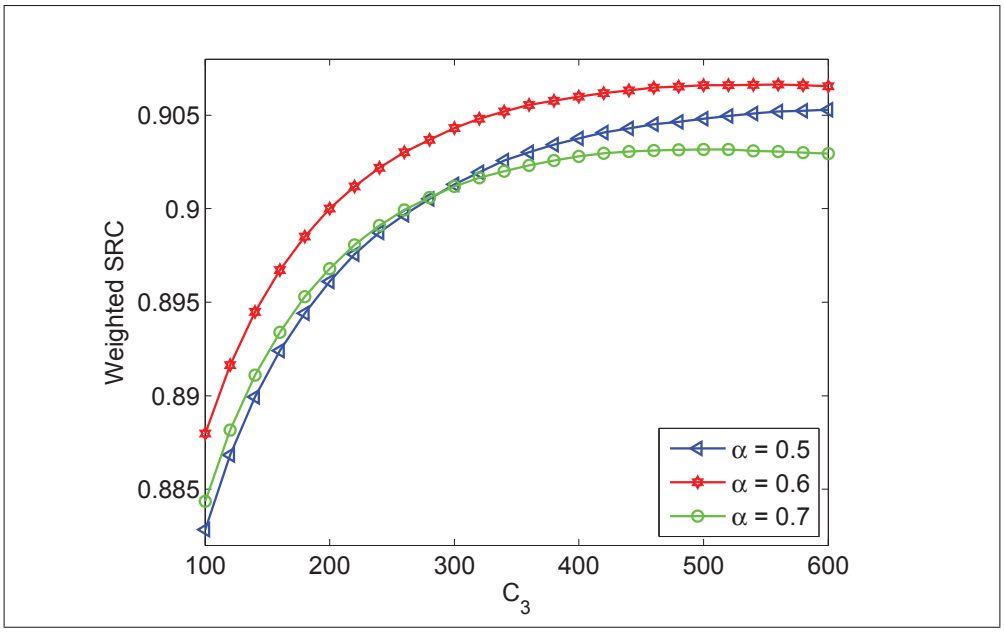

Figure 3.7 The weighted SRC performance of MDSI for different values of C_3 and α on eight datasets (TID2008, CSIQ, LIVE, TID2013, VCL@FER, CCID2014, ESPL, and DRIQ).

ues. Three pooling strategies were used in this experiment, e.g. mean pooling, mean absolute deviation (MAD) pooling and the standard deviation (SD) pooling. Figure 3.8 shows the SRC performance of the proposed index for different scenarios. From the plot in Figure 3.8, the following conclusions can be drawn. MAD pooling and both CS and $\widehat{\text{CS}}$ are good choices for MDSI. For almost every pooling strategy and parameter of C_3, the proposed $\widehat{\text{CS}}$ performs better than CS. This advantage is at the same time that the proposed $\widehat{\text{CS}}$ is more efficient than the existing CS.

3.3.8 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 (Ziaei Nafchi & Cheriet, 2016). The value of M is set to $[min(h,w)/256]$ (Wan), where h and w are image height and width, and $[.]$ is the round oper-

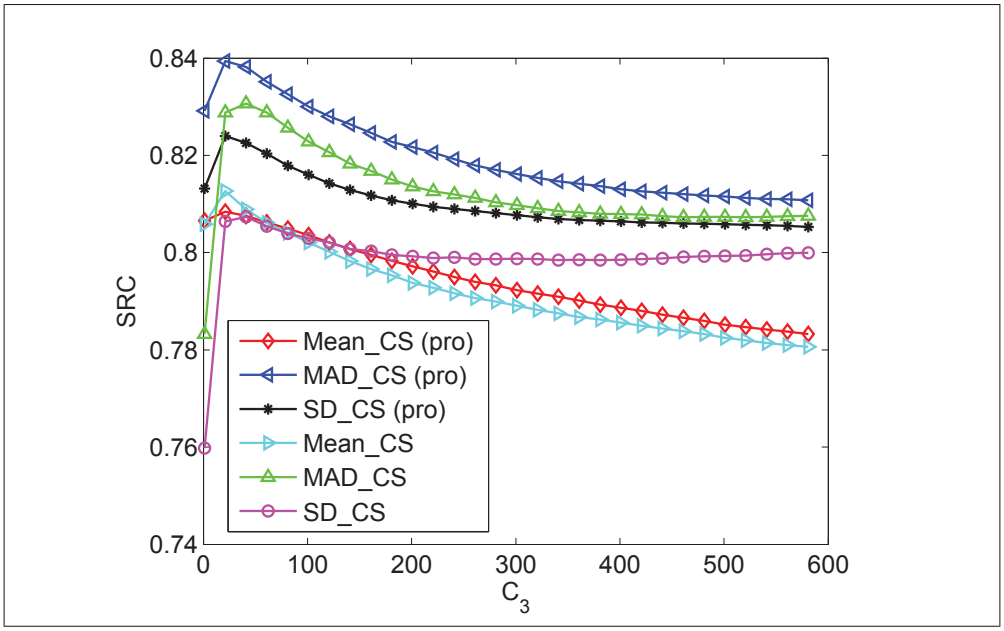

Figure 3.8 The SRC performance of the proposed index MDSI with two chromaticity similarity maps CS and $\widehat{\text{CS}}$ (proposed) for different values of C_3 and three pooling strategies on CCID2014 dataset (Gu *et al.*, 2015a).

ator. 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 3.6 lists the run times of fifteen IQA models when applied on images of size 384×512 and 1080×1920. The experiments were performed on a Core i7 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_c, the proposed index MDSI is about 3 to 6 times, 3 to 9 times, 4 to 5 times, 4 to 5 times, and 4 to 11 times faster, respectively. Another observation from the Table 3.6 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.

Table 3.6 Run time comparison of IQA models in terms of milliseconds

IQA model	384×512	1080×1920
PSNR	5.69	37.85
GMSD (Xue *et al.*, 2014b)	8.90	78.22
▷ MDSI	12.21	152.85
SSIM (Wang *et al.*, 2004)	14.97	80.23
SR_SIM (Zhang & Li, 2012)	17.02	100.06
MSSSIM (Wang *et al.*, 2003)	52.16	413.70
ADD-GSIM (Gu *et al.*, 2016a)	59.58	566.99
SFF (Chang *et al.*, 2013)	64.22	588.57
SCQI (Bae & Kim, 2016b)	71.68	524.01
VSI (Zhang *et al.*, 2014)	106.87	492.85
$FSIM_c$ (Zhang *et al.*, 2011)	145.02	590.84
IWSSIM (Wang & Li, 2011)	244.00	2538.43
DSCSI (Lee & Plataniotis, 2015)	423.73	4599.83
VIF (Sheikh & Bovik, 2006)	635.22	6348.67
MAD (Larson & Chandler, 2010)	847.54	8452.50

3.4 Conclusion

We proposed 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.

3.5 Acknowledgments

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

CHAPTER 4

FSITM: A FEATURE SIMILARITY INDEX FOR TONE-MAPPED IMAGES

Hossein Ziaei Nafchi[1], Atena Shahkolaei[1], Reza Farrahi Moghaddam[1], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Published in IEEE Signal Processing Letters
Volume 22, Issue 8, August 2015, Pages 1026-1029

Abstract

In this work, based on the local phase information of images, an objective index, called the feature similarity index for tone-mapped images (FSITM), is proposed. To evaluate a tone mapping operator (TMO), the proposed index compares the locally weighted mean phase angle map of an original high dynamic range (HDR) to that of its associated tone-mapped image calculated using the output of the TMO method. In experiments on two standard databases, it is shown that the proposed FSITM method outperforms the state-of-the-art index, the tone mapped quality index (TMQI). In addition, a higher performance is obtained by combining the FSITM and TMQI indices. The MATLAB source code of the proposed metric(s) is available at https://www.mathworks.com/matlabcentral/fileexchange/59814.

Keywords

High dynamic range, mean phase, objective quality assessment, tone-mapping operator.

4.1 Introduction

There is increasing interest in high dynamic range (HDR) images, HDR imaging systems, and HDR displays. The visual quality of high dynamic range images is vastly higher than that of conventional low-dynamic-range (LDR) images, and the significance of the move from LDR to HDR has been compared to the momentous move from black-and-white to color television (Reinhard *et al.*, 2010). In this transition period, and to guarantee compatibility in the future,

there has been a need to develop methodologies to convert an HDR image into its 'best' LDR equivalent. For this conversion, tone mapping operators (TMOs) have attracted considerable interest. Tone-mapping operators have been used to convert HDR images into their LDR associated images for visibility purposes on non-HDR displays.

Unfortunately, TMO methods perform differently, depending on the HDR image to be converted, which means that the best TMO method must be found for each individual case. A survey of various TMOs for HDR images and videos is provided in (Yeganeh & Wang, 2013a) and (Eilertsen *et al.*, 2013). Traditionally, TMO performance has been evaluated subjectively. In (Ledda *et al.*, 2005), a subjective assessment was carried out using an HDR monitor. Mantiuk et al. (Mantiuk *et al.*, 2005) propose an HDR visible difference predictor (HDR-VDP) to estimate the visibility differences of two HDR images, and this tool has also been extended to a dynamic range independent image quality assessment (Aydin *et al.*, 2008). However, the authors did not arrive at an objective score, but instead evaluated the performance of the assessment tool on HDR displays. Although subjective assessment provides true and useful references, it is an expensive and time-consuming process. In contrast, the objective quality assessment of tone mapping images enables an automatic selection and parameter tuning of TMOs (Yeganeh & Wang, 2010; Ma *et al.*, 2014). Consequently, objective assessment of tone-mapping images, which is proportional to the subjective assessment of the images, is currently of great interest.

Recently, an objective index, called the tone mapping quality index (TMQI) was proposed in (Yeganeh & Wang, 2013a) to objectively assess the quality of the individual LDR images produced by a TMO. The TMQI is based on combining an SSIM-motivated structural fidelity measure with a statistical naturalness:

$$\mathrm{TMQI}(H,L) = a[S(H,L)]^{\alpha} + (1-a)[N(L)]^{\beta}. \tag{4.1}$$

where S and N denote the structural fidelity and statistical naturalness, respectively. H and L denote the HDR and LDR images. The parameters α and β determine the sensitivities of the two factors, and a ($0 \leq a \leq 1$) adjusts their relative importance. Both S and N are upper bounded by 1, and so the TMQI is also upper bounded by 1 (Ma *et al.*, 2014). Although the TMQI clearly provides better assessment for tone-mapped images than the well-known image quality assessment metrics, like SSIM (Wang *et al.*, 2004), MS-SSIM (Wang *et al.*, 2003), and FSIM (Zhang *et al.*, 2011), its performance is not perfect. Liu et al. (Liu *et al.*, 2014b) replaced the pooling strategy of the structural fidelity map in the TMQI with various visual saliency-based strategies for better quality assessment of tone mapped images. They examined a number of visual saliency models and conclude that integrating saliency detection by combining simple priors (SDSP) into the TMQI provides better assessment capability than other saliency detection models.

In this paper, we first propose a feature similarity index for tone-mapped images (FSITM) which is based on the phase information of images. It has been observed that phase information of images prevails its magnitude (Oppenheim & Lim, 1981). Also, physiological evidence indicates that the human visual system responds strongly to points in an image where the phase information is highly ordered (Morrone & Burr, 1988). Based on this assumption, several quality assessment metrics have been proposed (Zhang *et al.*, 2011; Hassen *et al.*, 2013; Saha & Wu, 2013). In (Zhang *et al.*, 2011), the maximum moment of phase congruency covariance, which is an edge strength map, is used. Hassen et al. (Hassen *et al.*, 2013) used local phase coherence for image sharpness assessment. Saha et al. (Saha & Wu, 2013) proposed an image quality assessment using phase deviation sensitive energy features. Unfortunately, these metrics do not provide a reliable assessment for tone mapped images.

The FSITM images proposed in this paper uses the phase-derived feature type of the images in a different way from that proposed in (Zhang *et al.*, 2011; Hassen *et al.*, 2013; Saha & Wu, 2013). Our FSITM uses a locally weighted mean phase angle (LWMPA) (Kovesi, 1999b), which is a feature map based on the local-phase. This phase-derived map is noise independent,

and therefore there is no parameter to set for noise estimation. The proposed FSITM assesses both the appearance of the real world scene and the most pleasing image for human vision.

Given the FSITM and the TMQI, we also proposed a combined metric, FSITM_TMQI, which provides much better assessment of tone-mapped images. In the experiments, we compare the objective scores of our proposed similarity indices (FSITM, FSITM_TMQI), along with TMQI (Yeganeh & Wang, 2013a), on two major datasets (Yeganeh & Wang, 2013b; Čadík, 2008b).

4.2 The proposed similarity index

The proposed FSITM similarity index for tone-mapped images is based on a phase-derived feature map. As we mentioned before, phase-derived features have already been used successfully for quality assessment (Zhang *et al.*, 2011; Hassen *et al.*, 2013; Saha & Wu, 2013). However, their results for evaluating tone-mapped images is not reliable similar to other popular quality assessment metrics like the SSIM and its variations (Wang *et al.*, 2004, 2003). For this reason, we use the locally weighted mean phase angle (LWMPA) map in this paper, because it is a feature that marks locally dark/bright pixels, it is a rough indicator of the edges, and it is based on the directions that should remain unchanged in a tone-mapped image. Below, we briefly describe the theory and formulation of the LWMPA, and then discuss our proposed similarity index which is based on this feature map.

Let $M^e_{\rho r}$ and $M^o_{\rho r}$, which are known in the literature as quadratic pairs, denote the even symmetric and odd symmetric log-Gabor wavelets at a scale ρ and orientation r (Papari & Petkov, 2011). By considering $f(\mathbf{x})$ as a two-dimensional signal on the two-dimensional domain of $\mathbf{x}$, the response of each quadratic pair of filters at each image point $\mathbf{x}$ forms a response vector by convolving with $f(\mathbf{x})$:

$$\left[e_{\rho r}(\mathbf{x}), o_{\rho r}(\mathbf{x})\right] = \left[f(\mathbf{x}) * M^e_{\rho r}, f(\mathbf{x}) * M^o_{\rho r}\right] . \tag{4.2}$$

where the values $e_{\rho r}(\mathbf{x})$ and $o_{\rho r}(\mathbf{x})$ are real and imaginary parts of a complex-valued wavelet response at a scale ρ and an orientation r. We can now compute the local phase $\phi_{\rho r}(\mathbf{x})$ of the transform at a given wavelet scale ρ and orientation r:

$$\phi_{\rho r}(\mathbf{x}) = \text{arctan2}\Big(e_{\rho r}(\mathbf{x}), o_{\rho r}(\mathbf{x})\Big) \,, \tag{4.3}$$

where arctan2(x,y) $= 2\text{arctan}\frac{x}{\sqrt{x^2+y^2}+y}$. The locally-weighted mean phase angle ph($\mathbf{x}$) is obtained using the summation of all filter responses over all the possible orientations and scales:

$$\text{ph}(\mathbf{x}) = \text{arctan2}\Big[\sum_{\rho,r} e_{\rho r}(\mathbf{x}), \sum_{\rho,r} o_{\rho r}(\mathbf{x})\Big] \,. \tag{4.4}$$

The pixels of ph($\mathbf{x}$) take values between $-\pi/2$ (a dark line), $+\pi/2$ (a bright line), and 0 for steps. This classification of step and line features has been further studied in (Kovesi, 2002).

There are a few parameters to be considered in the calculation of ph($\mathbf{x}$). In our set of experiments, we determine the best fixed values for this operation (see section 4.3). Unlike the phase-derived edge map and local phase that are used in other research (Zhang *et al.*, 2011; Hassen *et al.*, 2013), the locally weighted mean phase angle ph($\mathbf{x}$) provides a good representation of image features, including the edges and shapes of objects. Since ph($\mathbf{x}$) indicates both dark and bright lines, it can be used to assess color changes, which is a popular feature of the TMOs. Moreover, the LWMPA is noise-independent, unlike the phase derived features used in (Zhang *et al.*, 2011; Hassen *et al.*, 2013; Saha & Wu, 2013), which are sensitive to noise, and therefore require an estimation of the noise. Some examples of ph($\mathbf{x}$) outputs are shown in Figure 4.1.

We use only ph($\mathbf{x}$) to calculate the FSITM. First, the HDR (H) image is converted into its LDR (L) by simply taking logarithm of its values (LoGH $= log(H)$). This rough LDR image is used as one of the reference images for computing the FSITM. Another reference image is the HDR image itself. The details of the FSITM calculations are provided below.

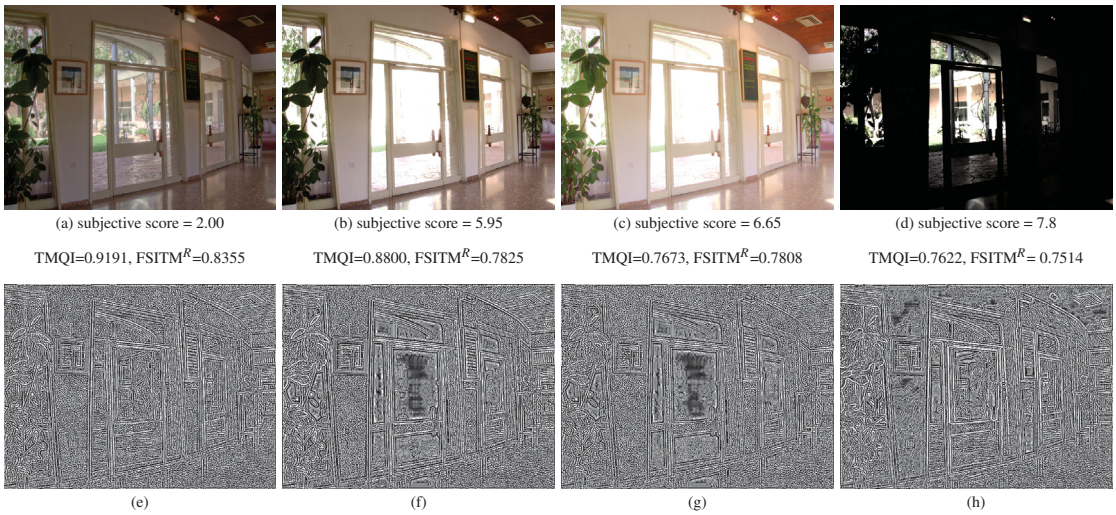

Figure 4.1 (a)-(d) LDR images using different TMOs (Yeganeh & Wang, 2013a), along with their corresponding TMQI and FSITM scores for each. (e)-(h) The associated LWMPA maps of their red channel.

Given the input images H and L, and LogH=$log(H)$ image, the ph$(\mathbf{x})$ for each channel C of these three images is calculated using equation (4.4). The FSITM is based on the simple fact that the features in the two corresponding channels should have remained the same in their ph$(\mathbf{x})$ maps. The FSITM is equal to 1 if all the feature types are the same, and 0 if they are all different. First, we define the feature similarity index for a channel C used in calculation of the FSITM:

$$F^C(L,H) = |P_L^C(\mathbf{x}) \wedge P_H^C(\mathbf{x})| \;/\; (row \times col), \tag{4.5}$$

where $P(\mathbf{x})$ denotes a binary image of ph$(\mathbf{x})$:

$$P(\mathbf{x}) = U\big(\text{ph}(\mathbf{x})\big), \tag{4.6}$$

where $U(\cdot)$ is the unit-step function. For the case of tone-mapped images, the FSITM for a channel C is defined as:

$$\text{FSITM}^C = \alpha F^C(H,L) + (1-\alpha)F^C(\text{LogH},L). \tag{4.7}$$

where α ($0 \leq \alpha \leq 1$), controls the impact factor of H and LogH in the calculation of the FSITM. Algorithm 4.1 lists all the steps in the process of calculating our proposed FSITM.

Algorithm 4.1 The feature similarity index for tone-mapped images (FSITM).

```
1    procedure FSITM(H, L, C) start
2        H: HDR, L: LDR, C∈{R, G, B}
3        LogH = log(H);
4        Calculate ph(x) for C channel of images L, H and LogH.
5        FSITM^C = αF^C(H,L) + (1−α)F^C(LogH,L);
6    return FSITM^C
7    end procedure
```

We also found that combining the FSITM and the TMQI provides a better assessment of tone-mapped images. Therefore, we proposed a combined index of the FSITM and the TMQI based on the following equation:

$$\text{FSITM}^C\text{\_TMQI} = (\text{FSITM}^C + \text{TMQI}) \ / \ 2 \tag{4.8}$$

In most of the cases, the different properties of these two indices cause them to moderate similarity estimation mistakes of each other.

4.3 Experimental results

To evaluate the proposed FSITM index, we used the dataset A introduced in (Yeganeh & Wang, 2013a) and (Čadík, 2008b). The first dataset contains 15 HDR images, along with 8 LDR images for each HDR image. The HDR images were produced using different TMOs. The quality of LDRs is ranked from 1 (best quality) to 8 (worst quality). The ranks were obtained based on a subjective assessment of 20 individuals. The second HDR dataset (dataset B) used

Table 4.1 Performance comparison of the proposed quality indices and TMQI (Yeganeh & Wang, 2013a,b) on the dataset A introduced in (Yeganeh & Wang, 2013a,b).

SRCC							
Index	TMQI	FSITM^{R}	FSITM^{R}_TMQI	FSITM^{G}	FSITM^{G}_TMQI	FSITM^{B}	FSITM^{B}_TMQI
Min	0.6826	0.6190	**0.7143**	0.5476	**0.7143**	0.1796	0.5509
Median	0.7857	0.8095	**0.8571**	0.8333	**0.8571**	**0.8571**	**0.8571**
Average	0.8058	0.8145	**0.8559**	0.8178	0.8424	0.7183	0.8097
STD	0.1051	0.1214	**0.0863**	0.1310	**0.0886**	0.2536	0.1229
KRCC							
Min	0.5455	0.5000	**0.5714**	0.3571	**0.5714**	0.2143	0.4001
Median	0.6429	0.7143	0.7143	0.7143	**0.7857**	0.7143	0.7143
Average	0.6840	0.7126	**0.7508**	0.6935	0.7317	0.5979	0.6838
STD	0.1221	0.1423	**0.1083**	0.1711	**0.1078**	0.2711	0.1436

Table 4.2 Performance comparison of the proposed quality indices and TMQI (Yeganeh & Wang, 2013a,b) on the dataset B introduced in (Čadík, 2008b).

SRCC							
Index	TMQI	FSITM^{R}	FSITM^{R}_TMQI	FSITM^{G}	FSITM^{G}_TMQI	FSITM^{B}	FSITM^{B}_TMQI
Min	0.7198	0.7363	0.8901	0.7692	**0.9231**	0.7637	0.8462
Average	0.7985	0.7692	0.9102	0.8461	**0.9267**	0.8241	0.8901
KRCC							
Min	0.5385	0.5897	0.6923	0.6154	**0.7692**	0.5385	0.6410
Average	0.6410	0.6410	0.7692	0.7265	**0.8119**	0.6410	0.7264

is also available along with subjective ranks for LDR images (Čadík, 2008b). That dataset contains three HDR images, and 14 LDR images for each HDR image.

To objectively evaluate the performance of the various similarity indices considered in our experiments, we use the Spearman rank-order correlation coefficient (SRCC) and the Kendall rank-order correlation coefficient (KRCC) metrics.

The proposed similarity indices (FSITM^{C}, FSITM^{C}_TMQI) are compared with the TMQI (Yeganeh & Wang, 2013a). The results are listed in Tables 4.1 and 4.2. The performance of the TMQI is listed based on the scores obtained by running the Matlab source code provided by Yeganeh and Wang in (Yeganeh & Wang, 2013b). The FSITM^{G} outperforms the TMQI in terms of SRCC and KRCC for both datasets. In general, there is less variation in TMQI performance than in FSITM performance. In contrast, the FSITM^{R}_TMQI and FSITM^{G}_TMQI are more robust, and also they outperform the FSITM and TMQI in terms of both the SRCC and KRCC scores.

It is worth to report the available results of other indices which have recently been proposed in the literature (Liu *et al.*, 2014b; Gu *et al.*, 2014b). In (Liu *et al.*, 2014b), the authors reported the SRCC performance of their proposed index for the dataset A (Yeganeh & Wang, 2013b). Their minimum and average SRCC performance is 0.6905 and 0.8408, respectively. Their standard deviation of SRCC scores is reported as 0.0907. For the same dataset, the median performance of the ref. (Gu *et al.*, 2014b) is reported as follows: SRCC=0.8106 and KRCC=0.5865.

A number of parameters impact the quality of the locally weighted mean phase angle $\mathrm{ph}(\mathbf{x})$, namely the number of filter scales N_ρ, the wavelength of the smallest scale filter wLen, and the scaling factor between successive filters mult. In the experiments, these parameters were set to $N_\rho = 2$, $wLen = 2$, and $mult = 2$ for the LogH image, while they were set to $N_\rho = 2$, $wLen = 8$, and $mult = 8$ for the original HDR image. The rational for using two different set of parameters is that the size of the image features could be different. Overall, it is the three parameters of $\mathrm{ph}(\mathbf{x})$ along with the value of α that influence the performance of the proposed indices.

In this work, we only used the original HDR image and its logarithm image LogH. It is worth mentioning that we have tried the same strategy used in defining FSITM in RGB color space in other color spaces, such as Lab and Yxy color spaces. However, we did not get a good performance.

We evaluated the run time of the FSITM and the TMQI as follows: our experiments were performed on a Core i7 3.4 GHz CPU with 16 GB of RAM. The FSITM algorithm was implemented in MATLAB 2012b running on Windows 7. The TMQI and the FSITM took 1.95 and 3.36 seconds respectively to assess images of size 1200×1600, while the run time for the FSITM^C_TMQI is simply obtained by adding the TMQI and FSITM^C run-times.

4.4 Conclusion

We have proposed an objective index, called the feature similarity index for tone-mapped images (FSITM), which is based on the local phase similarity of the original HDR and the target

converted LDR image. Unlike other studies in which different phase-derived feature maps are used, we have used the locally weighted mean phase angle, which is a robust and noise-independent feature map. The performance of the proposed similarity index is compared with the state-of-the-art TMQI on two datasets, and has been found to be promising. The proposed FSITM and the TMQI have been then combined to obtain a more accurate quality assessment. Further studies are required to develop more comprehensive HDR datasets, along with their subjective scores. Such datasets would allow us to develop better performing indices.

4.5 Acknowledgments

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

CHAPTER 5

MUG: A PARAMETERLESS NO-REFERENCE JPEG QUALITY EVALUATOR ROBUST TO BLOCK SIZE AND MISALIGNMENT

Hossein Ziaei Nafchi[1], Atena Shahkolaei[1], Rachid Hedjam[2], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Geography, McGill University,
805 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

Published in IEEE Signal Processing Letters
Volume 23, Issue 11, November 2016, Pages 1577-1581

Abstract

In this letter, a very simple no-reference image quality assessment (NR-IQA) model for JPEG compressed images is proposed. The proposed metric called median of unique gradients (MUG) is based on the very simple facts of unique gradient magnitudes of JPEG compressed images. MUG is a parameterless metric and does not need training. Unlike other NR-IQAs, MUG is independent to block size and cropping. A more stable index called MUG$^+$ is also introduced. The experimental results on six benchmark datasets of natural images and a benchmark dataset of synthetic images show that MUG is comparable to the state-of-the-art indices in literature. In addition, its performance remains unchanged for the case of the cropped images in which block boundaries are not known. The MATLAB source code of the proposed metrics is available at https://www.mathworks.com/matlabcentral/fileexchange/59810 and https://www.mathworks.com/matlabcentral/fileexchange/59813.

Keywords

JPEG compression, Blockiness artifact, JPEG quality assessment, No-reference quality assessment, MUG.

5.1 Introduction

JPEG lossy compression is one of the most common coding techniques to store images. It uses a block based coding scheme in frequency domain, e.g. discrete cosine transform (DCT), for compression. Since $B \times B$ (8×8) blocks are coded independent of each other, blocking artifacts are visible in JPEG compressed images specially under low bit rate compression. Several no-reference image quality assessment models (NR-IQAs) have been proposed to objectively assess the quality of the JPEG compressed images (Wu & Yuen, 1997; Tan & Ghanbari, 2000a,b; Wang *et al.*, 2000; Bovik & Liu, 2001; Wang *et al.*, 2002; Pan *et al.*, 2004; Perra *et al.*, 2005; Park *et al.*, 2007; Zhai *et al.*, 2008; Liu & Heynderickx, 2009; Chen & Bloom, 2010; Lee & Park, 2012; Golestaneh & Chandler, 2014; Li *et al.*, 2014a,b, 2015). NR-IQAs do not need any information of the reference image. NR-IQAs are of high interest because in most present and emerging practical real-world applications, the reference signals are not available (Wang & Bovik, 2011). In the following, we will have an overview on NR-IQAs for JPEG compressed images.

In (Wu & Yuen, 1997) for each block, horizontal and vertical difference at block boundaries are used to measure horizontal and vertical blockiness, respectively. The authors in (Tan & Ghanbari, 2000a) proposed a blockiness metric via analysis of harmonics. They used both the amplitude and the phase information of harmonics to compute a quality score. Harmonic analysis was also used to model another blockiness metric in (Tan & Ghanbari, 2000b).

Wang et. al. (Wang *et al.*, 2000) modeled the blocky image as a non-blocky image interfered with a pure blocky signal. Energy of the blocky signal is then used to calculate a quality score. In DCT domain, a metric was proposed in (Bovik & Liu, 2001) that models the blocking artifacts by a 2-D step function. The quality score is calculated following the human vision measurement of block impairments. The metric proposed in (Park *et al.*, 2007) measures blockiness artifact in both the pixel and the DCT domains. In (Golestaneh & Chandler, 2014), zero values DCT coefficients within each block are counted and a relevance map is estimated that

distinguishes between naturally uniform blocks and compressed uniform blocks. For this end, an analysis in both DFT and DCT domains is conducted.

Wang et. al. (Wang *et al.*, 2002) proposed an efficient metric that measures blockiness via horizontally and vertically computed features. These features are average differences across block boundaries, average absolute difference between in-block image samples, and zero crossing rate. Using a set of subjective scores, five parameters of this model are estimated via nonlinear regression analysis. In (Pan *et al.*, 2004), the edge orientation changes of blocks were used to measure severity of blockiness artifacts. Perra et. al. (Perra *et al.*, 2005) analyzed the horizontal, vertical and intra-block sets of 8×8 blocks after applying the Sobel operator to the JPEG compressed images.

The difference of block boundaries plus luminance adaptation and texture masking were used in (Zhai *et al.*, 2008) to form a noticeable blockiness map (NBM). From which, the quality score is calculated by a Minkowski summation pooling. In (Liu & Heynderickx, 2009), 1-D signal profile of gradient image is used to extract block sizes and then priodic peaks in DCT domain are analyzed to calculate a quality score. Chen et. al. (Chen & Bloom, 2010) proposed a very similar metric.

In (Li *et al.*, 2014a), three features including the corners, block boundaries (horizontal, vertical and intra-block), and color changes, together with the subjective scores are used to train a support vector regression (SVR) model. Li et. al. (Li *et al.*, 2014b) measured the blocking artifacts through weighting a set of blockiness scores calculated by Tchebichef moments of different orders.

Lee and Park (Lee & Park, 2012) proposed a blockiness metric that first identifies candidates of having blockiness artifacts. The degree of blockiness of these candidates is then used to compute a quality score. Recently a blockiness metric is proposed that performs in three steps (Li *et al.*, 2015). Block grids are extracted in the spatial domain and their strength and regularity is measured. Afterwards, a masking function is used that gives different weights to the smooth and textured regions.

The aforementioned indices have at least one of the following drawbacks. They might not be robust to block size and block misalignment (examples are (Wang *et al.*, 2002; Pan *et al.*, 2004; Perra *et al.*, 2005; Golestaneh & Chandler, 2014; Li *et al.*, 2014b; Zhai *et al.*, 2008)). They are complex (examples are (Bovik & Liu, 2001; Golestaneh & Chandler, 2014; Li *et al.*, 2015, 2014b,a; Liu & Heynderickx, 2009)), or have many parameters to set ((Wang *et al.*, 2002; Golestaneh & Chandler, 2014; Li *et al.*, 2015; Liu & Heynderickx, 2009; Li *et al.*, 2014b,a)). Indices like NJQA (Golestaneh & Chandler, 2014) and GridSAR (Li *et al.*, 2015) are too much slow. Some indices need training (Wang *et al.*, 2002; Li *et al.*, 2014a). Also, the range of quality scores provided by some of the indices like (Wang *et al.*, 2002) is not well defined, or they show other numerical issues (Li *et al.*, 2015).

In this letter, we propose a quality assessment model for JPEG compressed images that overcomes all aforementioned drawbacks. The proposed index is very simple and efficient, it is parameterless, and robust to block size and misalignment. The proposed metric called MUG is based on two simple facts about blockiness artifact. As a result of more JPEG compression, the number of unique gradient magnitude values decreases, and the median value of unique gradient magnitude values increases. The proposed blockiness metric MUG uses these two simple facts to provide accurate quality predictions for JPEG compressed images. Unlike other metrics that presume position of blocks beforehand or localize the position of blocks, MUG is not a local model and hence does not need any information on the position of blocks.

5.2 Proposed Metric (MUG)

The proposed index called MUG predicts the quality of JPEG compressed images as follows. Given the JPEG distorted image D, the Scharr gradient operator is used to approximate horizontal G_x and vertical G_y gradients of D: $G_x = h_x * D$ and $G_y = h_y * D$, where h_x and h_y are horizontal and vertical gradient operators, and $*$ denotes the convolution. From which, the gradient magnitude is computed as $G(\mathbf{x}) = \sqrt{G_x^2(\mathbf{x}) + G_y^2(\mathbf{x})}$. It is worth to mention that within the context of the proposed metrics, the Scharr operator performs better than the Sobel and Prewitt operators. The proposed metric works directly on the gradient magnitude instead of

directional gradients. Let's denote uG as the unique numerical values of $G(\mathbf{x})$. We show in the following that two properties of uG can be used to predict quality of JPEG compressed images: i) number of values in uG, and ii) median of uG values.

5.2.1 Number of unique gradients (NUG)

The number of unique gradients (NUG), e.g. the number of values in vector uG, indicates how many distinct edge strengths exist in JPEG compressed image D. It is very likely that a JPEG compressed image with blocking artifacts has smaller values of NUG than its uncompressed version. To verify this statement, JPEG compressed images of TID2013 dataset were chosen. For each of the 25 distortion-free images in TID2013, there are five JPEG compressed images of different distortion levels. The values of NUG for each of the 25 sets are found inversely proportional to the amount of distortion:

$$\text{Compression rate} \propto \frac{1}{\text{NUG}} \tag{5.1}$$

In other words, the Spearman Rank-order Correlation coefficient (SRCC) between NUG values and mean opinion score (MOS) values is equal to 1 for each of the 25 sets. This experiment shows that aforementioned statement holds true. Fig. 5.1 shows scatter plot of NUG scores against the subjective MOS on the LIVE dataset (Sheikh *et al.*, 2006) (see experimental results section to see how this plot is drawn). This plot shows that there is noticeable correlation between NUG scores and MOS on this dataset. Unfortunately, NUG does not take into account the content of original images. An image may originally have less edge strengths variation than another. Therefore, there are cases that NUG can not fairly judge images having different contents. This issue is solved through including median of unique gradients (MUG) into the proposed model.

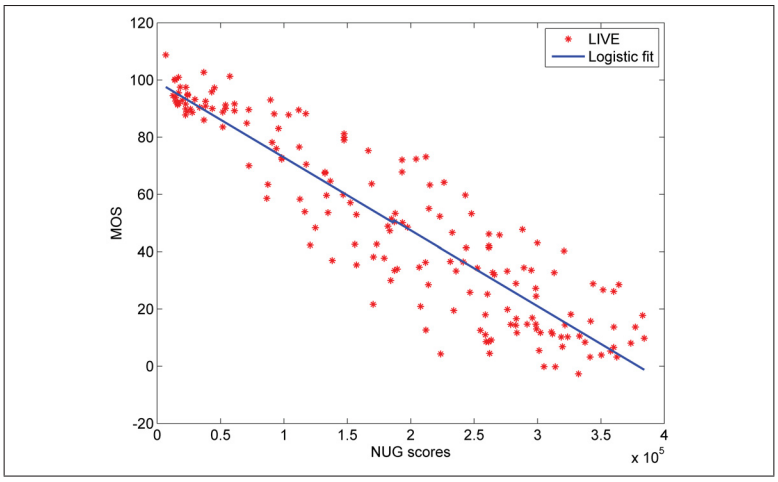

Figure 5.1 Scatter plot of NUG scores against the subjective MOS on the LIVE dataset. The Pearson linear Correlation Coefficient (PLCC) is equal to 0.9105.

5.2.2 Median of unique gradients (MUG)

As mentioned above, the image content is a factor that needs to be taken into account. Let's repeat the same experiment on JPEG compressed images of the TID2013 dataset, but this time for the median of unique gradients (MUG). The experiments show that the same statement holds true, e.g. the values of MUG for each of the 25 sets are proportional[1] to the amount of distortion:

$$\text{Compression rate} \propto MUG \tag{5.2}$$

In fact, MUG determines how strong is the middle value of unique gradients which helps in taking into account the content of images. However, the values of MUG are not always reliable because image quality is not only related to the edge strengths. The distribution of the unique gradients uG is another factor that can not be considered by direct median value. Therefore, a simple standard deviation normalization was applied on the uG values before median value being computed:

[1] Except for one case where SRCC is equal to 0.6, not 1.

$$uG' = \frac{uG}{\sqrt{\sigma(uG)}} \tag{5.3}$$

Unique gradients vector uG has different behavior for images having naturally uniform regions and block uniform regions. For images with mostly naturally uniform regions, the standard deviation in general decreases by more compression. In contrast, standard deviation value in general increases by more compression for images having less naturally uniform regions. Therefore, median of uG' takes into account the *content* of images. The effect of standard deviation normalization is visually shown in the scatter plots of Figure 5.2.

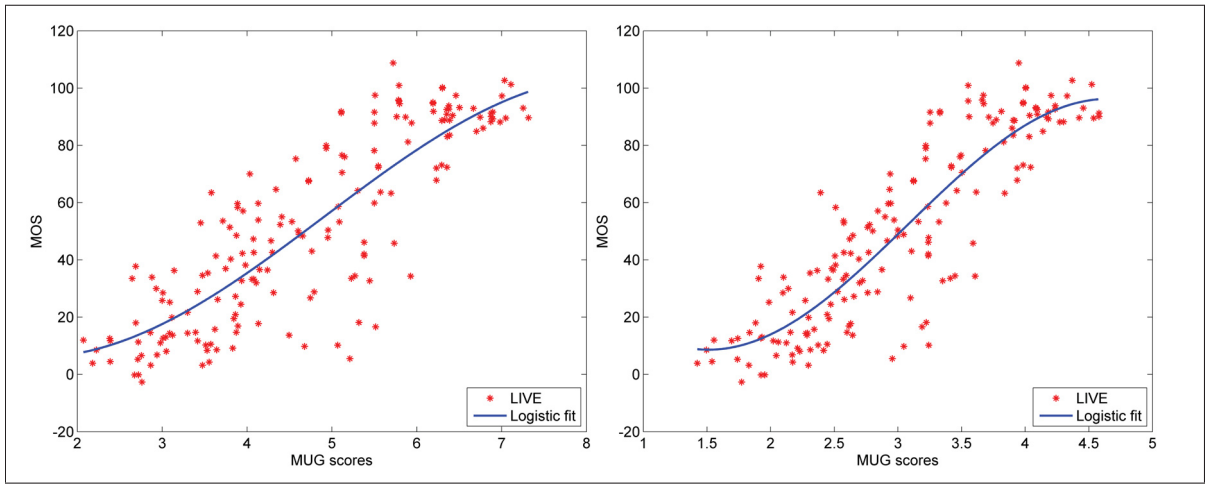

Figure 5.2 Scatter plots of MUG scores against the subjective MOS on the LIVE dataset. Left: MUG without normalization (PLCC = 0.8422), and right: MUG with standard deviation normalization (PLCC = 0.8768).

The proposed quality assessment model for JPEG compressed images (MUG) can be written by combining relations (5.1) and (5.2):

$$\text{MUG} = \frac{MUG}{\text{NUG}} \tag{5.4}$$

where, MUG (in italic) is the median value of uG'. It can be seen that the proposed metric is parameterless. To the best of our knowledge, MUG is the only parameterless metric in the

literature. MUG is therefore completely independent to the misalignment. This advantage is shown in the experimental results. Since the proposed metric is parameterless, it should be invariant to the block size as well. However, no dataset is available to experimentally verify this statement. It is worth to mention that when the input image is in color, MUG converts it to a luminance channel: L = 0.06R + 0.63G + 0.27B. According to (Geusebroek *et al.*, 2001), this conversion may be imperfect, but it is likely to offer accurate estimates of differential measurements. Therefore, image gradient computation from L should yield more accurate results. Since MUG only uses the median value of unique gradients, it might not be very accurate for images with different edge distributions. In the following, MUG is modified by adding a few more unique gradient values.

5.2.3 Stable MUG (MUG$^+$)

The distributions of unique gradient values can be very different for images having diverse edge information. This distribution might be skewed (usually right-skewed), bimodal, etc. Median value alone might not be sufficient for images with different edge distribution. Therefore, the MUG index can become more stable by considering a few more values in addition to the value of the median. These values must be smaller than the median value because larger values than median have much more variations and might be unreliable. Suppose that uG' values are sorted from smallest to largest. In this case NUG/2 is the index of median value in uG'. One easy way to add a few values as mentioned above is to use corresponding values of these indices: NUG$/i$, $i \in \{2,3,...,M+1\}$, where $i = 2$ is the index of median and M is the total number of values used ($M = 19$ in this paper). In fact by adding these extra values, the proposed metric becomes numerically more stable. Moreover, there are cases that there are not M unique values in the vector uG'. This property often happens when the majority of the input image or the whole image is naturally uniform or textured. Suppose that there are $1 \leq N \leq M$ of such values available. The stable MUG (called MUG$^+$) takes into account this behavior by the following formulation:

$$\mathrm{MUG}^{+} = \frac{\mathrm{MUG}}{M-N+1} \tag{5.5}$$

where $\mathrm{MUG}^{+} = \mathrm{MUG}$ for $N = M$.

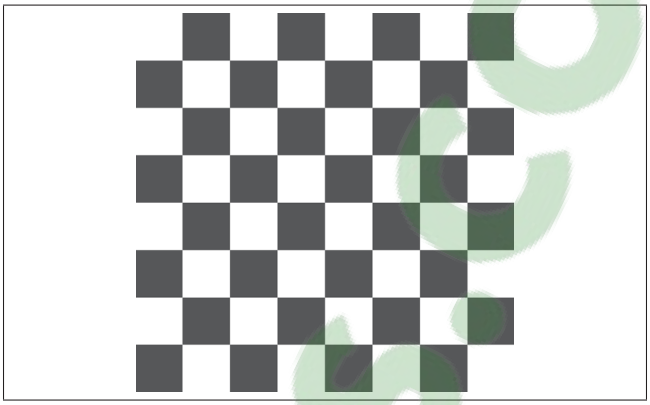

Figure 5.3 A high quality image of chessboard with naturally uniform and textured regions. The image size is 1024×1024 and block sizes are all 128×128.

Apart from the block misalignment problem, several JPEG quality assessment models like (Wang *et al.*, 2002; Golestaneh & Chandler, 2014) provide quite wrong predictions in special cases that image has large amount of naturally uniform regions and/or it is textured. Figure 5.3 shows a high quality image of chessboard. This image has a very bad quality according to the (Wang *et al.*, 2002) (Q = -245.89). NJQA (Golestaneh & Chandler, 2014) likewise assessed this image as being of bad quality (Q = 0.3414). GridSAR (Li *et al.*, 2015) was not able to provide a numerical value. MUG is equal to 0.8060 (very bad quality) which also provides wrong assessment. In contrast, MUG^{+} = 0.0448 which truly means that chessboard image has a very good quality. This is another advantage of the proposed index MUG^{+}. Note that the datasets used in this paper do not have any image sample with this behavior.

5.3 Experimental results

In the experiments, six standard datasets of natural images and a benchmark dataset of synthetic images are used. The TID2013 (Ponomarenko *et al.*, 2013) dataset contains 125 JPEG compressed images in total. The CSIQ dataset (Larson & Chandler, 2010) has 150, LIVE dataset (Sheikh *et al.*, 2006) has 175, VCL dataset (Zaric *et al.*, 2012) has 138, and the MICT dataset (Horita *et al.*) has 84 JPEG compressed images. ESPL dataset (Kundu & Evans, 2015) is a synthetic dataset which contains 100 JPEG compressed images. The TID2008 dataset (Ponomarenko *et al.*, 2009) is another dataset with 100 JPEG compressed images which is in fact a subset of TID2013.

For objective evaluation, two evaluation metrics were used in the experiments: the Spearman Rank-order Correlation coefficient (SRCC), and the Pearson linear Correlation Coefficient (PLCC). The SRCC and PLCC metrics measure prediction monotonicity and prediction linearity, respectively.

To get a visual observation, the scatter plots of the proposed NR-IQA models MUG and MUG$^+$ on the LIVE dataset are shown in Figure 5.4. The logistic function suggested in (Sheikh *et al.*, 2006) was used to fit a curve on each plot:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5 \tag{5.6}$$

where β_1, β_2, β_3, β_4 and β_5 are fitting parameters computed by minimizing the mean square error between quality predictions x and subjective scores MOS.

SSIM (Wang *et al.*, 2004) as an FR-IQA, as well as five NR-IQAs including (Wang *et al.*, 2002), NJQA (Golestaneh & Chandler, 2014), GridSAR (Li *et al.*, 2015), and the proposed indices MUG and MUG$^+$ were used in the experiments. (Wang *et al.*, 2002) was chosen because it shows outstanding performance, and NJQA because it follows a different approach with promising performance. GridSAR is recently introduced blockiness metric which is also

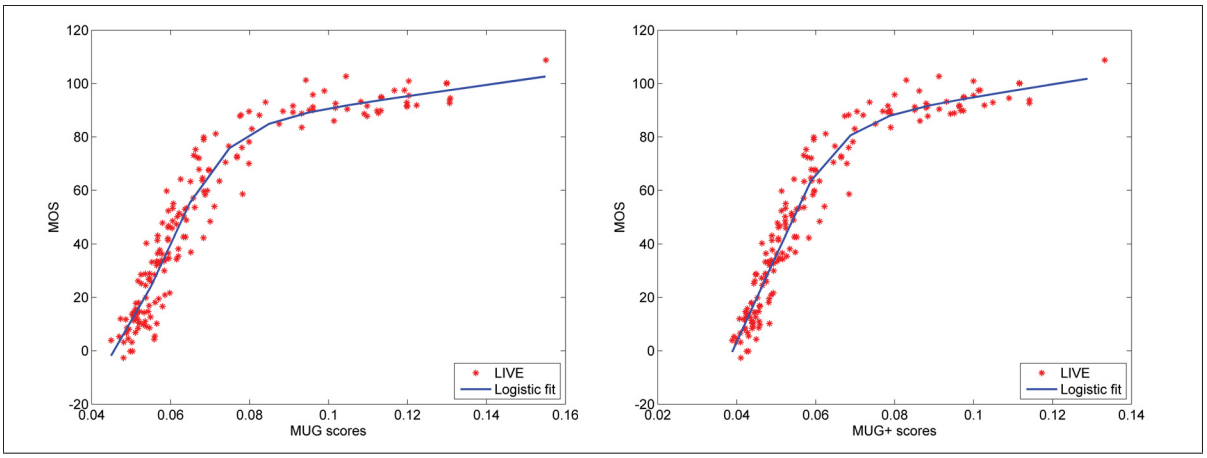

Figure 5.4 Scatter plots of MUG and MUG$^+$ scores against the subjective MOS on the LIVE dataset. Left: MUG (PLCC = 0.9649), and right: MUG$^+$ (PLCC = 0.9730).

Table 5.1 Performance comparison of the IQA models on JPEG compression distortion type of seven datasets in terms of SRCC and PLCC

Index		SSIM	(Wang *et al.*, 2002)	NJQA	(Li *et al.*, 2015)	MUG	MUG$^+$
Type		FR	NR	NR	NR	NR	NR
TID	PLCC	**0.9540**	0.9518	0.9442	0.9511	0.9408	**0.9529**
2008	SRCC	**0.9252**	0.9129	0.8993	0.9166	0.9169	**0.9239**
TID	PLCC	0.9544	0.9530	0.9477	**0.9545**	0.9419	**0.9546**
2013	SRCC	0.9200	**0.9267**	0.8860	**0.9309**	0.9077	0.9185
CSIQ	PLCC	**0.9786**	0.9751	0.9539	**0.9788**	0.9674	0.9717
	SRCC	0.9546	**0.9551**	0.9249	**0.9565**	0.9304	0.9372
LIVE	PLCC	**0.9790**	**0.9787**	0.9562	0.9756	0.9649	0.9730
	SRCC	**0.9764**	**0.9735**	0.9562	0.9726	0.9596	0.9677
VCL	PLCC	0.9257	**0.9433**	0.8611	**0.9304**	0.8683	0.8868
	SRCC	0.9236	**0.9403**	0.8445	**0.9313**	0.8659	0.8850
MICT	PLCC	0.8664	**0.8876**	**0.8746**	0.8305	0.8341	0.8503
	SRCC	0.8590	**0.8829**	**0.8728**	0.8333	0.8263	0.8513
ESPL	PLCC	0.9431	**0.9599**	0.8089	**0.9623**	0.9398	0.9370
	SRCC	0.9042	**0.9327**	0.7388	**0.9331**	0.9284	0.9265

able to handle block misalignment. Table 5.1 provides a performance comparison between the six aforementioned FR/NR-IQAs in terms of SRCC and PLCC. The same experiment is repeated on JPEG compressed images with misaligned blocks. JPEG compressed images with misaligned blocks are generated by cropping one pixel from the borders (four sides) of the images. Since only one pixel width is cropped from image borders, the MOS values should

remain unchanged. When block positions are known beforehand, the NR-IQA of (Wang *et al.*, 2002) shows the best overall performance for the seven datasets. The proposed indices show consistent prediction accuracy over different datasets and comparable to the GridSAR and SSIM. The proposed indices in general outperform NJQA (Golestaneh & Chandler, 2014). When block positions are not known, it can be seen from the Table 5.2 that the proposed indices, e.g. MUG and MUG$^+$, and GridSAR show almost the same prediction accuracy as in Table 5.1. This means that they are robust to the block misalignment. In contrast, (Wang *et al.*, 2002) provides predictions with low accuracy.

While GridSAR performs better than MUG$^+$ on more considered datasets, it should be noted that GridSAR is a complex metric with several parameters to set. It is also computationally inefficient and numerically unstable.

Table 5.2 Performance comparison of the IQA models on JPEG compression distortion type on seven datasets with block misalignment in terms of SRCC and PLCC

Index		SSIM	(Wang *et al.*, 2002)	NJQA	(Li *et al.*, 2015)	MUG	MUG$^+$
Type		FR	NR	NR	NR	NR	NR
TID	PLCC	0.9247	0.3742	0.8499	**0.9540**	0.9407	**0.9528**
2008	SRCC	0.8989	0.3146	0.8128	**0.9197**	0.9171	**0.9242**
TID	PLCC	0.9328	0.5087	0.8540	**0.9566**	0.9418	**0.9545**
2013	SRCC	0.9096	0.2372	0.8107	**0.9317**	0.9075	**0.9177**
CSIQ	PLCC	**0.9750**	0.6350	0.8899	**0.9790**	0.9676	0.9718
	SRCC	**0.9504**	0.5642	0.8694	**0.9560**	0.9303	0.9370
LIVE	PLCC	**0.9761**	0.5667	0.9214	**0.9762**	0.9646	0.9728
	SRCC	**0.9722**	0.4088	0.9131	**0.9727**	0.9593	0.9673
VCL	PLCC	**0.9043**	0.2949	0.6816	**0.9265**	0.8683	0.8867
	SRCC	**0.9017**	0.1923	0.6498	**0.9268**	0.8652	0.8845
MICT	PLCC	0.7967	0.4646	0.7647	0.8189	**0.8316**	**0.8475**
	SRCC	0.7865	0.4443	0.7450	0.8217	**0.8248**	**0.8474**
ESPL	PLCC	**0.9510**	0.6458	0.9414	**0.9626**	0.9398	0.9370
	SRCC	0.9144	0.6412	0.9154	**0.9333**	**0.9285**	0.9265

5.3.1 Complexity

To show the efficiency of the proposed indices, a run-time comparison between six IQAs is performed and shown in Table 5.3. The experiments were performed on a Core i7 3.40 GHz CPU with 16 GB of RAM. The IQA model was implemented in MATLAB 2013b running on Windows 7. It can be seen that MUG and MUG$^+$ have satisfactory run-times. Compared to the competing metric GridSAR, the proposed metric is about 250 times faster.

Table 5.3 Run time comparison of six IQA models when applied on an image of 1080×1920 size.

Index	Time (ms)
JPEGind (Wang *et al.*, 2002)	140.21
SSIM (Wang *et al.*, 2004)	187.85
MUG	222.06
MUG$^+$	225.52
GridSAR (Li *et al.*, 2015)	56810.53
NJQA (Golestaneh & Chandler, 2014)	79983.76

5.4 Conclusion

In this letter, two novel image quality assessment models for JPEG compressed images were proposed. The proposed indices are very simple and do not need training. They are based on the two simple facts of gradient magnitude of JPEG compressed images. As a result of more JPEG compression, the number of unique gradient magnitude values decreases and the median value of unique gradient magnitude values increases. The extensive experimental results shown that the proposed indices are robust to block misalignment and have consistent performance on seven benchmark datasets.

5.5 Acknowledgments

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

CHAPTER 6

EFFICIENT NO-REFERENCE QUALITY ASSESSMENT AND CLASSIFICATION MODEL FOR CONTRAST DISTORTED IMAGES

Hossein Ziaei Nafchi[1], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Submitted to IEEE Transactions on Broadcasting
November 2017

Abstract

In this paper, an efficient Minkowski Distance based Metric (MDM) for no-reference (NR) quality assessment of contrast distorted images is proposed. It is shown that higher orders of Minkowski distance and entropy provide accurate quality prediction for the contrast distorted images. The proposed metric performs predictions by extracting only three features from the distorted images followed by a regression analysis. Furthermore, the proposed features are able to classify type of the contrast distorted images with a high accuracy. Experimental results on the three datasets of CSIQ, TID2013, and CCID2014 show that the proposed metric with a very low complexity provides better quality predictions than the state-of-the-art NR metrics. The MATLAB source code of the proposed metric will be soon available to public at http://www.synchromedia.ca/system/files/MDM.zip.

Keywords

Image quality assessment, No-reference quality assessment, Contrast distortion, Minkowski distance.

6.1 Introduction

Image quality assessment (IQA) is a very important step in many image processing applications such as monitoring, benchmarking, restoration and parameter optimization (Wang *et al.*, 2004). Human visual system can easily have a fair judgment on the quality of the images.

However, subjective assessment of images is a very time consuming task. Hence, many IQA models (IQAs) have been proposed to automatically provide objective quality assessment of images (Wang *et al.*, 2004; Larson & Chandler, 2010; Sheikh & Bovik, 2006; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Zhang *et al.*, 2014; Lin & Kuo, 2011; Moorthy & Bovik, 2011b; Saad *et al.*, 2012b; Mittal *et al.*, 2012; Xue *et al.*, 2014a; Mittal *et al.*, 2013; Ye *et al.*, 2013). Among them, NR-IQAs (Moorthy & Bovik, 2011b; Saad *et al.*, 2012b; Mittal *et al.*, 2012; Xue *et al.*, 2014a; Mittal *et al.*, 2013; Ye *et al.*, 2013) are of high interest because in most present and emerging practical real-world applications, the reference signals are not available (Wang & Bovik, 2011). NR-IQAs do not need any information on the reference image. It is worth to mention that reduced-reference (RR) metrics (Wang & Simoncelli, 2005; Wang *et al.*, 2006; Li & Wang, 2009) need partial information about the reference image and that the full-reference (FR) metrics (Wang *et al.*, 2004; Larson & Chandler, 2010; Sheikh & Bovik, 2006; Zhang *et al.*, 2011; Xue *et al.*, 2014b; Lin & Kuo, 2011; Nafchi *et al.*, 2016) require the reference image.

Contrast distortion, which lies within the scope of this paper, is commonly produced in image acquisition setup. Poor and varying illumination conditions and poor camera's quality can drastically change image contrast and visibility. Figure 6.1 shows six examples of contrast distorted images. Several contrast enhancement methodologies have been proposed to adjust image contrast. These methods may over/under estimate the amount of contrast distortion and fail at enhancement accordingly. These methods, however, can use prior information provided by IQAs to overcome this wrong estimation.

With introduction of quality aware images (Wang *et al.*, 2006), RR-IQAs have shown their usefulness at assessment of image distortions caused by transmission in particular. Prior information about reference image is embedded inside the image to be transmitted, and the receiver decodes this information and uses it for quality assessment and even correction of distortions. The resulting metrics that eventually don't need training are good examples to illustrate RR-IQAs. In (Gu *et al.*, 2013), a RR-IQA called the RIQMC was proposed to assess the quality of contrast distorted images. RIMQC is a two-step model that uses entropy and four order statis-

Figure 6.1 Sample contrast distorted images from CCID2014 (Gu *et al.*, 2015a).

tics, e.g. mean, standard deviation, skewness and kurtosis. These are then linearly combined and a quality score is calculated. Seven parameters of the RIQMC are trained based on the 322 images of the CID2013 dataset that were also introduced in (Gu *et al.*, 2013). The performance of the RIQMC is very high and at the level of the leading FR-IQA models. The RIQMC was further modified in (Gu *et al.*, 2015a) by computing the phase congruency of the reference and distorted images. In (Gu *et al.*, 2014a), a more efficient RR-IQA called QMC was proposed that uses entropy and saliency features of the reference and distorted images for quality prediction. RCIQM is a more recent RR-IQA model that benefits from a bottom-up and top-down strategy (Liu *et al.*, 2017). It is based on bottom-up analysis of the free energy principle and top-down analysis of histograms of the reference and distorted images. RCIQM delivers a high performance for quality assessment of the contrast distorted images. The problem with these RR-IQAs is that they necessarily need reference or original image to be available.

There are limited methods in order to assess quality of the contrast distorted images (Fang *et al.*, 2015; Gu *et al.*, 2017). The authors in (Fang *et al.*, 2015) use a natural scene statistics (NSS) induced model to blindly predict the quality of contrast distorted images. They also use five features based on the NSS models of mean, standard deviation, skewness, kurtosis and entropy. Then, support vector regression is utilized to find a mapping function between these five feature set and subjective quality scores. They used 16873 images to train their NSS

model. The NR-IQA model in (Gu *et al.*, 2017) called NIQMC takes into account both local and global aspects of the contrast distorted images. In the local part, entropy of salient regions is computed. For the global part, a histogram analysis is proposed. NIQMC provides accurate quality predictions for contrast distorted images. The problem with this method is its high computational time.

In this paper, we propose a NR-IQA metric that is highly efficient and provides high prediction accuracy at the same time. We have found that the standard deviation (SD) alone provides a moderate quality prediction accuracy for global contrast distorted images. The SD when used to compare contrast level of two images is called root-mean-square (rms) contrast (Peli, 1990). The promising performance of rms contrast for global contrast changed images motivates us to use a variation of the Minkowski distance formulation along with the power-law transformation for no-reference quality assessment of contrast distorted images (NR-CDIQA). Power-law transformations are traditional image processing techniques that have been previously used for gamma correction and contrast manipulation. Previously, the Minkowski distance has been mainly used in IQA for two main purposes. The Minkowski metric has been used as a FR-IQA metric (Bovik, 2000), and the Minkowski pooling as a pooling strategy (Wang & Shang, 2006). Minkowski error metric between reference image R and distorted image D is defined as:

$$\mathrm{E}_\rho = \Big(\sum_{i=1}^{N} \left| R_i - D_i \right|^\rho \Big)^{1/\rho}. \tag{6.1}$$

where N is the number of image pixels, and $\rho \geq 1$ refers to the Minkowski power. Also, given any local similarity (S) map computed between a reference and distorted image by an IQA model, the Minkowski pooling is defined as

$$\mathrm{M} = \frac{1}{N} \sum_{i=1}^{N} S_i^\rho. \tag{6.2}$$

where M is the quality score of that IQA model. Except for the case $\rho = 1$ which is equal to the mean pooling, Minkowski pooling is rarely used in the literature (Xue *et al.*, 2014b; Nafchi *et al.*, 2016).

In this paper, we use higher orders of the Minkowski distance along with the power-law transformation and entropy to provide accurate quality predictions for contrast distorted images. In addition, the features of the proposed metric are able to classify type of the contrast distorted images. This information can be very useful in enhancing the contrast distorted images in real-time. To the best of our knowledge, classification of contrast distortion types has not been considered in the literature. In the following, the main contributions of the paper as well as its differences with respect to the previous works are briefly explained.

The proposed metric uses higher orders of Minkowski distance along with the power-law transformations, while in previous works like (Peli, 1990; Fang *et al.*, 2015), only the rms contrast or second order image statistic is used. To the best of our knowledge, Minkowski distance has not been used for the purpose of no-reference image quality assessment.

Entropy is widely used in previous studies for the purpose of contrast distortion assessment (Gu *et al.*, 2013; Fang *et al.*, 2015; Gu *et al.*, 2015a, 2014a; Liu *et al.*, 2017; Gu *et al.*, 2017). The proposed NR-IQA metric also uses entropy but despite having much lower complexity delivers higher and more consistent predictions than existing NR-IQA models on different datasets.

The three features of the proposed method are able to classify the type of contrast distorted images with a high accuracy, while features of existing method are not suitable for this task.

6.2 Proposed Metric (MDM)

Proposed NR-IQA of contrast distorted images follows the Minkowski distance formulation. Let's define the deviation as the variation of data values compared to a measure of central tendency (MCT) such as the mean, median, or mode. A deviation is in fact Minkowski distance of order ρ between an arbitrary vector $\mathbf{x}$ and its MCT:

$$\mathrm{D}(\mathbf{x},\rho) = \Big(\sum_{i=1}^{N} \big|\mathbf{x}_i - \mathrm{MCT}(\mathbf{x})\big|^{\rho}\Big)^{1/\rho}. \tag{6.3}$$

where, $\mathbf{x}_i$ denotes a vector value, MCT refers to the mean value of vector $\mathbf{x}$, and $\rho \geq 1$ indicates to the type of deviation. The proposed NR-IQA model uses a variation of the equation (6.3) as follows:

$$\widehat{\mathrm{D}}(\mathbf{x},\rho) = \Big(\frac{1}{N}\sum_{i=1}^{N} \big|\mathbf{x}_i - \mathrm{MCT}(\mathbf{x})\big|^{\rho}\Big)^{1/\rho}. \tag{6.4}$$

where, $\frac{1}{N}$ accounts for image resolution. Equation (6.4) is equivalent to the mean absolute deviation for $\rho = 1$ and equivalent to the standard deviation (rms contrast) for $\rho = 2$. Let D denotes the distorted image and D^q denotes the pixel-wise distorted image D to the power q, which is known as power-law transformation. Also let MCT^q denotes the mean value of the image D^q. The proposed NR metric MDM for distorted image D is computed by the following equation:

$$\mathrm{MDM}^{\rho,q}(D) = \sqrt[4]{\Big(\frac{1}{N}\sum_{i=1}^{N} \big|D_i^q - \mathrm{MCT}(D^q)\big|^{\rho}\Big)^{1/\rho}}. \tag{6.5}$$

where, D_i^q denotes one pixel of distorted image to the power of q. The fourth root in above equation is used for better numerical stability and visualization of quality scores. The reason for inclusion of parameter q is that contrast distorted images may follow the gamma transfer function in the form of $D = R^q$. In this paper, a large value of q is used. This large value of q will most likely increase the severity of distortion of D. From one prospective, this effect can be compared with the strategy proposed in (Crete-Roffet *et al.*, 2007). In (Crete-Roffet *et al.*, 2007), the input image is blurred and the result is compared with the input image in order to blindly assess its blurriness. Here, a similar strategy is used except that D^q is not compared with D. Figure 6.2 illustrates the impact of parameter q on the input intensity level.

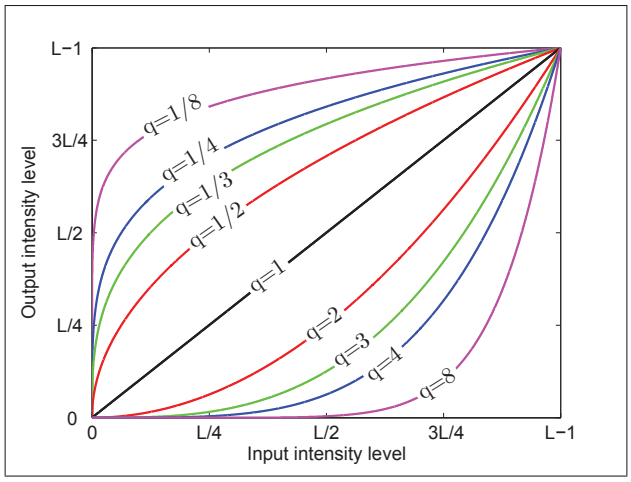

Figure 6.2 Output intensity level versus input intensity level ($q = 1$) for different values of q.

Equation 6.5 computes the first feature of the proposed metric for a distorted image D. The proposed metric also computes a second feature by the same equation from the complement of a contrast distorted image, e.g. $\overline{D} = 255 - D$. Except for some special cases, $\text{MDM}^{\rho,q}(D) \neq \text{MDM}^{\rho,q}(\overline{D})$. Note that rms contrast of D and $\overline{D}$ are equal. In experiments, the values of ρ and q are set to 128 and 8.

The two MDM based features are highly suitable for quality assessment of global contrast change ($D = R^q$) and mean shift ($D = R \pm \triangle$) distorted images, where $\triangle$ is a scalar within dynamic range of R. Figure 6.3 shows values of these two features versus MOS values for 250 contrast distorted images of TID2013 dataset (Ponomarenko *et al.*, 2013). This plot shows that values of the proposed Minkowski-based features are proportional to the MOS values.

In addition, the proposed Minkowski-based features can be used to classify contrast distortion types. Figure 6.4 shows plot of the first versus second Minkowski-based feature. The points of each distortion type on the plot can be separated with a high accuracy which shows the ability of these two features for classifying contrast distortions.

Additionally, the proposed metric uses entropy of the distorted image along with the MDM features as the third feature. Entropy is a common statistical measure of randomness which is

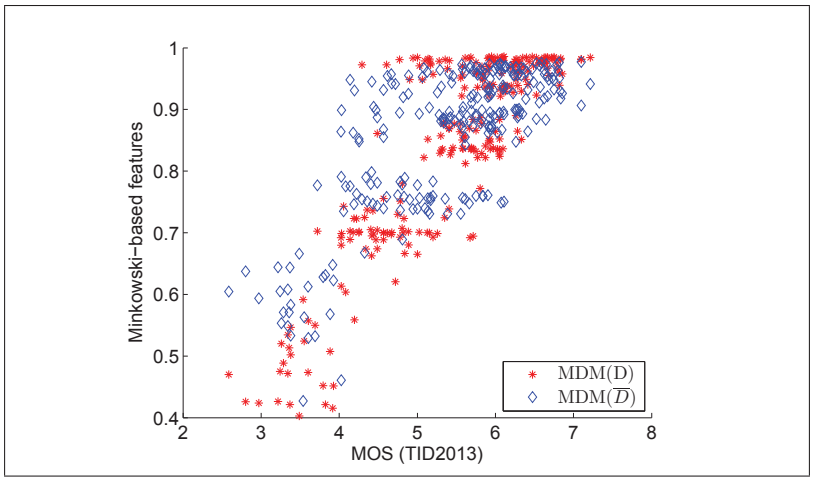

Figure 6.3 Two Minkowski-based features versus MOS for TID2013 dataset.

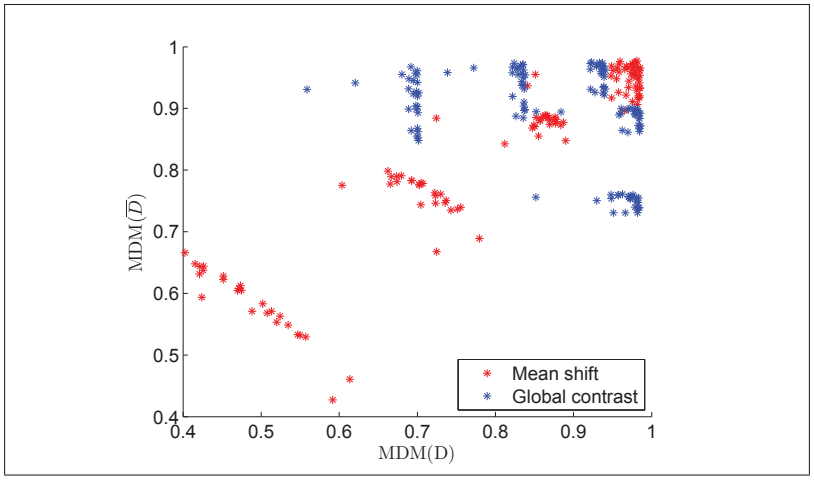

Figure 6.4 Visualization of the two Minkowski based features MDM(D) and MDM($\overline{D}$) for two contrast distortion types.

useful in analyzing texture of the images. Previous study (Gu *et al.*, 2015a) states that high-contrast image often has large entropy. The entropy is defined as:

$$\mathrm{H}(D) = -\sum_{L=0}^{255} \mathrm{P}_L(D) \log_2 \mathrm{P}_L(D) \tag{6.6}$$

where, $P_L(D)$ is the probability density of L-th intensity level. Therefore, two Minkowski-based features and entropy form the feature vector of the proposed metric. Support vector regression (SVR) is used to map these three features to the mean opinion scores (MOS). For the purpose of contrast distortion classification, support vector classifier (SVC) is used to assign a label to each image which indicates to the type of the contrast distortion.

6.3 Experimental results

6.3.1 Contrast distorted datasets

In the experiments, contrast distorted images of three standard datasets are used. The TID2013 (Ponomarenko *et al.*, 2013) dataset contains 125 global contrast changed images, and 125 images with mean shift distortion. CSIQ (Larson & Chandler, 2010) is another dataset that contains 116 global contrast distorted images in total. CCID2014 is a dedicated dataset of contrast distorted images (Gu *et al.*, 2015a). It 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. Please refer to ref. (Gu *et al.*, 2015a) for detailed explanation. The TID2008 (Ponomarenko *et al.*, 2009) and CID2013 (Wang *et al.*, 2013) datasets are not used in this paper because they are subsets of TID2013 and CCID2014, respectively.

6.3.2 Objective evaluation

For objective evaluation, two evaluation metrics were used in the experiments: the Spearman Rank-order Correlation coefficient (SRC), and the Pearson linear Correlation Coefficient (PCC) after a nonlinear regression analysis. The SRC and PCC metrics measure prediction monotonicity and prediction accuracy, respectively. The reported PCC values in this paper are computed after mapping quality scores to MOS based on the following logistic function:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5 \tag{6.7}$$

where β_1, β_2, β_3, β_4 and β_5 are fitting parameters computed by minimizing the mean square error between quality predictions x and subjective scores MOS.

Table 6.1 Performance comparison of the proposed NR-IQA model MDM and thirteen popular/competing indices on three benchmark datasets of contrast distorted images

Index	Type	PSNR	SSIM	VIF	IWSSIM	FSIM$_c$	MDSI	RIQMC	QMC	RCIQM	QAC	NIQE	NSS	NIQMC	MDM
		FR	FR	FR	FR	FR	FR	RR	RR	RR	NR	NR	NR	NR	NR
TID	PCC	0.4755	0.5735	**0.8458**	0.6919	0.6468	0.7028	0.8619	0.7710	**0.8866**	0.1683	-0.0734	0.5317	0.7225	**0.9285**
2013	SRC	0.5020	0.4992	**0.7716**	0.4528	0.4398	0.4859	0.8010	0.7071	**0.8541**	0.0278	-0.0652	0.4053	0.6458	**0.8989**
CSIQ	PCC	0.8888	0.7891	0.9439	**0.9614**	0.9452	0.9580	0.9605	0.9622	**0.9645**	0.3737	0.3025	0.8265	0.8747	**0.9665**
	SRC	0.8621	0.7922	0.9345	**0.9539**	0.9438	0.9446	0.9501	0.9554	**0.9569**	0.2533	0.2284	0.7994	0.8533	**0.9486**
CCID	PCC	0.4112	0.8308	**0.8588**	0.8353	0.8204	0.8576	0.8701	**0.8960**	0.8845	-0.2765	0.4458	0.7878	0.8438	**0.8717**
2014	SRC	0.6743	0.8174	**0.8349**	0.7822	0.7657	0.8128	0.8430	**0.8722**	0.8565	-0.1419	0.3655	0.7753	0.8113	**0.8363**

Six FR-IQAs including the PSNR, SSIM (Wang *et al.*, 2004), VIF (Sheikh & Bovik, 2006), IWSSIM (Wang & Li, 2011), FSIM$_c$ (Zhang *et al.*, 2011), MDSI (Nafchi *et al.*, 2016), and three RR-IQAs, e.g. RIQMC (Gu *et al.*, 2015a), QMC (Gu *et al.*, 2014a) and RCIQM (Liu *et al.*, 2017), and four NR-IQAs including QAC (Xue *et al.*, 2013), NIQE (Mittal *et al.*, 2013), NSS (Fang *et al.*, 2015) and NIQMC (Gu *et al.*, 2017) were used in the experiments.

Table 6.1 provides a performance comparison between proposed NR-IQA, e.g. MDM, and thirteen FR/RR/NR-IQAs in terms of SRC and PCC. The best performing FR/RR/NR IQAs are highlighted for each category. It can be seen that RR-IQAs that are designated to assess contrast distorted images provide relatively good prediction accuracy on different datasets. Among FR-IQAs, the performance of VIF is noticeable. With a comparison between NR-IQAs, the following conclusions can be drawn. First, the proposed index MDM performs very well on the three datasets. MDM outperforms NR-IQAs on the three datasets. The recently proposed NR metric for contrast distorted images NIQMC is only comparable to the proposed metric on the CCID2014 dataset. MDM outperforms all the indices listed in Table 6.1 on 250 contrast distorted images of the TID2013 dataset. On the other datasets, the proposed index MDM is comparable to the best performing RR metrics RIQMC, QMC and RCIQM. However, the popular NR-IQA model NSS (Fang *et al.*, 2015) shows inconsistent predictions on different

datasets. Also, multi-purpose NR-IQAs like QAC and NIQE have major difficulty in quality assessment of contrast distorted images. It can be concluded that three features of the proposed method are more powerful than the five features of NSS for the purpose of assessing quality of contrast distorted images.

Table 6.2 Performance comparison of the proposed metric (MDM) and NSS for different train-test setups on the three datasets.

NR index	TID2013		CSIQ		CCID2014	
	SRC	PCC	SRC	PCC	SRC	PCC
20%-80%						
NSS	0.2507	0.3239	0.7347	0.7491	0.7686	0.7525
MDM	**0.8707**	**0.9103**	**0.9237**	**0.9316**	**0.8215**	**0.8564**
50%-50%						
NSS	0.3514	0.4702	0.7737	0.7884	0.7807	0.7663
MDM	**0.8810**	**0.9184**	**0.9348**	**0.9477**	**0.8273**	**0.8620**
80%-20%						
NSS	0.4053	0.5317	0.7994	0.8265	0.7878	0.7753
MDM	**0.8989**	**0.9285**	**0.9486**	**0.9665**	**0.8363**	**0.8717**

In Table 6.2, the performance of the NR metric NSS (Fang *et al.*, 2015) and the proposed metric are listed for different train and test setups. Each dataset is divided into different randomly chosen subsets and the results are reported on the basis of the median value of 1000 times train-test for three cases: 20% train 80% test, 50% train 50% test, and 80% train 20% test. The splits are done in a way that image contents are different for train and test. Hence, for CCID2014 and TID2013 datasets, 0.5333% train 0.4667% test and 52% train 48% test is used respectively instead of the 50%-50% train-test. From the results of the Table 6.2, it can be seen that the proposed metric performs very well with small number of training data. Also, the proposed metric with three features outperforms the five-features metric NSS on the three datasets.

6.3.3 Contrast distortion classification

While no-reference image quality assessment of contrast distorted images is of great interest, classification of contrast distortion types provides very useful additional information that can be used for automatic and fast contrast enhancement. In this paper, the three features of the proposed metric are used to classify type of contrast distortions. The only dataset with more than one type of contrast distortion and with known labels for each distortion type is TID2013 (Ponomarenko *et al.*, 2013). TID2013 contains 125 distorted images with global contrast change and 125 distorted images with mean shift. Table 6.3 lists accuracy results of the proposed method and NSS for contrast distortion classification on the TID2013 dataset. In this experiment, image contents for train and test has no overlap. The three features of the proposed method can fairly classify contrast distortions even with small number of training data. However, five features of NSS do not have enough discriminative power to be used for this classification task. Results of Table 6.3 verify these statements.

Table 6.3 Contrast distortion classification accuracy of the three features of the proposed method and five features of NSS for different setups of train and test.

NR index	20%-80%	50%-50%	80%-20%
NSS	0.5925	0.6250	0.6400
MDM	**0.8525**	**0.9000**	**0.9200**

6.3.4 Parameters

The proposed index MDM has two parameters to set, e.g. q and ρ. Experimentally, we found that MDM has its maximum prediction accuracy for $q = \{8, 10\}$ and some $50 \leq \rho \leq 130$. Further increasing the value of q will have little effect on the performance. Apart from the performance, being a power of 2 was another consideration on the choice of the parameters ρ and q because MDM runs faster in this case (please refer to subsection Complexity).

6.3.5 Complexity

To show the efficiency of the proposed metric, a run-time comparison between fourteen IQAs is performed and shown in Table 6.4. The experiments were performed on a Core i7 3.40 GHz CPU with 16 GB of RAM. The IQA model was implemented in MATLAB 2013b running on Windows 7. It can be seen that PSNR and MDM are the top two fastest indices for images with different resolution, respectively. Depending on the image resolution, MDM runs faster than PSNR because the code is optimized to calculate power operations in O($\log \rho$) and O($\log q$) instead of O(ρ) and O(q) respectively, and that the distorted image is downsampled by a factor of $M = max(2, [min(h,w)/512])$. Here, h and w are image height and width, and $[.]$ is the round operator. In addition, the proposed method only processes the distorted image, while PSNR processes both reference and its distorted version. In comparison with the most competing NR metric NIQMC which is also proposed for contrast distortion assessment, the proposed method is about 180 to 550 times faster. Clearly, the proposed index is highly efficient and can be used in real-time applications.

Table 6.4 Run time comparison of IQA models in terms of milliseconds

IQA model	384×512	1080×1920	2160×3840
PSNR	5.61	37.51	145.83
SSIM (Wang *et al.*, 2004)	14.99	77.59	287.92
VIF (Sheikh & Bovik, 2006)	572.93	6162.70	25381.32
IWSSIM (Wang & Li, 2011)	228.11	2499.94	10471.56
$FSIM_c$ (Zhang *et al.*, 2011)	142.06	600.23	1562.81
MDSI (Nafchi *et al.*, 2016)	12.77	153.47	781.41
RIQMC (Gu *et al.*, 2015a)	743.90	2868.79	6313.60
QMC (Gu *et al.*, 2014a)	9.40	51.38	232.42
RCIQM (Liu *et al.*, 2017)	N/A	N/A	N/A
QAC (Xue *et al.*, 2013)	151.88	1706.15	7180.60
NIQE (Mittal *et al.*, 2013)	187.80	1726.97	6878.10
NSS (Fang *et al.*, 2015)	23.92	247.13	976.11
NIQMC (Gu *et al.*, 2017)	2897.52	10580.00	34190.62
MDM	5.23	56.91	60.40

6.4 Conclusion

In this paper, an image quality assessment and classification model for contrast distorted images was proposed. The proposed index is very simple and runs in real-time. The proposed index (MDM) uses two Minkowski distance based features and entropy information to assess simple and complex types of contrast distortions. For the first time, the features of the proposed metric were used to classify type of the contrast distortions with a high accuracy. A comparison with the state-of-the-art no-reference IQAs verifies that the proposed metric MDM runs much faster and provides better prediction accuracy on different benchmark datasets than existing NR metrics. In addition, compared to the existing state-of-the-art full reference and reduced reference IQAs, the proposed index shows comparable or better prediction accuracy.

6.5 Acknowledgments

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

CHAPTER 7

CORRC2G: COLOR TO GRAY CONVERSION BY CORRELATION

Hossein Ziaei Nafchi[1], Atena Shahkolaei[1], Rachid Hedjam[2], Mohamed Cheriet[1]

[1] Département de Génie de la production automatisée, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Department of Geography, McGill University,
805 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

Published in IEEE Signal Processing Letters
Volume 24, Issue 11, November 2017, Pages 1651-1655

Abstract

In this letter, a novel decolorization method is proposed to convert color images into grayscale. The proposed method, called CorrC2G, estimates the three global linear weighting parameters of the color to gray conversion by correlation. These parameters are estimated directly from the correlations between each channel of the RGB image and a contrast image. The proposed method works directly on the RGB channels; it does not use any edge information nor any optimization or training. The objective and subjective experimental results on three available benchmark datasets of color to gray conversion, e.g. Cadik, CSDD and Color250, show that the proposed decolorization method is highly efficient and comparable to recent state-of-the-art decolorization methods. The MATLAB source code of the proposed method is available at: https://www.mathworks.com/matlabcentral/fileexchange/64628.

Keywords

Decolorization, color to gray conversion, correlation, luminance, grayscale, RGB.

7.1 Introduction

In many real-world image/video processing and computer vision applications, the 3D color image needs to be transformed into a 1D grayscale image. This is a lossy but a necessary conver-

sion for several applications (Kanan & Cottrell, 2012). Recent years have seen several efforts in developing novel decolorization methods that are more likely to follow human perception of brightness and contrast (Gooch *et al.*, 2005; Neumann *et al.*, 2007; Grundland & Dodgson, 2007; Smith *et al.*, 2008; Kim *et al.*, 2009; Lu *et al.*, 2012; Song *et al.*, 2013, 2014; Du *et al.*, 2015; Liu *et al.*, 2015, 2016; Tao *et al.*, 2017). Color to gray (C2G) conversion methods can be categorized into global, local, and hybrid. The global mapping approach has the potential to produce natural looking grayscale outputs. In contrast, local mapping techniques (Neumann *et al.*, 2007; Smith *et al.*, 2008) that better preserve the local contrast may produce unnatural outputs. In local mapping methods, the same color pixel within an image might be mapped into different grayscale values, which is generally not desired. Therefore, several methods consider global and local contrast or features for conversion (Kuk *et al.*, 2011; Jin *et al.*, 2014; Du *et al.*, 2015). Besides, video decolorization methods such as (Song *et al.*, 2014; Tao *et al.*, 2017) are specifically developed in order to maintain temporal coherence of videos.

Since the proposed method belongs to the category of global mapping, we focus on these methods. Gooch et al. (Gooch *et al.*, 2005) proposed a method to maintain color contrast between pixel pairs by optimizing an objective contrast function. Kim et al. (Kim *et al.*, 2009) proposed a non-linear parametric model in which the parameters are estimated by minimizing an objective function that preserves color differences. In several recent global mapping methods, the input color image I is converted into a grayscale output g by linear weighting of the R, G, and B channels, i.e. $g(i,j) = \sum_{c=R,G,B} \lambda_c \mathrm{I}_c(i,j)$, where $\sum_{c=R,G,B} \lambda_c = 1$. Here, the three linear weighting parameters λ, should be estimated on the basis of some models. In (Lu *et al.*, 2012), a gradient error energy function is minimized to compute the three linear weighting parameters. This interesting approach was given notable consideration and some variations of this method has been proposed (Liu *et al.*, 2015, 2016). While the method of (Liu *et al.*, 2015) objectively preserves the contrast and run in real-time, it may produce grayscale outputs with an unnatural appearance. In contrast, the method proposed in (Liu *et al.*, 2016) produces mostly natural outputs but at the cost of being several times slower.

In this letter, we propose a novel decolorization method that estimates the three global linear weighting parameters λ directly from the R, G, and B channels. The correlations between each channel of the color image with a base image map, which is very likely to preserve contrast, are mapped to λ. To the best of our knowledge, correlation has not been used for the purpose of C2G conversion. The proposed method takes into account both the magnitude and sign of the correlation values to adjust the weighting parameters. The proposed training-free method is very simple, it runs in real-time and offers perceptually consistent grayscale outputs with good contrast preservation.

7.2 Proposed Decolorization method

The proposed decolorization method is a global mapping approach that estimates the three linear weighting parameters λ from correlation. Correlation is a measure of association between variables (Rodgers & Nicewander, 1988). Here, we use Pearson's measure of correlation between two variables X and Y, which is commonly defined as:

$$\rho_{X,Y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2\right]^{1/2}} \tag{7.1}$$

where, $\bar{X}$ and $\bar{Y}$ are means of variables X and Y, respectively. It is worth noting that an equivalent formula for ρ is $s_{XY}/s_X s_Y$, where s_{XY} is the sample covariance, and s_X and s_Y are sample standard deviations. Given the R, G, and B channels of a color image, the correlation between each channel with an image map that is likely to preserve contrast is computed and normalized in order to estimate the three weighting parameters λ.

In order to produce a 2D image which reflects the contrast of the color image, the following two images are combined. The first image is simply the mean image μ, and the second is the standard deviation image σ:

$$\mu(i,j) = \frac{1}{3} \sum_{c=R,G,B} \mathrm{I}_c(i,j) \tag{7.2}$$

$$\sigma(i,j) = \left[\frac{1}{2} \sum_{c=R,G,B} \left| \mathrm{I}_c(i,j) - \mu(i,j) \right|^2 \right]^{1/2} \tag{7.3}$$

The values of σ are further divided by 147.2243, which is the maximum possible value of σ. In what follows, the rationale to use the two images, μ and σ, and their combination are explained. Considering the [0 255] range for the 3D RGB image, there are 256^3 = 16,777,216 possible color values. However, possible values for the 2D images μ and σ (1D vectors) are just 766 and 16,365, respectively. This clearly indicates the probability of contrast loss as a result of the 3D to 1D conversion. In order to reduce the probability of contrast loss, we propose to use the pointwise product of μ and σ, which provides more than 2 million possible values:

$$\mathrm{Q}(i,j) = \mu(i,j) \times \sigma(i,j) \tag{7.4}$$

Figure 7.1 shows a color image with its mean μ, standard deviation σ, and contrast map Q.

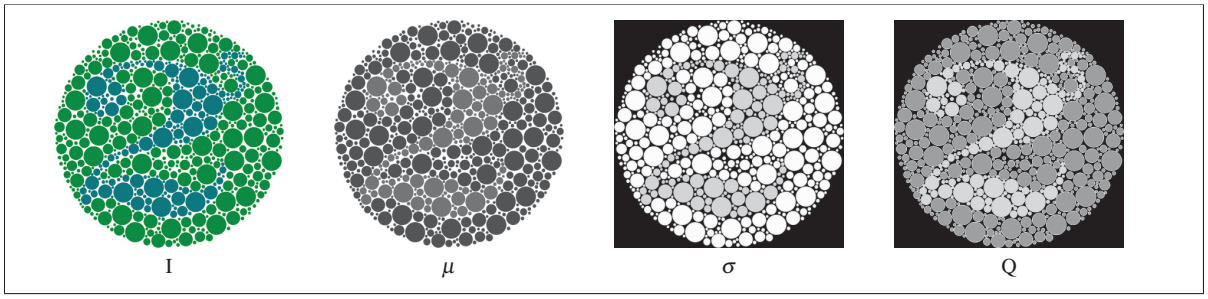

Figure 7.1 An example of the mean image μ, standard deviation image σ, and contrast map Q for a color image I.

The three Pearson correlation values between RGB channels and Q are denoted as: $P = \{\rho_{RQ}, \rho_{GQ}, \rho_{BQ}\}$. Each correlation value lies in [-1 1]. Given these three correlations, the purpose is to map them to the three weighting parameters, i.e. $P \longrightarrow \lambda$. For mapping, the unsigned correlations (absolute values) are mapped to β, and original values of the correlations are mapped to γ. Finally, λ is computed from β and γ. The reason for the consideration of the absolute correlations along with the signed correlations is to avoid mapping larger negative correlations to the smaller weighting parameters. The parameters of λ are computed using the following simple calculations. The first assumption is that a channel with a higher correlation with the contrast map Q should take a larger weighting parameter:

$$\beta_c = \frac{|P_c|}{\sum |P|} \tag{7.5}$$

where, c is the channel index and $\sum \beta = 1$. At the same time, a channel with inverse correlation with Q should take a lower weighting parameter:

$$\gamma_c = \frac{P_c - min_P}{max_P - min_P} - 0.5 \tag{7.6}$$

where, min_P is the minimum value of P, max_P is the maximum value of P, and $\gamma \in [-0.5 \; +0.5]$. In the above equation, the constant 0.5 can be replaced with any other value in range [0 1] to control the contribution of the inverse correlations. λ is estimated by adding $min(\beta, \gamma)$ to β:

$$\lambda_c = |\beta_c + min(\beta_c, \gamma_c)| \tag{7.7}$$

Finally, λ is normalized so that $\sum_{c=R,G,B} \lambda_c = 1$. In the above equation, γ is used only when it decreases weighting parameters. We recall that γ was used to decrease the weighting parameters of those channels with inverse correlations. In contrast to the other methods that estimate the weighting parameters λ by optimizing an objective function (Gooch *et al.*, 2005;

Lu *et al.*, 2012; Liu *et al.*, 2015, 2016), the proposed method directly estimates λ from the correlation values. The problem with such methods is that the defined objective function does not necessarily follow human perception of brightness and contrast.

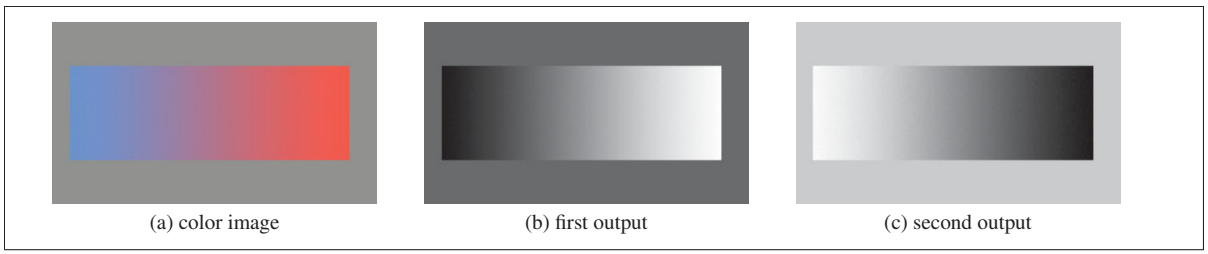

Figure 7.2 Two possible grayscale outputs of the proposed method for a color image. (b) is produced by using the standard deviation image σ, and (c) is produced by using the complement of σ.

Depending on the preference of users for the perceived color, more saturated colors are perceived to be either brighter or darker than their luminance (Kim *et al.*, 2009). In this regard, the standard deviation image σ can be replaced with its complement image $(1-\sigma)$ and parameters of λ can be estimated accordingly. In this approach, the proposed method has two grayscale outputs. Figure 7.2 gives an example of these two grayscale outputs. Some users may prefer one or the other of them, while others may evaluate them as equal. Since the proposed method should produce a single output, the one with more and larger peaks at the middle of its histogram is chosen as the final output. This two-output strategy slightly improves the objective and subjective results. The objective performance can be greatly improved if the final output of the method is chosen according to the C2G evaluation metrics, such as the E-score (Lu *et al.*, 2014) and C2G-SSIM (Ma *et al.*, 2015). We did not use these metrics because they are several times slower than the histogram analysis approach.

7.3 Experimental results

In the experiments, three available datasets for evaluation of the color to gray methods are used. The Cadik dataset (Čadík, 2008a) contains 25 (originally 24) saturated color images. The Color250 dataset (Lu *et al.*, 2014) comprises 250 color images with a wide range of natural and

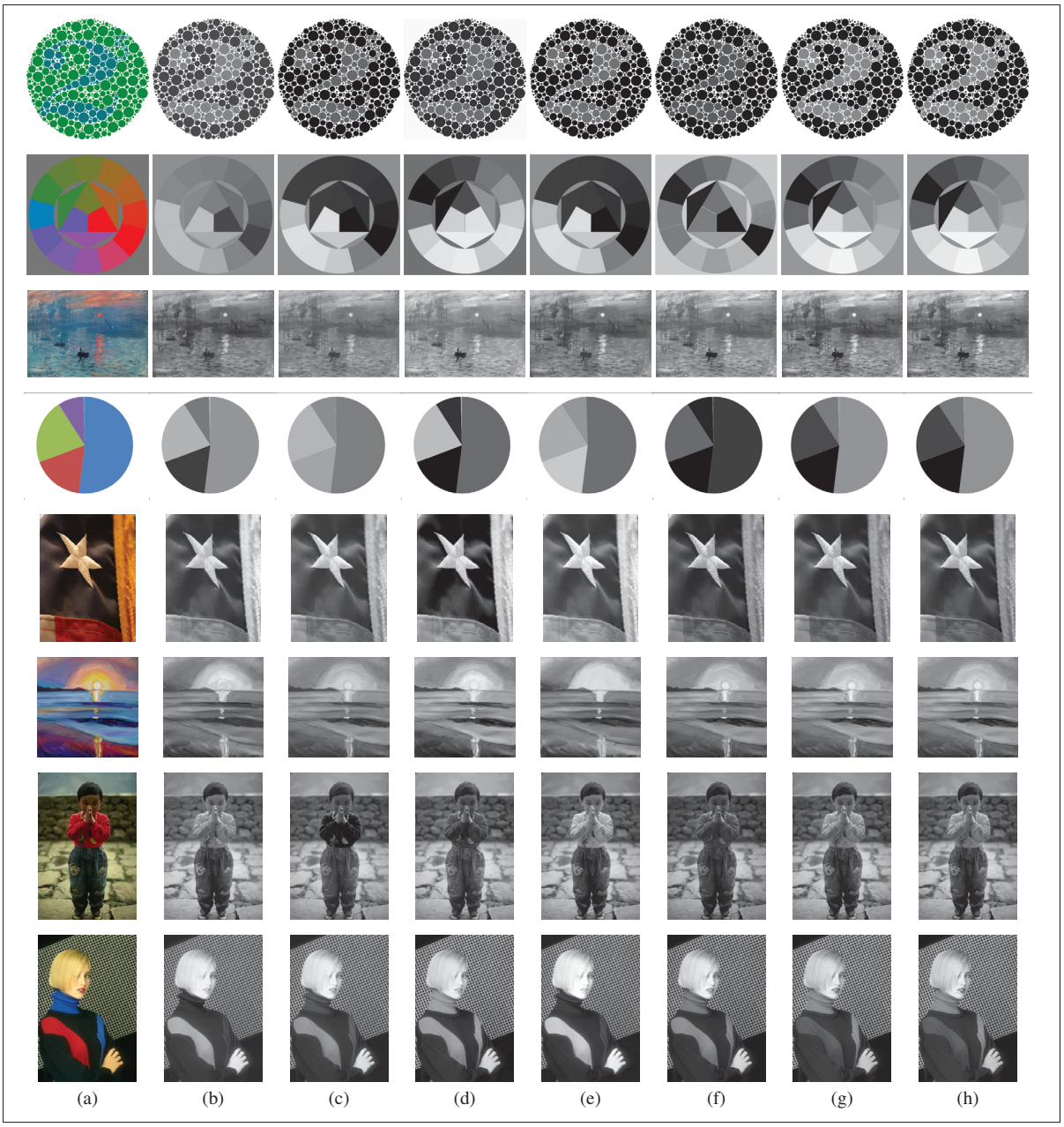

Figure 7.3 Visual comparison of six color to gray conversion methods. (a) color image, (b) Decolorize (Grundland & Dodgson, 2007), (c) RTCP (Lu *et al.*, 2012), (d) Saliency (Du *et al.*, 2015), (e) GcsDecolor (Liu *et al.*, 2015), (f) SPDecolor (Liu *et al.*, 2016), (g) CorrC2G (r=512), (h) CorrC2G (r=256). To view finer detail, please zoom in on the electronic version.

synthetic images. The third dataset is CSDD (Du *et al.*, 2015), which contains 22 color images with abundant colors and patterns. For objective evaluation, two objective quality assessment

metrics for color to gray image conversion are used: E-score (Lu *et al.*, 2014) and C2G-SSIM (Ma *et al.*, 2015). E-score evaluates both the color contrast preservation ratio (CCPR) and color content fidelity ratio (CCFR). C2G-SSIM is a more recent C2G evaluation metric based on the popular image quality assessment metric SSIM (Wang *et al.*, 2004). In comparison with the E-score, C2G-SSIM provides a useful quality map and shows higher correlation with human subjective evaluations. Five state-of-the-art C2G methods were chosen for comparison (Grundland & Dodgson, 2007; Lu *et al.*, 2012; Du *et al.*, 2015; Liu *et al.*, 2015, 2016). Each of these methods has shown very promising performance in comparison to the other existing methods. In Figure 7.3, outputs of the six C2G methods for eight color images are shown. Here, r is a downsampling parameter (see subsection Complexity). It can be seen from Figure 7.3 that the proposed method fairly shows the color differences.

According to the results of Table 7.1, GcsDecolor provides highest performance based on E-score, and the proposed method shows highest performance based on C2G-SSIM. In terms of E-score, the proposed method shows better results than the Decolorize, Saliency, and SPDecolor methods.

Table 7.1 The average performance of six C2G methods for 297 images

C2G method (297 images)	E-score ($\tau = 15$)	C2G-SSIM
Decolorize (Grundland & Dodgson, 2007)	0.8972	0.8639
RTCP (Lu *et al.*, 2012)	0.9115	0.8770
Saliency (Du *et al.*, 2015)	0.8965	0.8705
GcsDecolor (Liu *et al.*, 2015)	**0.9162**	0.8707
SPDecolor (Liu *et al.*, 2016)	0.8952	0.8775
CorrC2G (r = 512)	0.8981	**0.8796**
CorrC2G (default, r = 256)	0.8987	**0.8796**
CorrC2G (r = 128)	0.8957	0.8774
CorrC2G (r = 64)	0.8944	0.8777

It is common to report qualitative performance based on the CCPR by varying its parameter τ, a threshold below which the color differences become almost invisible to the human visual system (Lu *et al.*, 2014). Figure 7.4 shows the results for six C2G methods. We can see that

GcsDecolor yields the best results, and that the proposed method is comparable with the other methods.

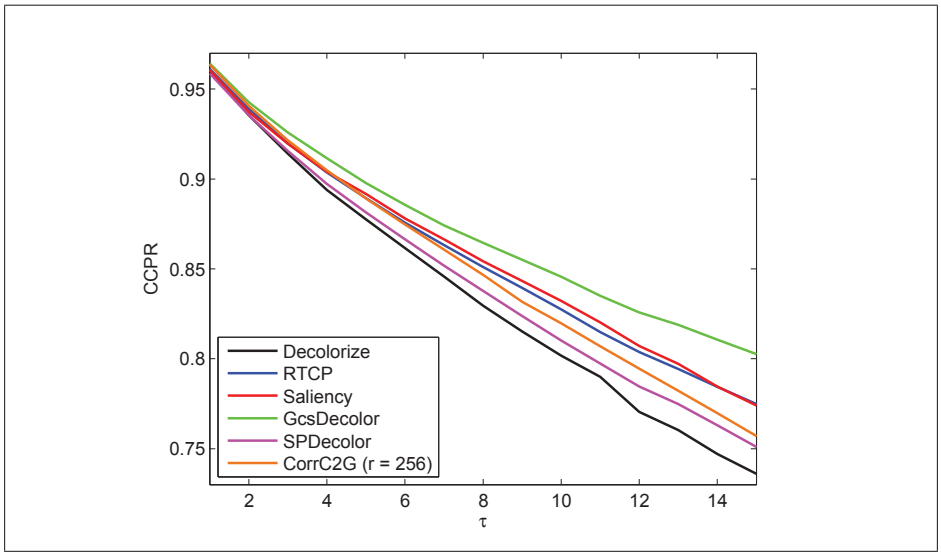

Figure 7.4 Comparison of six C2G methods based on the CCPR metric.

We also conducted a subjective evaluation in which the outputs of the proposed C2G method are compared with those of other methods. In this experiment, three subjects with a background in image processing were asked to discuss why they preferred the output of one method to that of another. After discussion, only one vote was given for each pair-comparison: 'worse', 'equal', or 'better'. The subjects had no prior knowledge of the compared methods and that the grayscale pairs were randomly placed on the screen. In this experiment, 297 color images of the three datasets were considered. In total, 2970 comparisons were recorded. The results of the five C2G methods against the proposed method are listed in Table 7.2. We can see that the proposed method shows better performance than the other methods. For example, outputs of GcsDecolor are rated worse than the proposed method CorrC2G (r = 256) for 127 images, they are rated equal for 139 images, and better for 31 images. GcsDecolor, which has the highest E-score performance, shows the worst results on the basis of subjective evaluation. In fact, the correlation between C2G-SSIM and the subjective evaluations is higher than that for the E-score.

Table 7.2 Results of subjective evaluation for five C2G methods against the proposed method for 297 color images

C2G method	CorrC2G (r = 512)			CorrC2G (r = 256)		
	worse	equal	better	worse	equal	better
Decolorize (Grundland & Dodgson, 2007)	81	166	50	80	162	55
RTCP (Lu *et al.*, 2012)	69	200	28	68	198	31
Saliency (Du *et al.*, 2015)	91	168	38	91	167	39
GcsDecolor (Liu *et al.*, 2015)	129	140	28	127	139	31
SPDecolor (Liu *et al.*, 2016)	49	209	39	50	206	41

7.3.1 Complexity

To show the efficiency of the proposed method, a run-time comparison between six C2G methods was performed and is shown in Table 7.3. The experiments were performed on a Core i7 3.40 GHz CPU with 16 GB of RAM. The proposed method was implemented in MATLAB 2013b running on Windows 7. The proposed method first downsamples the color image with a factor of $f = r/min(h,w)$, where h and w are image height and width, and constant $r = 256$ is used by default. Then, it estimates the weighting parameters from the downsampled image. It can be seen from Table 7.3 that CorrC2G runs faster than the other methods for images with different resolution. Also, the proposed method runs faster by reducing its downsampling parameter r. This speedup is smaller for larger images because the majority of the run-time is spent on the common operations that are dependent to the image size but independent from the value of r. For majority of the images, reducing the value of r to some threshold does not affect the visual appearance of the outputs. Figure 7.5 verifies this behavior for eight images.

7.4 Conclusion

This letter introduces a novel correlation-based decolorization method to convert color images into grayscale. The Pearson correlations between channels of color images with a contrast map are mapped directly to the three linear weighting parameters. The proposed method is very simple and runs in real-time, yet it offers perceptually consistent outputs with a natural appearance. Extensive objective and subjective experiments on the three benchmark datasets

Table 7.3 Run time comparison of C2G methods in terms of milliseconds

C2G method	128×128	384×512	1080×1920	2160×3840
Decolorize (Grundland & Dodgson, 2007)	4.28	65.63	771.15	3078.83
RTCP (Lu *et al.*, 2012)	11.87	19.04	67.20	219.96
Saliency (Du *et al.*, 2015)	-	-	-	-
GcsDecolor (Liu *et al.*, 2015)	16.43	25.40	73.28	226.97
SPDecolor (Liu *et al.*, 2016)	25.51	225.97	2239.70	7883.76
CorrC2G (r = 512)	2.87	37.67	138.18	230.36
CorrC2G (r = 256)	2.87	16.76	50.46	147.46
CorrC2G (r = 128)	2.87	7.45	39.75	137.41
CorrC2G (r = 64)	2.78	5.60	35.16	134.37

Figure 7.5 Outputs of the proposed method given the color images of Figure 7.3 for r = 128 (first and third columns) and r = 64 (second and fourth columns). Except for image 'sunrise', other outputs are quite similar.

verified the efficiency of the proposed method. Several possible ways of improving the current method would be to use a better numerically compatible contrast map, more accurate mapping of the correlations to the three weighting parameters, and to modify the double-output behavior of the proposed method.

7.5 Acknowledgments

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

CHAPTER 8

GENERAL DISCUSSION

This thesis has addressed several problems related to the image quality assessment. The introduction and literature review (Chapter 1) showed limitations of current features, similarity maps and derived IQA metrics. Specifically, the following question was investigated: what are efficient similarity maps and features that can improve performance and efficiency of image quality assessment metrics? We established five research objectives in Chapter 2 that led to the development of two novel full-reference and two novel no-reference IQA models as we as a perceptually consistent color to gray image conversion method. These methods made their own contributions and were presented, evaluated and discussed in Chapter 3, Chapter 4, Chapter 5, Chapter 6, and Chapter 7. All of the aspects studied in this thesis were related to aspects of the human visual system. Our contributions are now discussed in the following sections by considering their advances made in the state of the art image quality assessment, with a focus on their strength and limitations.

8.1 Efficient and reliable full-reference image quality assessment for natural, synthetic and photo-retouched images

No or few FR-IQA models can deal with various image distortions, image contents, and run in real-time. These models usually use computationally expensive features like phase congruency and saliency, and may follow a multi-scale approach which is also inefficient. This has led to our attempt to introduce similarity maps that are more likely to follow HVS (Chapter 3). The new gradient similarity is computed by a fusion technique that enables more accurate measurement of structural distortions. The color similarity formula is maximally optimized for efficiency. The proposed deviation pooling formula is used to compute a quality map from the gradient and color similarity maps. Deviation pooling considers both magnitude and spread of distortions which is more close to the way HVS judges the image quality. Considering hand-held devices that use different display technologies, gamut mapping has become very popular. Gamut mapping usually introduces specific color distortions that are challenging to be assessed

by traditional general-purpose FR-IQA models. The proposed method uses simple color space, and may not be able to accurately predict the quality of images with gamut mapping effects. In general, the color similarity of the proposed metric is potentially its weakness for some specific applications.

8.2 Full-reference image quality assessment for tone-mapped images

Very few metrics are available for quality assessment of tone-mapped images. The existing metric converts HDR and corresponding LDR (tone-mapped) images into a color space and measure their differences by an SSIM-induced approach. This approach is promising, but an approach with the ability to directly compare HDR with LDR can be of high interest. We proposed a new metric in Chapter 4 that compares local phase information of HDR with that of LDR. It shows good performance compared to the state of the art metric TMQI. The proposed metric uses local phase of a log-transformed version of HDR. While this technique is shown to provide better performance, still HDR is not directly compared to the LDR. This means that the proposed metric is not yet a dynamic range invariant metric. However, it can be further improved by removing the comparisons based on the log image. Also, the proposed metric provides channel by channel assessment. A method to combine computed scores from the three channels can improve the performance at the cost of more computational cost.

8.3 Block-size and misalignment invariant no-reference image quality assessment model for JPEG compressed images

Too many NR-IQA metrics for JPEG compressed images are available, but none of them is parameterless nor robust to block size and misalignment. The only available metric which is robust to misalignment first localizes the blocks inside an image and then measure its blockiness. In addition, some better performing blockiness metrics are numerically unstable. The lack of such a robust and invariant metric has led to our first attempt to propose a parameterless metric which is invariant to block size and misalignment (Chapter 5). We considered two simple facts about blockiness artifact. As a result of more JPEG compression, the num-

ber of unique gradient magnitude values decreases, and the median value of unique gradient magnitude values increases. A potential flaw of the proposed metric is when edge distributions of images are highly different. We use a simple normalization strategy, and then compute median and number of gradient values. An alternative normalization can help to improve the aforementioned limitation of the proposed metric.

8.4 Efficient no-reference quality assessment and classification of contrast distorted images

There are a limited number of NR-IQA metrics for contrast distorted images. The available metrics may not deliver high performance or cannot be used in real-time applications. To fill this gap, we proposed the first image contrast assessment metric that is highly efficient and delivers high performance as well (Chapter 6). High orders of Minkowski distance and power transformation are used for feature extraction. The metric is optimized to run in real-time. Unlike features used by an existing metric, the proposed features are able to classify contrast distortions with a high accuracy. Large datasets of contrast distorted images with labeled distortion type are not available. Such dataset can help to better evaluate the proposed method and others for distortion classification task.

8.5 Efficient color to gray image conversion by correlation

We have seen the quality assessment of tone-mapped images in Chapter 4. Given a color image, quality assessment of its gray-scale image is a challenging task. Few metrics are available to be used for quality assessment of color to gray image conversion. In Chapter 7, we proposed a perceptually consistent C2G method based on the correlation. This new C2G method was evaluated based on subjective ratings. Then, we found that C2G-SSIM better correlates with subjective ratings. Still, there is room to propose better performing C2G assessment metrics. While efficiency and high performance of the proposed C2G method are shown, it has some limitations. It does not consider coherency of video frames. It is a double-output method which is not desirable in general. Therefore, considering video coherency, modifying double output

behavior, and using better numerically compatible contrast maps can improve the proposed C2G method.

CONCLUSION AND RECOMMENDATIONS

In this thesis, we have presented original contributions to the state of the art in the field of image quality assessment. Similarity maps and features are the ground on which the image quality assessment models are built. These similarity maps and features should be effective and efficient to enable development of efficient image quality assessment models. We have introduced several features and similarity maps most of which were efficient. We also have shown their effectiveness in different full-reference and no-reference image quality assessment scenarios.

For full-reference image quality assessment of low dynamic range images, the contributions of this thesis show directions for efficient full-reference image quality assessment design. Quality assessment with expensive image transforms or complex color spaces increase the computational complexity and leave several parameters to set. Therefore, derived similarity maps from fast differential operators and color spaces with very low complexity are necessary to design real-time full-reference image quality assessment models. For full-reference image quality assessment of tone-mapped images, the contribution of this thesis can open path to a dynamic range invariant full-reference image quality assessment model.

No-reference image quality assessment metrics are of very high interest because in real-world applications, usually, the original signal is not available. Usually, multi-purpose NR-IQA metrics are computationally expensive. Therefore, efficient distortion-specific NR-IQA metrics can be used for specific applications. The contributions of this thesis on distortion-specific no-reference image quality assessment emphasize on the efficiency of the models by indicating that even real-time image processing systems can benefit from these metrics.

Future work

Despite achieving outstanding speed and performance, the proposed full-reference metric for low dynamic range images is not perfect especially for quality assessment of color distortions and sparse images. One solution would be adjusting the contribution of the color similarity and the gradient similarity in an adaptive way. This has to be done with cautious to avoid numerical instability.

Features that are invariant to the dynamic range are of high interest to be used for quality assessment of tone-mapped images. Such features should take into account both achromatic and chromatic distortions that are common in case of tone-mapped images.

One flaw in the proposed blockiness metric is that image content is not strongly considered by the proposed model. Proper normalization of the edge distribution can help to avoid potential inaccurate assessments.

Current features of no-reference image quality assessment model for contrast distorted images are tested to classify two types of contrast distortions. This is because existing datasets with several contrast distortions are not labeled. It would be interesting to develop such datasets. Joint assessment and classification models have not been studied before in the literature. More research in this direction is of high interest.

Extension of the proposed image decolorization method to maintain coherency in video frames can be considered in future works. Correlation values are robust under global contrast changes, therefore the difference between correlations of video frames can be used to maintain coherency.

Summary of contributions

In the following, we briefly highlight the major contributions of this thesis.

- Efficient similarity maps and deviation pooling for reliable full-reference image quality assessment,
- A feature similarity index for quality assessment of tone-mapped images,
- Parameterless no-reference image quality assessment metric for JPEG compressed images which is robust to block size and misalignment,
- Highly efficient features to assess and classify contrast distorted images,
- Efficient perceptually consistent color to gray image conversion method based on the correlation.

Articles in peer reviewed journals (contributions of the thesis)

1. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam and Mohamed Cheriet: Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. IEEE Access, vol. 4, pp. 5579-5590, 2016.

2. Hossein Ziaei Nafchi, Atena Shahkolaei, Reza Farrahi Moghaddam and Mohamed Cheriet: FSITM: A Feature Similarity Index For Tone-Mapped Images. IEEE Signal Processing Letters, vol. 22, no. 8, pp. 1026-1029, 2015.

3. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam and Mohamed Cheriet: MUG: A Parameterless No-Reference JPEG Quality Evaluator Robust to Block Size and Misalignment. IEEE Signal Processing Letters, vol. 23, no. 11, pp. 1577-1581, 2016.

4. Hossein Ziaei Nafchi and Mohamed Cheriet: Efficient No-Reference Quality Assessment and Classification Model for Contrast Distorted Images. *submitted to* IEEE Transactions on Broadcasting, (November 2017).

5. Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam and Mohamed Cheriet: CorrC2G: Color to Gray Conversion by Correlation. IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1651-1655, 2017.

Other articles in peer reviewed journals

1. Atena Shahkolaei, Hossein Ziaei Nafchi, Somaya Al-Maadeed and Mohamed Cheriet: Subjective and objective quality assessment of degraded document images. Journal of Cultural Heritage, (in press).

2. Rachid Hedjam, Margaret Kalacska, Max Mignotte, Hossein Ziaei Nafchi and Mohamed Cheriet: Iterative Classifiers Combination Model for Change Detection in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, 2016.

3. Rachid Hedjam, Hossein Ziaei Nafchi, Margaret Kalacska and Mohamed Cheriet: Influence of Color-to-Gray Conversion on the Performance of Document Image Binarization: Toward a Novel Optimization Problem. IEEE Transactions on Image Processing, vol. 24, no. 11, 2015.

4. Hossein Ziaei Nafchi, Reza Farrahi Moghaddam and Mohamed Cheriet: Phase-Based Binarization of Ancient Document Images: Model and Applications. IEEE Transactions on Image Processing, vol. 23, no. 7, 2014.

BIBLIOGRAPHY

Aydin, T. O., Mantiuk, R., Myszkowski, K. & Seidel, H.-P. (2008). Dynamic Range Independent Image Quality Assessment. *ACM Trans. Graph.*, 27(3), 69:1–69:10.

Bae, S. H. & Kim, M. (2016a). DCT-QM: A DCT-Based Quality Degradation Metric for Image Quality Optimization Problems. *IEEE Transactions on Image Processing*, 25(10), 4916-4930. doi: 10.1109/TIP.2016.2598492.

Bae, S. H. & Kim, M. (2016b). A Novel Image Quality Assessment with Globally and Locally Consilient Visual Quality Perception. *IEEE Transactions on Image Processing*, 25(5), 2392-2406. doi: 10.1109/TIP.2016.2545863.

Bovik, A. & Liu, S. (2001). DCT-domain blind measurement of blocking artifacts in DCT-coded images. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3, 1725-1728. doi: 10.1109/ICASSP.2001.941272.

Bovik, A. (Ed.). (2000). *Handbook of Image and Video Processing*. Academic Press.

Canny, J. (1986). A computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6), 679-698.

Chandler, D. M. (2013). Seven Challenges in Image Quality Assessment: Past, Present, and Future Research. *ISRN Signal Processing*, 2013, 1-53. Article ID. 905685.

Chandler, D. & Hemami, S. (2007). VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Transactions on Image Processing*, 16(9), 2284-2298. doi: 10.1109/TIP.2007.901820.

Chang, H.-W., Yang, H., Gan, Y. & Wang, M.-H. (2013). Sparse Feature Fidelity for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, 22(10), 4007-4018. doi: 10.1109/TIP.2013.2266579.

Chen, C. & Bloom, J. (2010). A Blind Reference-Free Blockiness Measure. In Qiu, G., Lam, K., Kiya, H., Xue, X.-Y., Kuo, C.-C. & Lew, M. (Eds.), *Advances in Multimedia Information Processing* (vol. 6297, pp. 112-123). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-15702-8_11.

Chen, G.-H., Yang, C.-L. & Xie, S.-L. (2006, Oct). Gradient-Based Structural Similarity for Image Quality Assessment. *IEEE International Conference on Image Processing*, pp. 2929-2932. doi: 10.1109/ICIP.2006.313132.

Cisco Systems. (2015). Cisco Visual Networking Index: Forecast and Methodology, 2014-2019.

Crete-Roffet, F., Dolmiere, T., Ladret, P. & Nicolas, M. (2007). The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric. *SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*.

Damera-Venkata, N., Kite, T., Geisler, W., Evans, B. & Bovik, A. (2000). Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4), 636-650. doi: 10.1109/83.841940.

Daniel L Ruderman. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5(4), 517-548.

Du, H., He, S., Sheng, B., Ma, L. & Lau, R. W. H. (2015). Saliency-Guided Color-to-Gray Conversion Using Region-Based Optimization. *IEEE Transactions on Image Processing*, 24(1), 434-443. doi: 10.1109/TIP.2014.2380172.

Eilertsen, G., Wanat, R., Mantiuk, R. K. & Unger, J. (2013). Evaluation of Tone Mapping Operators for HDR-Video. *Computer Graphics Forum*, 32(7), 275–284. doi: 10.1111/cgf.12235.

Fang, Y., Ma, K., Wang, Z., Lin, W., Fang, Z. & Zhai, G. (2015). No-Reference Quality Assessment of Contrast-Distorted Images Based on Natural Scene Statistics. *IEEE Signal Processing Letters*, 22(7), 838-842. doi: 10.1109/LSP.2014.2372333.

Geusebroek, J.-M., Van den Boomgaard, R., Smeulders, A. & Geerts, H. (2001). Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12), 1338-1350. doi: 10.1109/34.977559.

Ghadiyaram, D. & Bovik, A. C. (2017). Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach. *Journal of Vision (in press)*.

Golestaneh, S. & Chandler, D. (2014). No-Reference Quality Assessment of JPEG Images via a Quality Relevance Map. *IEEE Signal Processing Letters*, 21(2), 155-158. doi: 10.1109/LSP.2013.2296038.

Gooch, A. A., Olsen, S. C., Tumblin, J. & Gooch, B. (2005). Color2Gray: Salience-preserving Color Removal. *ACM Trans. Graph.*, 24(3), 634–639. doi: 10.1145/1073204.1073241.

Grundland, M. & Dodgson, N. A. (2007). Decolorize: Fast, contrast enhancing, color to grayscale conversion. *Pattern Recognition*, 40(11), 2891 - 2896. doi: 10.1016/j.patcog.2006.11.003.

Gu, K., Zhai, G., Yang, X., Zhang, W. & Chen, C. (2014a). Automatic Contrast Enhancement Technology with Saliency Preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(9), 1480-1494. doi: 10.1109/TCSVT.2014.2372392.

Gu, K., Zhai, G., Lin, W. & Liu, M. (2015a). The Analysis of Image Contrast: From Quality Assessment to Automatic Enhancement. *IEEE Transactions on Cybernetics*, 46(1), 284-297. doi: 10.1109/TCYB.2015.2401732.

Gu, K., Wang, S., Zhai, G., Lin, W., Yang, X. & Zhang, W. (2016a). Analysis of Distortion Distribution for Pooling in Image Quality Prediction. *IEEE Transactions on Broadcasting*, 62(2), 446-456. doi: 10.1109/TBC.2015.2511624.

Gu, K., Lin, W., Zhai, G., Yang, X., Zhang, W. & Chen, C. W. (2017). No-Reference Quality Metric of Contrast-Distorted Images Based on Information Maximization. *IEEE Transactions on Cybernetics*, 47(12), 4559-4565. doi: 10.1109/TCYB.2016.2575544.

Gu, K., Zhai, G., Yang, X., Zhang, W. & Liu, M. (2013, Sept). Subjective and objective quality assessment for images with contrast change. *2013 20th IEEE International Conference on Image Processing (ICIP)*, pp. 383-387. doi: 10.1109/ICIP.2013.6738079.

Gu, K., Zhai, G., Liu, M., Yang, X. & Zhang, W. (2014b, June). Details preservation inspired blind quality metric of tone mapping methods. *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, pp. 518-521. doi: 10.1109/ISCAS.2014.6865186.

Gu, K., Liu, M., Zhai, G., Yang, X. & Zhang, W. (2015b). Quality Assessment Considering Viewing Distance and Image Resolution. *IEEE Transactions on Broadcasting*, 61(3), 520-531. doi: 10.1109/TBC.2015.2459851.

Gu, K., Zhai, G., Lin, W., Yang, X. & Zhang, W. (2016b). Learning a blind quality evaluation engine of screen content images. *Neurocomputing*, 196, 140 - 149. doi: 10.1016/j.neucom.2015.11.101.

Hassen, R., Wang, Z. & Salama, M. (2013). Image Sharpness Assessment Based on Local Phase Coherence. *IEEE Trans. on Image Processing*, 22(7), 2572-2581.

Horita, Y., Shibata, K., Kawayoke, Y. & Sazzad, Z. M. P. MICT Image Quality Evaluation Database 2000.

I. Sobel, G. F. (1968). *A 3x3 isotropic gradient operator for image processing*. Presented at a talk at the Stanford Artificial Project.

ITU-T P. 1401. (2012). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.

Jin, Z., Li, F. & Ng, M. K. (2014). A Variational Approach for Image Decolorization by Variance Maximization. *SIAM Journal on Imaging Sciences*, 7(2), 944-968. doi: 10.1137/130935197.

Kanan, C. & Cottrell, G. W. (2012). Color-to-Grayscale: Does the Method Matter in Image Recognition? *PLOS ONE*, 7(1), 1-7. doi: 10.1371/journal.pone.0029740.

Kim, D.-O., Han, H.-S. & Park, R.-H. (2010). Gradient information-based image quality metric. *IEEE Transactions on Consumer Electronics*, 56(2), 930-936. doi: 10.1109/TCE.2010.5506022.

Kim, Y., Jang, C., Demouth, J. & Lee, S. (2009). Robust Color-to-gray via Nonlinear Global Mapping. *ACM Trans. Graph.*, 28(5), 161:1–161:4. doi: 10.1145/1618452.1618507.

Kovesi, P. (1999a). Phase Preserving Denoising of Images. *International Conference on Digital Image Computing: Techniques and Applications*.

Kovesi, P. (1999b). Image Features from Phase Congruency. *Videre: Journal of Computer Vision Research*, 1, 1-26.

Kovesi, P. (2000). Phase congruency: A low-level image invariant. *Psychological Research*, 64, 136-148.

Kovesi, P. (2003). Phase Congruency Detects Corners and Edges. *International Conference on Digital Image Computing: Techniques and Applications*.

Kovesi, P. (2002). Edges Are Not Just Steps. *Asian Conference on Computer Vision*, pp. 1-6.

Kuk, J. G., Ahn, J. H. & Cho, N. I. (2011). A Color to Grayscale Conversion Considering Local and Global Contrast. In Kimmel, R., Klette, R. & Sugimoto, A. (Eds.), *Computer Vision – ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part IV* (pp. 513–524). Asian Conference on Computer Vision. doi: 10.1007/978-3-642-19282-1_41.

Kundu, D. & Evans, B. L. (2015). Full-reference visual quality assessment for synthetic images: A subjective study. *IEEE International Conference on Image Processing*, pp. 2374-2378.

Larson, E. C. & Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1), 011006. doi: 10.1117/1.3267105.

Ledda, P., Chalmers, A., Troscianko, T. & Seetzen, H. (2005). Evaluation of Tone Mapping Operators Using a High Dynamic Range Display. *ACM Trans. Graph.*, 24(3), 640–648.

Lee, D. & Plataniotis, K. N. (2015). Towards a Full-Reference Quality Assessment for Color Images Using Directional Statistics. *IEEE Transactions on Image Processing*, 24(11), 3950-3965. doi: 10.1109/TIP.2015.2456419.

Lee, S. & Park, S. J. (2012). A new image quality assessment method to detect and measure strength of blocking artifacts. *Signal Processing: Image Communication*, 27(1), 31-38. doi: 10.1016/j.image.2011.08.002.

Li, C. & Bovik, A. C. (2010). Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, 25(7), 517-526. doi: 10.1016/j.image.2010.03.004. Special Issue on Image and Video Quality Assessment.

Li, L., Lin, W. & Zhu, H. (2014a). Learning Structural Regularity for Evaluating Blocking Artifacts in JPEG Images. *IEEE Signal Processing Letters*, 21(8), 918-922. doi: 10.1109/LSP.2014.2320743.

Li, L., Zhu, H., Yang, G. & Qian, J. (2014b). Referenceless Measure of Blocking Artifacts by Tchebichef Kernel Analysis. *IEEE Signal Processing Letters*, 21(1), 122-125. doi: 10.1109/LSP.2013.2294333.

Li, L., Zhou, Y., Wu, J., Lin, W. & Li, H. (2015). GridSAR: Grid strength and regularity for robust evaluation of blocking artifacts in JPEG images. *Journal of Visual Communication and Image Representation*, 30, 153-163. doi: 10.1016/j.jvcir.2015.04.001.

Li, Q. & Wang, Z. (2009). Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2), 202-211. doi: 10.1109/JSTSP.2009.2014497.

Li, S., Zhang, F., Ma, L. & Ngan, K. N. (2011). Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. *IEEE Transactions on Multimedia*, 13(5), 935-949. doi: 10.1109/TMM.2011.2152382.

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

Liu, A., Lin, W. & Narwaria, M. (2012). Image Quality Assessment Based on Gradient Similarity. *IEEE Transactions on Image Processing*, 21(4), 1500-1512. doi: 10.1109/TIP.2011.2175935.

Liu, H. & Heynderickx, I. (2009). A Perceptually Relevant No-reference Blockiness Metric Based on Local Image Characteristics. *EURASIP J. Adv. Signal Process*, 2009, 2:1–2:14. doi: 10.1155/2009/263540.

Liu, L., Dong, H., Huang, H. & Bovik, A. C. (2014a). No-reference image quality assessment in curvelet domain. *Signal Processing: Image Communication*, 29(4), 494 - 505. doi: 10.1016/j.image.2014.02.004.

Liu, M. & Ndjiki-Nya, P. (2012, July). A new perceptual-based no-reference contrast metric for natural images based on human attention and image dynamic. *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 254-259. doi: 10.1109/QoMEX.2012.6263887.

Liu, M., Gu, K., Zhai, G., Callet, P. L. & Zhang, W. (2017). Perceptual Reduced-Reference Visual Quality Assessment for Contrast Alteration. *IEEE Transactions on Broadcasting*, 63(1), 71-81. doi: 10.1109/TBC.2016.2597545.

Liu, Q., Liu, P. X., Xie, W., Wang, Y. & Liang, D. (2015). GcsDecolor: Gradient Correlation Similarity for Efficient Contrast Preserving Decolorization. *IEEE Transactions on Image Processing*, 24(9), 2889-2904. doi: 10.1109/TIP.2015.2423615.

Liu, Q., Liu, P., Wang, Y. & Leung, H. (2016). Semi-Parametric Decolorization with Laplacian-based Perceptual Quality Metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(9), 1856-1868. doi: 10.1109/TCSVT.2016.2555779.

Liu, X., Zhang, L., Li, H. & Lu, J. (2014b). Integrating Visual Saliency Information into Objective Quality Assessment of Tone-Mapped Images. In et al., D.-S. H. (Ed.), *LNCS 8588 (International Conference on Intelligent Computing, ICIC'14)* (pp. 376-386). Taiyuan, Shanxi, China: Springer.

Lu, C., Xu, L. & Jia, J. (2012). Real-time Contrast Preserving Decolorization. *SIGGRAPH Asia 2012 Technical Briefs*, (SA '12), 34:1–34:4. doi: 10.1145/2407746.2407780.

Lu, C., Xu, L. & Jia, J. (2014). Contrast Preserving Decolorization with Perception-Based Quality Metrics. *International Journal of Computer Vision*, 110(2), 222–239. doi: 10.1007/s11263-014-0732-6.

Ma, K., Zhao, T., Zeng, K. & Wang, Z. (2015). Objective Quality Assessment for Color-to-Gray Image Conversion. *IEEE Transactions on Image Processing*, 24(12), 4673-4685. doi: 10.1109/TIP.2015.2460015.

Ma, K., Yeganeh, H., Zeng, K. & Wang, Z. (2014). High Dynamic Range Image Tone Mapping by Optimizing Tone Mapped Image Quality Index. *IEEE International Conference on Multimedia & Expo*.

Mantiuk, R., Daly, S., Myszkowski, K. & Seidel, S. (2005). Predicting visible differences in high dynamic range images-model and its calibration. *Proc. SPIE*, 5666, 204-214.

Marziliano, P., Dufaux, F., Winkler, S. & Ebrahimi, T. (2002). A no-reference perceptual blur metric. *Image Processing. 2002. Proceedings. 2002 International Conference on*, 3, III-57-III-60 vol.3. doi: 10.1109/ICIP.2002.1038902.

Mittal, A., Moorthy, A. K. & Bovik, A. C. (2012). No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12), 4695-4708. doi: 10.1109/TIP.2012.2214050.

Mittal, A., Soundararajan, R. & Bovik, A. (2013). Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3), 209-212. doi: 10.1109/LSP.2012.2227726.

Moorthy, A. K. & Bovik, A. C. (2011a). Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, 20(12), 3350-3364. doi: 10.1109/TIP.2011.2147325.

Moorthy, A. & Bovik, A. (2009a). Visual Importance Pooling for Image Quality Assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2), 193-201. doi: 10.1109/JSTSP.2009.2015374.

Moorthy, A. & Bovik, A. (2011b). Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, 20(12), 3350-3364. doi: 10.1109/TIP.2011.2147325.

Moorthy, A. K. & Bovik, A. C. (2009b). Perceptually significant spatial pooling techniques for image quality assessment. *Proc. SPIE*, 7240, 724012-724012-11.

Morrone, M. & Burr, D. (1988). Feature detection in human vision: a phase-dependent energy model. *Royal Society of London*, B 235, 221-245.

Morrone, M., Ross, J., Burr, D. & Owens, R. (1986). Mach bands are phase dependent. *Nature*, 324(6049), 250-253.

Nafchi, H. Z., Shahkolaei, A., Hedjam, R. & Cheriet, M. (2016). Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. *IEEE Access*, 4, 5579-5590. doi: 10.1109/ACCESS.2016.2604042.

Narwaria, M. & Lin, W. (2010). Objective Image Quality Assessment Based on Support Vector Regression. *IEEE Transactions on Neural Networks*, 21(3), 515-519. doi: 10.1109/TNN.2010.2040192.

Neumann, L., Čadík, M. & Nemcsics, A. (2007). An Efficient Perception-based Adaptive Color to Gray Transformation. *Proceedings of the Third Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, (Computational Aesthetics'07), 73–80. doi: 10.2312/COMPAESTH/COMPAESTH07/073-080.

Obara, B., Fricker, M., Gavaghan, D. & Garu, V. (2012). Contrast-Independent Curvilinear Structure Detection in Biomedical Images. *IEEE Trans. on Image Processing*, 21(5), 2572-2581.

Oppenheim, A. V. & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69, 529-541.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Systems, Man, and Cybernetics*, 9(2), 62-66.

Pan, F., Lin, X., Rahardja, S., Ong, E. & Lin, W. (2004, June). Measuring blocking artifacts using edge direction information [image and video coding]. *IEEE International Conference on Multimedia and Expo*, 2, 1491-1494. doi: 10.1109/ICME.2004.1394519.

Papari, G. & Petkov, N. (2011). Edge and line oriented contour detection: State of the art. *Image and Vision Computing*, 29, 79-103.

Park, C.-S., Kim, J.-H. & Ko, S.-J. (2007). Fast Blind Measurement of Blocking Artifacts in both Pixel and DCT Domains. *Journal of Mathematical Imaging and Vision*, 28(3), 279-284. doi: 10.1007/s10851-007-0019-4.

Peli, E. (1990). Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10), 2032-2040.

Perra, C., Massidda, F. & Giusto, D. (2005, Sept). Image blockiness evaluation based on Sobel operator. *IEEE International Conference on Image Processing*, 1, I-389-92. doi: 10.1109/ICIP.2005.1529769.

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M. & Battisti, F. (2009). TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics*, 10, 30-45.

Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F. & Kuo, C.-C. (2013, June). Color image database TID2013: Peculiarities and preliminary results. *4th European Workshop on Visual Information Processing (EUVIP)*, pp. 106-111.

Reinhard, E., Ward, G., Pattanaik, S. & Debevec, P. (2010). *High Dynamic Range Imaging: Acquisition, Display, and Image-based Lighting*. Morgan Kaufmann.

Rezazadeh, S. & Coulombe, S. (2013). A novel discrete wavelet transform framework for full reference image quality assessment. *Signal, Image and Video Processing*, 7(3), 559–573. doi: 10.1007/s11760-011-0260-6.

Rodgers, J. L. & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), 59-66.

Saad, M. A., Bovik, A. C. & Charrier, C. (2012a). Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. *IEEE Transactions on Image Processing*, 21(8), 3339-3352. doi: 10.1109/TIP.2012.2191563.

Saad, M., Bovik, A. & Charrier, C. (2012b). Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. *IEEE Transactions on Image Processing*, 21(8), 3339-3352. doi: 10.1109/TIP.2012.2191563.

Sadr, J. & Sinha, P. (2004). Object recognition and Random Image Structure Evolution. *Cognitive Science*, 28(2), 259–287. doi: 10.1207/s15516709cog2802_7.

Saha, A. & Wu, Q. J. (2013). Perceptual image quality assessment using phase deviation sensitive energy features. *Signal Processing*, 93(11), 3182-3191.

Sampat, M., Wang, Z., Gupta, S., Bovik, A. & Markey, M. (2009). Complex Wavelet Structural Similarity: A New Image Similarity Index. *IEEE Transactions on Image Processing*, 18(11), 2385-2401. doi: 10.1109/TIP.2009.2025923.

Santhaseelan, V. & Asari, V. (2011). Phase Congruency Based Technique for the Removal of Rain from Video. *International Conference on Image Analysis and Recognition*.

Sheikh, H. & Bovik, A. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2), 430-444. doi: 10.1109/TIP.2005.859378.

Sheikh, H., Wang, Z., Cormack, L. & Bovik, A. LIVE Image Quality Assessment Database Release 2.

Sheikh, H., Bovik, A. & de Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12), 2117-2128. doi: 10.1109/TIP.2005.859389.

Sheikh, H., Sabir, M. & Bovik, A. (2006). A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *Image Processing, IEEE Transactions on*, 15(11), 3440-3451. doi: 10.1109/TIP.2006.881959.

Shnayderman, A., Gusev, A. & Eskicioglu, A. (2006). An SVD-based grayscale image quality measure for local and global assessment. *IEEE Transactions on Image Processing*, 15(2), 422-429. doi: 10.1109/TIP.2005.860605.

Smith, K., Landes, P., Thollot, J. & Myszkowski, K. (2008). Apparent greyscale: a simple and fast conversion to perceptually accurate images and video. *Comput Graph Forum*, 27(2), 193-200.

Song, Y., Bao, L. & Yang, Q. (2014, March). Real-time video decolorization using bilateral filtering. *IEEE Winter Conference on Applications of Computer Vision*, pp. 159-166. doi: 10.1109/WACV.2014.6836106.

Song, Y., Bao, L., Xu, X. & Yang, Q. (2013). Decolorization: Is Rgb2Gray() out? *SIGGRAPH Asia 2013 Technical Briefs*, (SA '13), 15:1–15:4. doi: 10.1145/2542355.2542374.

Tan, K. & Ghanbari, M. (2000a). Frequency domain measurement of blockiness in MPEG-2 coded video. *International Conference on Image Processing*, 3, 977-980. doi: 10.1109/ICIP.2000.899621.

Tan, K. & Ghanbari, M. (2000b). Blockiness detection for MPEG2-coded video. *IEEE Signal Processing Letters*, 7(8), 213-215. doi: 10.1109/97.855443.

Tao, Y., Shen, Y., Sheng, B., Li, P. & Lau, R. W. H. (2017). Video Decolorization Using Visual Proximity Coherence Optimization. *IEEE Transactions on Cybernetics*, PP(99), 1-14. doi: 10.1109/TCYB.2017.2695655.

Čadík, M. (2008a). Perceptual Evaluation of Color-to-Grayscale Image Conversions. *Computer Graphics Forum*, 27(7), 1745–1754. doi: 10.1111/j.1467-8659.2008.01319.x.

Čadík, M. (2008b). *Perceptually Based Image Quality Assessment and Image Transformations*. (Ph.D. thesis, Czech Technical University in Prague).

Video Quality Experts Group. (2003). Final report from the video quality experts group on the validation of objective models of video quality assessment.

Vu, C., Phan, T., Singh, P., & Chandler, D. M. (2012). Digitally Retouched Image Quality (DRIQ) Database.

Wang, S., Deng, C., Lin, W., Zhao, B. & Chen, J. (2013, Sept). A novel SVD-based image quality assessment metric. *20th IEEE International Conference on Image Processing (ICIP)*, pp. 423-426. doi: 10.1109/ICIP.2013.6738087.

Wang, Z. & Shang, X. (2006, Oct). Spatial Pooling Strategies for Perceptual Image Quality Assessment. *IEEE International Conference on Image Processing*, pp. 2945-2948. doi: 10.1109/ICIP.2006.313136.

Wang, Z., Simoncelli, E. & Bovik, A. (2003, Nov). Multiscale structural similarity for image quality assessment. *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, 2, 1398-1402. doi: 10.1109/ACSSC.2003.1292216.

Wang, Z. (2011). Applications of Objective Image Quality Assessment Methods [Applications Corner]. *IEEE Signal Processing Magazine*, 28(6), 137-142. doi: 10.1109/MSP.2011.942295.

Wang, Z. & Bovik, A. (2009). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*, 26(1), 98-117. doi: 10.1109/MSP.2008.930649.

Wang, Z. & Bovik, A. (2011). Reduced- and No-Reference Image Quality Assessment. *IEEE Signal Processing Magazine*, 28(6), 29-40. doi: 10.1109/MSP.2011.942471.

Wang, Z. & Li, Q. (2011). Information Content Weighting for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(5), 1185-1198. doi: 10.1109/TIP.2010.2092435.

Wang, Z. & Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. *in Proc. of SPIE Human Vision and Electronic Imaging*, pp. 149–159.

Wang, Z., Bovik, A. & Evan, B. (2000). Blind measurement of blocking artifacts in images. *International Conference on Image Processing*, 3, 981-984. doi: 10.1109/ICIP.2000.899622.

Wang, Z., Sheikh, H. R. & Bovik, A. (2002). No-reference perceptual quality assessment of JPEG compressed images. *International Conference on Image Processing*, 1, I-477-I-480. doi: 10.1109/ICIP.2002.1038064.

Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612. doi: 10.1109/TIP.2003.819861.

Wang, Z., Wu, G., Sheikh, H., Simoncelli, E., Yang, E.-H. & Bovik, A. (2006). Quality-aware images. *IEEE Transactions on Image Processing*, 15(6), 1680-1689. doi: 10.1109/TIP.2005.864165.

Wu, H. & Yuen, M. (1997). A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11), 317-320. doi: 10.1109/97.641398.

Xue, W., Mou, X., Zhang, L., Bovik, A. & Feng, X. (2014a). Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features. *IEEE Transactions on Image Processing*, 23(11), 4850-4862. doi: 10.1109/TIP.2014.2355716.

Xue, W., Zhang, L. & Mou, X. (2013, June). Learning without Human Scores for Blind Image Quality Assessment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 995-1002. doi: 10.1109/CVPR.2013.133.

Xue, W., Zhang, L., Mou, X. & Bovik, A. (2014b). Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing*, 23(2), 684-695. doi: 10.1109/TIP.2013.2293423.

Ye, P., Kumar, J., Kang, L. & Doermann, D. (2013, June). Real-Time No-Reference Image Quality Assessment Based on Filter Learning. *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 987-994. doi: 10.1109/CVPR.2013.132.

Yeganeh, H. & Wang, Z. (2010, Sept). Objective assessment of tone mapping algorithms. *17th IEEE International Conference on Image Processing (ICIP)*, pp. 2477-2480. doi: 10.1109/ICIP.2010.5651778.

Yeganeh, H. & Wang, Z. (2013a). Objective Quality Assessment of Tone-Mapped Images. *IEEE Transactions on Image Processing*, 22(2), 657-667. doi: 10.1109/TIP.2012.2221725.

Yeganeh, H. & Wang, Z. (2013b). Subject-rated image database of tone-mapped images.

Zaric, A., Tatalovic, N., Brajkovic, N., Hlevnjak, H., Loncaric, M., Dumic, E. & Grgic, S. (2012). VCL@FER Image Quality Assessment Database. *AUTOMATIKA*, 53, 344-354.

Zhai, G., Zhang, W., Yang, X., Lin, W. & Xu, Y. (2008). No-reference noticeable blockiness estimation in images. *Signal Processing: Image Communication*, 23(6), 417-432. doi: 10.1016/j.image.2008.04.007.

Zhang, L., Zhang, L., Zhang, D. & Guo, Z. (2012). Phase congruency induced local features for finger-knuckle-print recognition. *Pattern Recognition*, 45, 2522-2531.

Zhang, L. & Li, H. (2012, Sept). SR-SIM: A fast and high performance IQA index based on spectral residual. *19th IEEE International Conference on Image Processing (ICIP)*, pp. 1473-1476. doi: 10.1109/ICIP.2012.6467149.

Zhang, L., Zhang, D. & Mou, X. (2010, Sept). RFSIM: A feature based image quality assessment metric using Riesz transforms. *17th IEEE International Conference on Image Processing (ICIP)*, pp. 321-324. doi: 10.1109/ICIP.2010.5649275.

Zhang, L., Zhang, D., Mou, X. & Zhang, D. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8), 2378-2386. doi: 10.1109/TIP.2011.2109730.

Zhang, L., Shen, Y. & Li, H. (2014). VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, 23(10), 4270-4281. doi: 10.1109/TIP.2014.2346028.

Zhang, L., Zhang, L. & Bovik, A. (2015). A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing*, 24(8), 2579-2591. doi: 10.1109/TIP.2015.2426416.

Ziaei Nafchi, H., Shahkolaei, A., Farrahi Moghaddam, R. & Cheriet, M. (2015). FSITM: A Feature Similarity Index For Tone-Mapped Images. *IEEE Signal Processing Letters*, 22(8), 1026-1029. doi: 10.1109/LSP.2014.2381458.

Ziaei Nafchi, H. & Cheriet, M. (2016). A Note on Efficiency of Downsampling and Color Transformation in Image Quality Assessment. *CoRR*, abs/1606.06152.