Objective Quality Assessment of Tone-Mapped Images

Hojatollah Yeganeh, *Student Member, IEEE*, and Zhou Wang, *Member, IEEE*

***Abstract*— Tone-mapping operators (TMOs) that convert high dynamic range (HDR) to low dynamic range (LDR) images provide practically useful tools for the visualization of HDR images on standard LDR displays. Different TMOs create different tone-mapped images, and a natural question is which one has the best quality. Without an appropriate quality measure, different TMOs cannot be compared, and further improvement is directionless. Subjective rating may be a reliable evaluation method, but it is expensive and time consuming, and more importantly, is difficult to be embedded into optimization frameworks. Here we propose an objective quality assessment algorithm for tone-mapped images by combining: 1) a multiscale signal fidelity measure on the basis of a modified structural similarity index and 2) a naturalness measure on the basis of intensity statistics of natural images. Validations using independent subject-rated image databases show good correlations between subjective ranking score and the proposed tone-mapped image quality index (TMQI). Furthermore, we demonstrate the extended applications of TMQI using two examples—parameter tuning for TMOs and adaptive fusion of multiple tone-mapped images.[1]**

***Index Terms*— High dynamic range image, image fusion, image quality assessment, naturalness, perceptual image processing, structural similarity, tone mapping operator.**

I. Introduction

THERE has been a growing interest in recent years in high dynamic range (HDR) images, where the range of intensity levels could be on the order of 10,000 to 1 [1]. This allows for accurate representations of the luminance variations in real scenes, ranging from direct sunlight to faint starlight [1]. With recent advances in imaging and computer graphics technologies, HDR images are becoming more widely available. A common problem that is often encountered in practice is how to visualize HDR images on standard display devices that are designed to display low dynamic range (LDR) images. To overcome this problem, an increasing number of tone mapping operators (TMOs) that convert HDR to LDR images have been developed, for examples [2]–[5]. Because of the reduction in dynamic range, tone mapping procedures inevitably cause information loss. With multiple TMOs available, one would ask which TMO faithfully preserves the structural information in the original HDR images, and which TMO produces natural-looking realistic LDR images.

TMO assessment in the past mostly relied on human subjective evaluations. In [6], perceptual evaluations of 6 TMOs were conducted with regard to similarity and preferences. An overview and a subjective comparison of 8 TMOs were reported in [7]. HDR capable monitor was employed in [8] to compare 6 TMOs in a subjective experiment using a paired comparison method. In [9], 14 subjects were asked to rate 2 architectural interior scenes produced by 7 TMOs based on basic image attributes as well as the naturalness of the LDR images. A more comprehensive subjective evaluation was carried out in [10], where tone mapped images generated by 14 TMOs were shown to 2 groups of 10 human observers to rate LDR images, concerning overall quality, brightness, contrast, detail reproduction and color. In [11], subjects were asked to choose the best LDRs derived from 2 TMOs with different parameter settings to optimally tune the algorithms. The value of subjective testing cannot be overestimated. However, they have fundamental limitations. First, it is expensive and time consuming. Second, it is difficult to be incorporated into an optimization framework to automatically improve TMOs and adjust their parameter settings. Furthermore, important image structures contained in HDR images may be missing in tone mapped images, but human observers may not be aware of their existence. In this sense, subjective evaluation should not be regarded as a golden standard for the quality of tone mapped images.

Typical objective image quality assessment (IQA) approaches assume the reference and test images to have the same dynamic range [12], and thus cannot be directly applied to evaluate tone mapped images. Only a few objective assessment methods have been proposed for HDR images. The HDR visible differences predictor (HDR-VDP) [1], [13] is a human visual system (HVS) based fidelity metric that aims to distinguish between visible (suprathreshold) and invisible (subthreshold) distortions. The metric reflects the perception of distortions in terms of detection probability. Since HDR-VDP is designed to predict the visibility of differences between two HDR images of the same dynamic range, it is not applicable to compare an HDR image with an LDR image. A dynamic range independent approach was proposed in [14], which improves upon HDR-VDP and

Manuscript received August 22, 2011; revised September 7, 2012; accepted September 7, 2012. Date of publication October 2, 2012; date of current version January 10, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Eli Peli.

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: hyeganeh@uwaterloo.ca; zhouwang@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2221725

[1]A MATLAB code of the proposed algorithm is available online at https://ece.uwaterloo.ca/~z70wang/research/tmqi/. Partial preliminary results of this work were presented at International Conference on Image Analysis and Recognition, Burnaby, BC, Canada, June 2011.

produces three types of quality maps that indicate the loss of visible features, the amplification of invisible features, and reversal of contrast polarity, respectively. These quality maps show good correlations with subjective classifications of image degradation types including blur, sharpening, contrast reversal, and no distortion. However, it does not provide a single quality score for an entire image, making it impossible to be validated with subjective evaluations of overall image quality.

The purpose of the current work is to develop an objective IQA model for tone mapped LDR images using their corresponding HDR images as references. Our work is inspired by the success of two design principles in IQA literature. The first is the structural similarity (SSIM) approach [15] and its multi-scale derivations [16], [17], which asserts that the main purpose of vision is to extract structural information from the visual scene and thus structural fidelity is a good predictor of perceptual quality. The second is the natural scene statistics (NSS) approach, which maintains that the visual system is highly adapted to the natural visual environment and uses the departure from natural image statistics as a measure of perceptual quality [18]. Here we propose a method that combines a multi-scale structural fidelity measure and a statistical naturalness measure, leading to Tone Mapped image Quality Index (TMQI). Moreover, we demonstrate that TMQI can be employed for optimizing parameters in TMOs and for adaptively fusing multiple tone mapped images.

II. Quality Assessment Method

Due to the reduction in dynamic range, TMOs cannot preserve all information in HDR images, and human observers of the LDR versions of these images may not be aware of this. Therefore, structural fidelity plays an important role in assessing the quality of tone-mapped images [19]. On the other hand, structural fidelity alone does not suffice to provide an overall quality evaluation. A good quality tone mapped image should achieve a good compromise between structural fidelity preservation and statistical naturalness, which are sometimes competing factors.

A. Structural Fidelity

The SSIM approach provides a useful design philosophy as well as a practical method for measuring structural fidelities between images [20]. The original SSIM algorithm is applied locally and contains three comparison components − luminance, contrast and structure. Since TMOs are meant to change local intensity and contrast, direct comparisons of local and contrast are inappropriate. Let x and y be two local image patches extracted from the HDR and the tone-mapped LDR images, respectively. We define our local structural fidelity measure as

$$S_{\text{local}}(x, y) = \frac{2\sigma_x'\sigma_y' + C_1}{\sigma_x'^2 + \sigma_y'^2 + C_1} \cdot \frac{\sigma_{xy} + C_2}{\sigma_x\sigma_y + C_2} \tag{1}$$

where σ_x, σ_y and σ_{xy} are the local standard deviations and cross correlation between the two corresponding patches in HDR and LDR images, respectively, and C_1 and C_2 are positive stabilizing constants. Compared with the SSIM definition [15], the luminance comparison component is missing, and the structure comparison component (the second term in (1)) is exactly the same. The first term in (1) compares signal strength and is modified from that of the SSIM definition based on two intuitive considerations. First, the difference of signal strength between HDR and LDR image patches should not be penalized when their signal strengths are both significant (above visibility threshold) or both insignificant (below visibility threshold). Second, the algorithm should penalize the cases that the signal strength is significant in one of the image patches but insignificant in the other. This is different from the corresponding term in the original SSIM definition where any change in signal strength is penalized.

To distinguish between significant and insignificant signal strength, we pass the local standard deviation σ through a nonlinear mapping, which results in the σ' value employed in (1). The nonlinear mapping should be designed so that significant signal strength is mapped to 1 and insignificant signal strength to 0, with a smooth transition in-between. Therefore, the nonlinear mapping is related to the visual sensitivity of contrast, which has been an extensively studied subject in the literature of visual psychophysics [21]. Practically, the HVS does not have a fixed threshold of contrast detection, but typically follows a gradual increasing probability in observing contrast variations. Psychometric functions describing the detection probability of signal strength have been employed to model the data taken from psychophysical experiments. Generally, the psychometric function resembles a sigmoid shape [22], [23] and the sensory threshold is usually defined at the level of 50% of detection probability. A commonly adopted psychometric function is known as Galton's ogive [21], which takes the form of a cumulative normal distribution function given by

$$p(s) = \frac{1}{\sqrt{2\pi}\theta_s} \int_{-\infty}^{s} \exp\left[-\frac{(x-\tau_s)^2}{2\theta_s^2}\right] dx \tag{2}$$

where p is the detection probability density, s is the amplitude of the sinusoidal stimulus, τ_s is the modulation threshold, and θ_s is the standard deviation of the normal distribution that controls the slope of detection probability variation. It was found that the ratio

$$k = \frac{\tau_s}{\theta_s} \tag{3}$$

is roughly a constant, known as Crozier's law [21], [24]. Typical values of k ranges between 2.3 and 4, and $k = 3$ makes the probability of false alarm considerably small [21].

The reciprocal of the modulation threshold τ_s is often used to quantify visual contrast sensitivity, which is a function of spatial frequency, namely the contrast sensitivity function (CSF) [21]. A CSF formula that fits well with data collected in various psychological experiments is given by [25]

$$A(f) \approx 2.6[0.0192 + 0.114f]\exp[-(0.114f)^{1.1}] \tag{4}$$

where f denotes spatial frequency. This function is normalized to have peak value 1, and thus only provides relative sensitivity across the frequency spectrum. In practice, it needs to be scaled by a constant λ to fit psychological data. In our

implementation, we follow Kelly's CSF measurement [26]. Combining this with (4), we obtain

$$\tau_s(f) = \frac{1}{\lambda A(f)}. \tag{5}$$

This threshold value is calculated based on contrast sensitivity measurement assuming pure sinusoidal stimulus. To convert it to a signal strength threshold measured using the standard deviation of the signal, we need to take into account that signal amplitude scales with both contrast and mean signal inensity, and there is a $\sqrt{2}$ factor between the amplitude and standard deviation of a sinusoidal signal. As a result, a threshold value defined on signal standard deviation can be computed as

$$\tau_\sigma(f) = \frac{\overline{\mu}}{\sqrt{2}\lambda A(f)} \tag{6}$$

where $\overline{\mu}$ is the mean intensity value. Based on Crozier's law [21], [24], we have

$$\theta_\sigma(f) = \frac{\tau_\sigma(f)}{k}. \tag{7}$$

We can then define the mapping between σ and σ' as

$$\sigma' = \frac{1}{\sqrt{2\pi}\theta_\sigma}\int_{-\infty}^{\sigma}\exp\left[-\frac{(x-\tau_\sigma)^2}{2\theta_\sigma^2}\right]dx \tag{8}$$

In (1), σ_x' and σ_y' are the mapped versions of σ_x and σ_y, respectively. They are bounded between 0 and 1, where 0 and 1 represent completely insignificant and completely significant signal strengths, respectively.

The local structural fidelity measure S_{local} is applied to an image using a sliding window that runs across the image space. This results in a map that reflects the variation of structural fidelity across space. The visibility of image details depends on the sampling density of the image, the distance between the image and the observer, the resolution of the display, and the perceptual capability of the observer's visual system. A single-scale method cannot capture such variations. Following the idea used in multi-scale [16] and information-weighted SSIM [17], we adopt a multi-scale approach, where the images are iteratively low-pass filtered and downsampled to create an image pyramid structure [27], as illustrated in Fig. 1. The local structural fidelity map is generated at each scale. Fig. 2 shows two examples of such maps computed at multiple scales for the LDR images created from two different TMOs. It is interesting to observe these fidelity maps and examine how they correlate with perceived image fidelity. For example, the structural details of the brightest window regions are missing in Image (b), but are more visible in Image (a). For another example, there are detailed structures in the top-right dark regions that are not easily discerned in Image (a), but are better visualized in Image (b). All of these are clearly reflected in the structural fidelity maps.

At each scale, the map is pooled by averaging to provide a single score:

$$S_l = \frac{1}{N_l}\sum_{i=1}^{N_l} S_{\text{local}}(x_i, y_i) \tag{9}$$

where x_i and y_i are the i-th patches in the HDR and LDR images being compared, respectively, and N_l is the number of patches in the l-th scale. In the literature, advanced pooling strategies such as information content based pooling [17] have been shown to improve the performance of IQA algorithms. However, in our current experiment, these advanced pooling methods did not result in notable performance gain in the proposed structural fidelity measure. The overall structural fidelity is calculated by combining scale level structural fidelity scores using the method in [16]

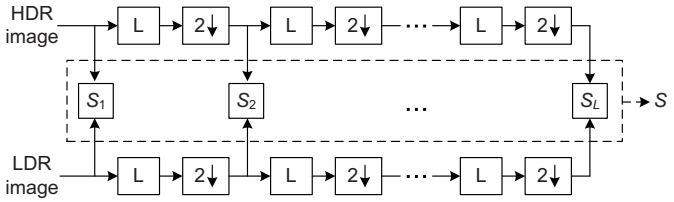


Fig. 1. Framework of multiscale structural fidelity assessment.

$$S = \prod_{l=1}^{L} S_l^{\beta_l} \tag{10}$$

where L is the total number of scales and β_l is the weight assigned to the l-th scale.

There are several parameters in the implementation of our structural fidelity model. First, when computing S_{local}, we set $C_1 = 0.01$ and $C_2 = 10$, and we find that the overall performance of the structural fidelity model is insensitive to these parameters within an order of magnitude. Second, to create the fidelity map at each scale, we adopt the same setting as in the SSIM algorithm [15] by employing a Gaussian sliding window of size 11×11 with standard deviation 1.5. Third, as in [16], we assume a viewing distance of 32 cycles/degree, which can represent signals up to 16 cycles/degree of resolution without aliasing, and thus we use 16 cycles/degree as the spatial frequency parameter when applying the CSF in (4) to the finest scale measurement. The spatial frequency parameters applied to the subsequent finer scales are then 8, 4, 2, 1 cycles/degree, respectively. Fourth, the mean intensity value in (6) is set to be the mean of the dynamic range of LDR images, i.e., $\overline{\mu} = 128$. Fifth, when combining the measures across scales, we set $L = 5$ and $\{\beta_l\} = \{0.0448, 0.2856, 0.3001, 0.2363, 0.1333\}$, which follows the psychophysical experiment results reported in [16]. Finally, in order to assess the quality of color images, we first convert them from RGB color space to Yxy space and then apply the proposed structural fidelity measure on the Y component only.

B. Statistical Naturalness

A high quality tone mapped LDR image should not only faithfully preserve the structural fidelity of the HDR image, but also look natural. Nevertheless, naturalness is a subjective quantity that is difficult to define quantitatively. A large literature has been dedicated to the statistics of natural images which have important significance to both image processing applications and the understanding of biological vision [28]. An interesting study of naturalness in the context of subjective evaluation of tone mapped images was carried out in [29],

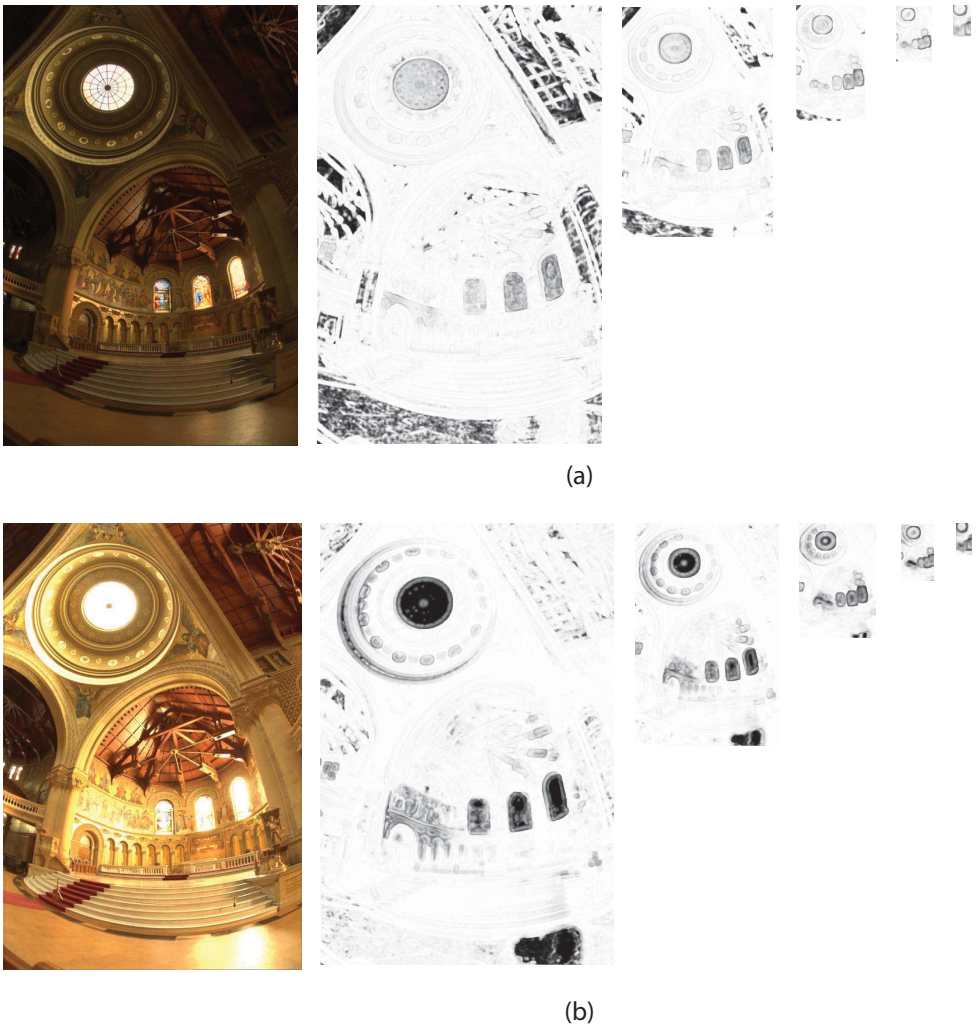


Fig. 2. Tone-mapped LDR images and their structural fidelity maps in five scales. The images were created using Adobe Photoshop "Highlight compression" and "Exposure and Gamma" methods (not optimized for quality), respectively. (a) $S = 0.9152$ ($S_1 = 0.8940$; $S_2 = 0.9341$; $S_3 = 0.9428$; $S_4 = 0.9143$; $S_5 = 0.8277$). (b) $S = 0.8614$ ($S_1 = 0.9161$; $S_2 = 0.9181$; $S_3 = 0.8958$; $S_4 = 0.8405$; $S_5 = 0.7041$).

which provided useful information regarding the correlations between image naturalness and different image attributes such as brightness, contrast, color reproduction, visibility and reproduction of details. The results showed that among all attributes being tested, brightness and contrast have more correlation with perceived naturalness. This motivates us to build our statistical naturalness model based on these two attributes. This choice may be oversimplifying in defining the general concept of statistical image naturalness (and may not generalize to other image processing applications that uses the concept of naturalness), but it provides an ideal compromise between the simplicity of our model and the capability of capturing the most important ingredients of naturalness that are related to the tone mapping evaluation problem we are trying to solve, where brightness mapping is an inevitable issue in all tone mapping operations. It also best complements the structural fidelity measure described in Section II-A, where brightness modeling and evaluation are missing.

Our statistical naturalness model is built upon statistics conducted on about 3,000 8bits/pixel gray-scale images obtained from [30], [31] that represent many different types of natural scenes. Fig. 3 shows the histograms of the means and standard deviations of these images, which are useful measures that reflect the global intensity and contrast of images. We found that these histograms can be well fitted using a Gaussian and a Beta probability density functions given by

$$P_m(m) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{m-\mu_m}{2\sigma_m^2}\right] \tag{11}$$

and

$$P_d(d) = \frac{(1-d)^{\beta_d - 1} d^{\alpha_d - 1}}{\mathrm{B}(\alpha_d, \beta_d)} \tag{12}$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function. The fitting curves are shown in Fig. 3, where the model parameters are estimated by regression, and the best values we found are $\mu_m = 115.94$ and $\sigma_m = 27.99$ in (11), and $\alpha_d = 4.4$ and $\beta_d = 10.1$ in (12), respectively.

Recent studies suggested that brightness and contrast are largely independent quantities in terms of both natural image statistics and biological computation [32]. As a result, their joint probability density function would be the product of the

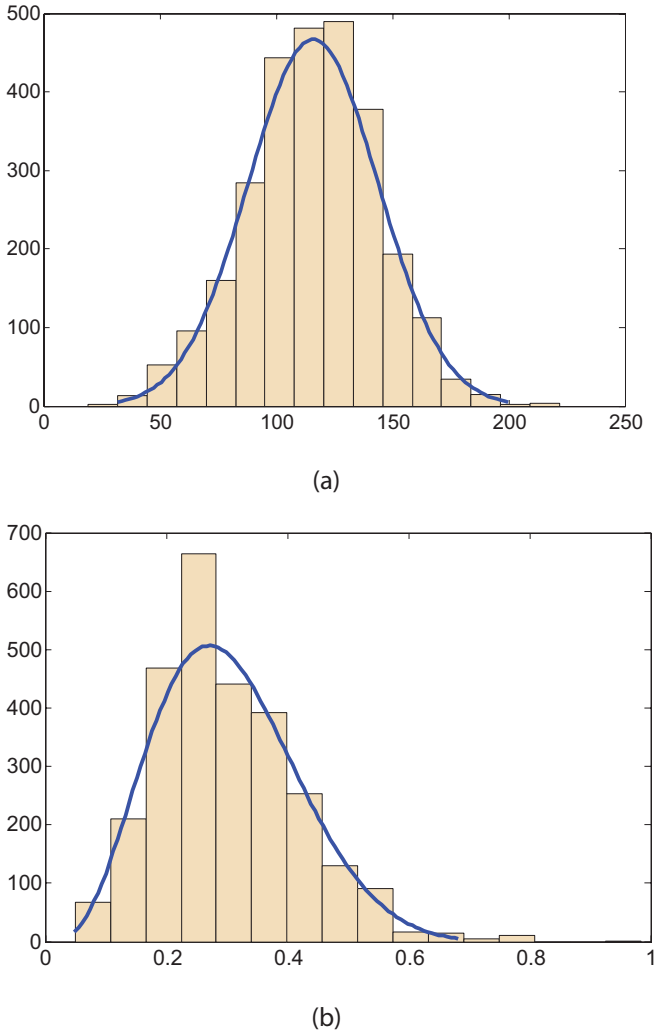


Fig. 3. Histograms of (a) means (fitted by Gaussian PDF) and (b) standard deviations (fitted by Beta PDF) of natural images.

two. Therefore, we define our statistical naturalness measure as

$$N = \frac{1}{K} P_m \, P_d \tag{13}$$

where K is a normalization factor given by $K = \max\{P_m \, P_d\}$. This constrains the statistical naturalness measure to be bounded between 0 and 1.

C. Quality Assessment Model

The structural fidelity measure S introduced in Section II-A and the statistical naturalness measure N described in Section II-B characterizes different aspects of the quality of tone mapped images. They may be used individually or jointly as a vector valued measure. In many practical applications, however, users prefer a single score that indicates the overall quality of the image. Therefore these parameters should be combined in some manner. In the literature of IQA, there had been earlier work that combines image statistics and measures of structure and contrast [33], though in a different context. Here we define a three-parameter function to scalarize the joint measure, resulting in a Tone Mapped image Quality Index (TMQI)

$$Q = aS^{\alpha} + (1 - a)N^{\beta} \tag{14}$$

where $0 \leq a \leq 1$ adjusts the relative importance of the two components, and α and β determine their sensitivities, respectively. Since both S and N are upper-bounded by 1, the overall quality measure is also upper-bounded by 1.

The parameters in (14) are left to be determined. In our implementation, they are tuned to best fit the subjective evaluation data provided by the authors of [34]. In their experiments, the subjects were instructed to look simultaneously at two LDR images created by two different TMOs applied upon the same HDR image, and then pick the one with better overall quality. Two studies have been done, involving two groups of subjects. The first study was carried out at Zheijang University, where 59 naive volunteers were invited to do the pair-wise comparison task and fill the preference matrix. The second study was conducted using Amazon Mechanical Turk, an online service of subjective evaluation. Each paired comparison was assigned to 150 anonymous subjects. The database includes 6 data sets, each of which contains images generated by 5 well-known TMOs, introduced by Drago *et. al.* [4], Durand & Dorsey [35], Fattal *et. al.* [5], Reinhard *et. al.* [2] and Mertens *et. al.* [36]. The subjective ranking scores in each folder are then computed using the preference matrix.

Finding the best parameters in (14) using subjective data is essentially a regression problem. The major difference from traditional regression problems is that here we are provided with relative ranking data between images only, but not quality scores associated with individual images. We developed a learning method where the parameters are learnt from an iterative method. At each iteration, one pair of images is randomly selected from one randomly selected data set. If the model generates objective scores that give the same order of the pair as the subjective rank order, then there is no change to the model parameters; Otherwise, each parameter is updated towards the direction of correcting the model error by a small step. The iteration continues until convergence. In our experiment, we observe good convergence property of this iterative learning process. Furthermore, to ensure the robustness of our approach, we conducted a leave-one-out cross validation procedure, where the database (of 6 data sets) was divided into 5 training sets and 1 testing set, and the same process was repeated 6 times, each with a different division between training and testing sets. Although each time ends up with a different set of parameters, they are fairly close to each other and result in the same ranking orders for all the training and testing sets. In the end, we select $a = 0.8012$, $\alpha = 0.3046$, and $\beta = 0.7088$ as our final model parameters.

III. Validation of Quality Assessment Method

The validation process is conducted by comparing our objective quality assessment results with subjective data. Two evaluation metrics are employed which are given as follows.

1) Spearman's rank-order correlation coefficient (SRCC) is defined as

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{15}$$

where d_i is the difference between the i-th image's ranks in subjective and objective evaluations. SRCC is a non-parametric rank-order based correlation metric, independent of any monotonic nonlinear mapping between subjective and objective scores.

2) Kendall's rank-order correlation coefficient (KRCC) is another non-parametric rank correlation metric computed as

$$\text{KRCC} = \frac{N_c - N_d}{\frac{1}{2}N(N-1)} \tag{16}$$

where N_c and N_d are the numbers of concordant (of consistent rank order) and discordant (of inconsistent rank order) pairs in the data set, respectively.

The proposed TMQI is the only objective quality measure being tested. To the best of our knowledge, almost no other method has been proposed to compare images with different dynamic ranges. The only exception is the method proposed in [14], which creates probability maps to distinguish between visible (suprathreshold) and invisible (subthreshold) degradations. The probability maps are shown to be useful in classifying image distortion types but are not meant to be pooled to produce an overall quality score of a tone mapped image. As a result, direct comparison with the proposed method is not possible.

Three experiments have been carried out in our validation process, each uses a different subject-ranked database. The first database is from [34], which was also used in the parameter training step discussed in Section II-C. Our leave-one-out cross validation method described in Section II-C creates SRCC and KRCC values for each of the six testing data sets, where for each data set, the parameters were trained using the other five data sets. Table I shows the means and standard deviations of KRCC and SRCC values between subjective rankings and our model predictions, respectively.

In the second experiment, we use the database introduced in [10], [37], from which we employ the overall quality ranking data by 10 naive subjects of 14 tone mapped images created from the same HDR image. The KRCC and SRCC values between subjective rankings of the images and our structural fidelity, statistical naturalness and overall quality scores are given in Table II, where we observe that the structural fidelity measure alone can provide reasonable predictions of subjective rankings. The statistical naturalness measure by itself is not a good predictor of the overall quality ranking, but it complements the structural fidelity measure. When the two measures are combined, better prediction of the overall image quality is achieved. It is worth mentioning that the test data here is not used in the training process, but the resulting KRCC and SRCC values are comparable with those obtained in the test using the first database, which is used for training. This implies good generalization capability of the training method described in Section II-C.

TABLE I
CROSS-VALIDATION RESULTS USING DATA FROM [34]

	KRCC	SRCC
Mean	0.7333	0.8333
Std	0.1632	0.1211

TABLE II
PERFORMANCE EVALUATION USING DATA FROM [10], [37]

	KRCC	SRCC
Structural Fidelity	0.6923	0.7912
Statistical Naturalness	0.3846	0.5385
Overall Quality	0.7179	0.8187

The third experiment is conducted using a database developed by ourselves. Twenty subjects were provided with 15 sets of tone mapped images, each of which includes 8 images generated by 8 TMOs from the same HDR image. The results created by five of the TMOs developed by Reinhard *et al.* [2], Drago *et. al.* [4], Durand & Dorsey [35], Mantiuk *et. al.* [38] and Pattanaik *et. al.* [39] are computed using the publicly available software Qtpfsgui [40]. In addition, three other images were created using the built-in TMOs in Adobe Photoshop, namely "Exposure and Gamma," "Equalize Histogram," and "Local Adaptation," respectively. The parameters used in all 8 TMOs are set as their default values and are not optimized. The reference HDR images are selected to represent different indoor and outdoor scenes and are all available online [10], [41]–[43]. In the subjective test, each of the 20 observers was asked to rank the 8 images in each image set from the best to the worst. The subjective rankings for each image is then averaged, resulting in its mean ranking score within the set.

To evaluate the TMQI method, we calculate the KRCC and SRCC values between the mean ranking scores and the objective quality measures for each image set. The results are given in Table III. To provide an anchor in evaluating the performance of TMQI, we compare it with the behavior of an average subject. To do this, we first compute the KRCC and SRCC values between the mean ranking scores and the ranking scores given by each individual subject for each image set. We then compute the mean and standard deviation of these KRCC and SRCC values over subjects, which are shown in Table III. The average KRCC and SRCC values over all 15 image set are given in the last row. It can be seen that for all image sets, the KRCC and SRCC values of TMQI are well within the range of ± 1 standard deviation from the KRCC and SRCC values of the mean over all subjects. This indicates that TMQI behaves quite similarly to an average subject.

Since the TMQI algorithm does not involve any expensive search or iterative procedure, it is computationally efficient. Our unoptimized MATLAB implementation on an Intel Quad-Core 2.67 GHz computer takes on average around 0.75 and 2.7 seconds to evaluate images of sizes 512×512 and 1024×1024, respectively. Fig. 4 illustrates the scatter plot of runtime versus the number of image pixels for 20 HDR-LDR comparisons. It shows that the computational complexity of the TMQI algorithm is approximately linear with respect to the number

TABLE III
PERFORMANCE EVALUATIONS USING 15 IMAGE SETS AND 8 TMOS

Image Set	KRCC			SRCC		
	Mean Subject Performance	STD of Subject Performance	TMQI Performance	Mean Subject Performance	STD of Subject Performance	TMQI Performance
1	0.8071	0.1038	0.7857	0.9071	0.0650	0.9048
2	0.7269	0.2072	0.6429	0.8251	0.1709	0.7857
3	0.7642	0.1064	0.6429	0.8797	0.0758	0.8095
4	0.8107	0.1141	0.7143	0.9130	0.0746	0.8571
5	0.4714	0.2116	0.6429	0.6000	0.2030	0.7381
6	0.6464	0.1646	0.7857	0.7630	0.1707	0.9048
7	0.7250	0.1275	0.5714	0.8285	0.1006	0.6905
8	0.7000	0.1862	0.5714	0.8023	0.1813	0.6905
9	0.6607	0.1978	0.5714	0.7857	0.1625	0.7619
10	0.8418	0.0991	0.7857	0.9276	0.0581	0.9048
11	0.7428	0.1815	0.7143	0.8523	0.1352	0.8810
12	0.6250	0.2084	0.5714	0.7595	0.2055	0.7143
13	0.5637	0.2298	0.5455	0.6970	0.2343	0.6587
14	0.6214	0.1720	0.6429	0.7702	0.1474	0.7381
15	0.8142	0.0994	0.7857	0.9035	0.0705	0.9048
Average	**0.7014**	**0.1606**	**0.6649**	**0.8143**	**0.1368**	**0.7963**

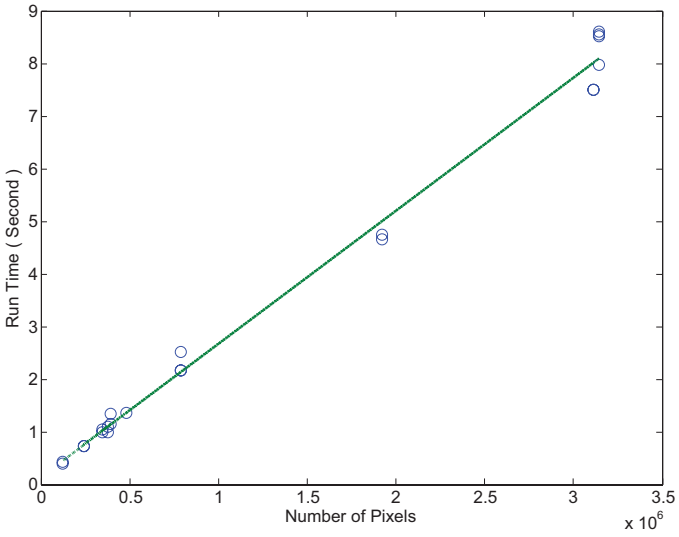


Fig. 4. Run time versus the number of image pixels of the proposed algorithm.

of pixels in the image. The relatively low computational cost makes it easily adapted to practical applications that involve iterative optimization processes.

IV. APPLICATIONS OF QUALITY ASSESSMENT METHOD

The application scope of objective IQA measures is beyond evaluating images and comparing algorithms. A wider range of applications extends to developing novel image processing algorithms optimized for the novel IQA measures. In this section, we use two examples to demonstrate the potentials of TMQI.

A. Parameter Tuning in TMO Algorithm

Many TMOs contain one or more parameters whose optimal values are often image-dependent. Without human interference, it is often a challenging task to choose these parameters, which could lead to drastically different results. An objective

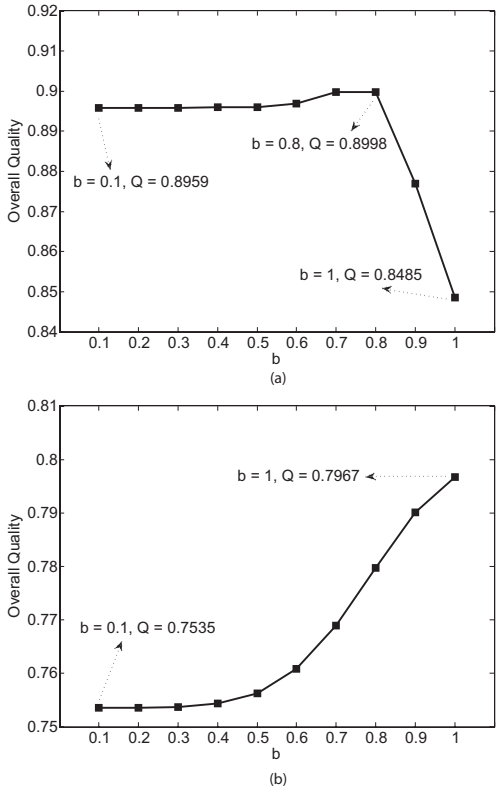


Fig. 5. Overall quality measure Q versus parameter b for (a) *Desk* and (b) *Bristol Bridge* images. The tone-mapped images corresponding to selected b values are shown in Figs. 6 and 7, respectively.

quality measure provides a useful tool to pick these parameters automatically. Here we use the TMO proposed in [4] as an example, which uses logarithmic function with varying bases in different locations to change the dynamic range adaptively. The algorithm is given by

$$L_{\mathrm{d}} = \frac{L_{\mathrm{dmax}} \cdot 0.01}{\log_{10}(L_{\mathrm{wmax}}+1)} \cdot \frac{\log(L_{\mathrm{w}}+1)}{\log\left(2+\left(\left(\frac{L_{\mathrm{w}}}{L_{\mathrm{wmax}}}\right)^{\frac{\log(b)}{\log(0.5)}}\right)\cdot 8\right)} \quad (17)$$

(a) (b) (c)

Fig. 6. LDR images generated with different parameter b in (17). (a) $b = 0.1$, $S = 0.8344$, $N = 0.4599$, and $Q = 0.8959$. (b) $b = 0.8$, $S = 0.8448$, $N = 0.4874$, and $Q = 0.8998$. (c) $b = 1.0$, $S = 0.8337$, $N = 0.1423$, and $Q = 0.8485$.

(a) (b) (c)

Fig. 7. LDR images generated with different parameter b in (17). (a) $b = 0.1$, $S = 0.5214$, $N = 0.0249$, and $Q = 0.7535$. (b) $b = 0.7$, $S = 0.8137$, $N = 0.1136$, and $Q = 0.7690$. (c) $b = 1.0$, $S = 0.8856$, $N = 0.2923$, and $Q = 0.7967$.

where L_{w} and L_{wmax} are world luminance and maximum luminance of the scene, L_{d} and L_{dmax} are display luminance and maximum luminance of display, respectively, and b is a tuning parameter. The perceptual quality of the tone mapped image varies significantly with b. However, in the literature, the b value is typically fixed around 0.8 through empirical experimenting with multiple images [4], [40].

In Figs. 5(a) and 5(b), we plot how TMQI varies as a function of b for images "Desk" and "Bristol Bridge," respectively (No computation beyond $b = 1$ is conducted because it is beyond the suggested value range by the algorithm). It appears that the quality score behaves quite differently as a function of b. Based on the plots, $b = 0.8$ and $b = 1$ are picked as the optimal values for the two images, respectively. These results confirm that the optimal b value is close to the empirical value (around 0.8) selected in previous studies, but varies for different images. The tone mapped LDR images corresponding to three selected b values are shown in Fig. 6 and Fig. 7, respectively. Careful inspection of these images shows that the best b values lead to good balance between preserving structural details and producing natural looking images.

B. Adaptive Fusion of Tone-Mapped Images

When experimenting with different TMOs on different HDR images, we often find it difficult to pick a single TMO that produces the best results for all HDR images. Furthermore, within a single HDR image, the best TMO may also vary when different regions in the image are under consideration. To take the advantages of multiple TMOs, image fusion techniques may be employed to combine multiple tone mapped images and an objective quality measure can play an important role in this process.

Given multiple tone mapped images created by different TMOs, we first apply a Laplacian pyramid transform that decomposes these images into different scales. In the pyramid domain, this results in multiple coefficients at the same scale and the same spatial location, each corresponds to a different TMO. Examples are given in the first two rows in Fig. 8, which demonstrate four-scale Laplacian pyramid decompositions, where the fine scale coefficients (Scales 1–3) represent image details and the coarsest scale coefficients (Scale 4) preserve local mean intensities across space. A fusion strategy can then be applied to combine multiple coefficients into one at each location in each scale before an inverse Laplacian pyramid transform is employed to reconstruct a fused image. Typical fusion schemes aim to locally select the most salient image features [44]. The most widely adopted approaches include averaging the coefficients or picking one of the coefficients with the largest absolute value.

Here we propose a different fusion scheme. The general idea is to use the TMQI as the weighting factors in the fusion process. Let S_j and c_j be the local structural fidelity measure and the Laplacian pyramid transform coefficient computed from the j-th tone mapped image being fused, respectively. The fused coefficient is computed as

$$c^{(fused)} = \frac{\sum_j S_j c_j}{\sum_j S_j}. \tag{18}$$

This is applied to all scales except for the coarsest scale, for which we use the statistical naturalness measure as the

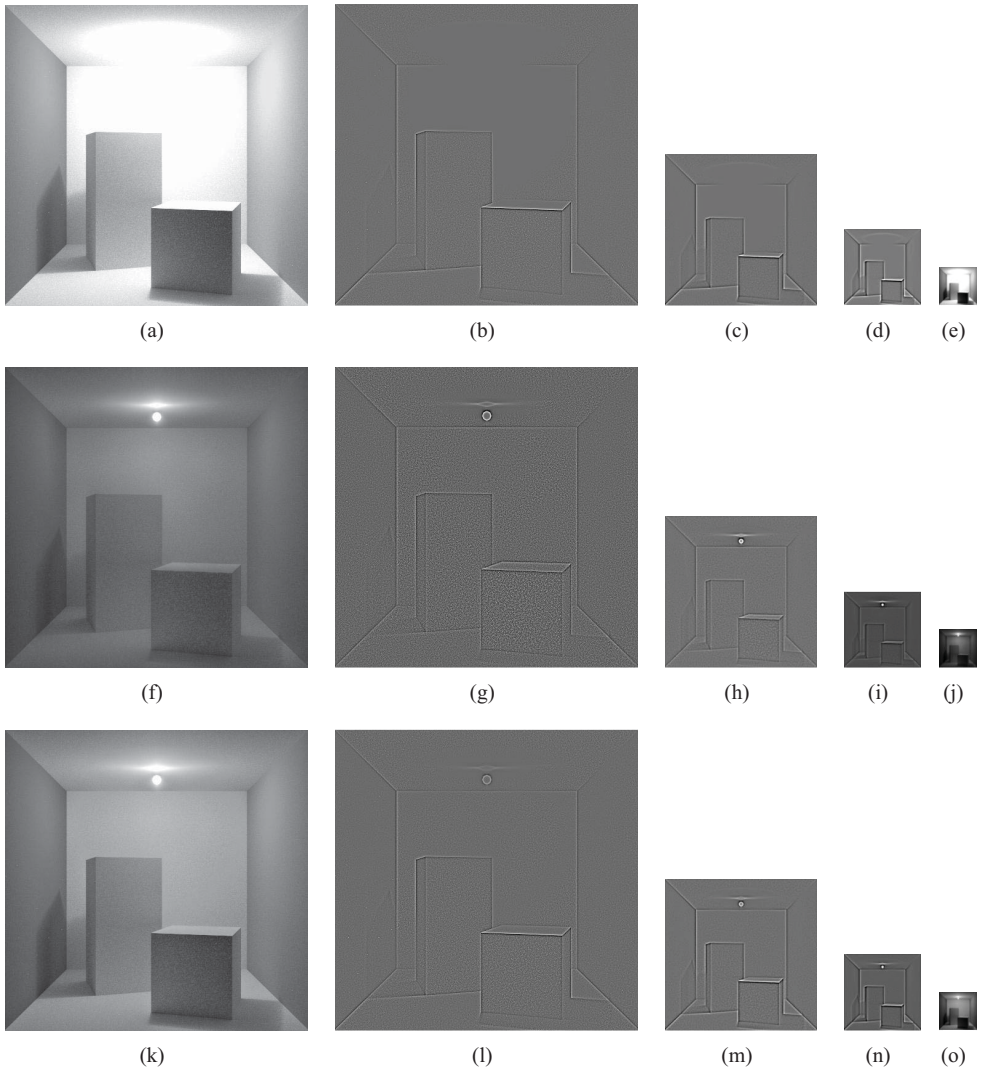


Fig. 8. Image fusion in Laplacian pyramid domain. Top row: first tone-mapped image (a) created by TMO proposed in [38], and its (b)–(e) Laplacian pyramid subbands, $S = 0.5034$, $N = 0.1263$, $Q = 0.6937$. Middle row: second tone-mapped image (f) using "Exposure and Gamma" method in Adobe Photoshop, and its (g)–(j) Laplacian pyramid subbands, $S = 0.6642$, $N = 0.0786$, and $Q = 0.7386$. Bottom row: fused image by (k) the proposed method, and its (l)–(o) Laplacian pyramid domain representation, $S = 0.7419$, $N = 0.3080$, and $Q = 0.8167$.

TABLE IV
AVERAGE RANKING SCORES MADE BY 10 SUBJECTS FOR EACH SET

Image Set	Source 1	Source 2	Fused Image
1	4.3	7	1.8
2	5.2	4	1.5
3	3.7	5.9	2.3
4	4.1	6.1	2.2
5	2.7	6.9	3

weighting factor:

$$c^{(fused)} = \frac{\sum_j N_j c_j}{\sum_j N_j} \tag{19}$$

where N_j denotes the statistical naturalness score of the j-th tone mapped image.

The proposed Laplacian pyramid domain fusion method is demonstrated in the bottom row of Fig. 8, where the fused image preserves the details in the brightest region (light area on the top) as in (f), while at the same time maintains higher contrast in relatively darker regions, as in (a). Fig. 9 provides an example with natural scene, where one tone mapped image (a) better preserves structural details, and another (b) gives more natural overall appearance (but loses structural information, especially at the brightest areas). Three fused images created by three different image fusion algorithms are given in (c), (d) and (e), respectively. The image created by the proposed method achieves the best balance between structure preserving and statistical naturalness, and also results in the best quality score using TMQI.

To further validate the proposed fusion scheme, we have conducted an additional subjective experiment, where ten subjects were invited to rank five sets of tone-mapped images, each of which includes eight images. Seven of these images are generated using the TMOs employed in the third experiment in Section III. Two of these seven TMOs are chosen to produce the eighth image using the proposed fusion method. Table IV compares average subjective rankings of the source images and their corresponding fused images, where lower ranking

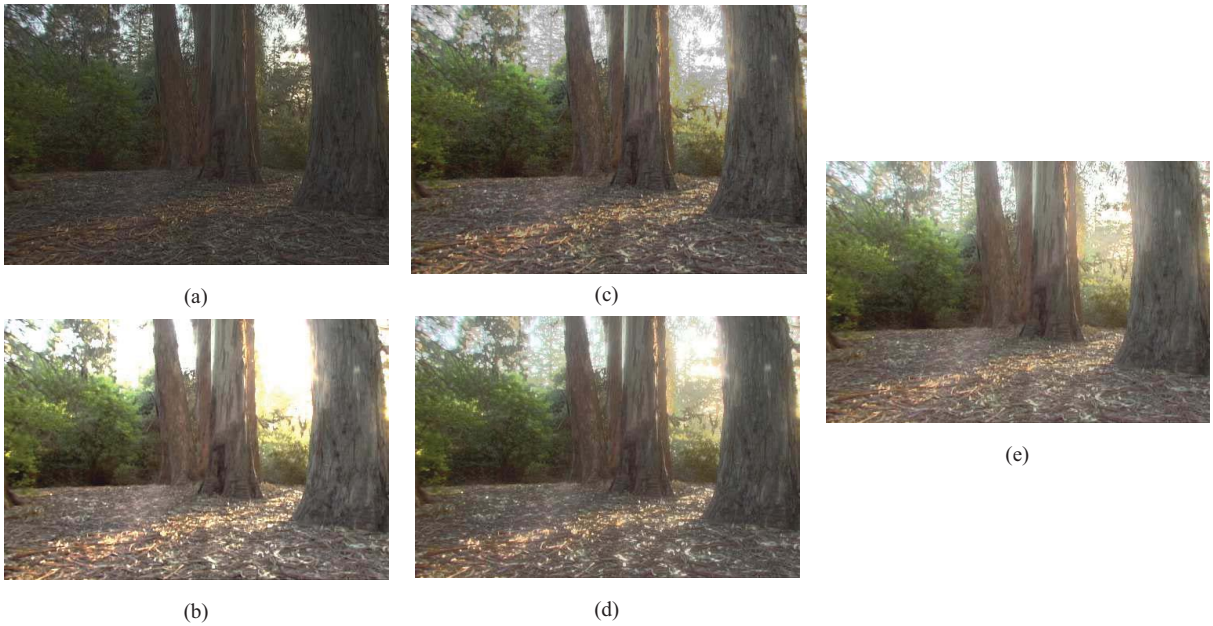


Fig. 9. Fusion of tone-mapped images. (a) First tone-mapped image using TMO proposed in [35], $S = 0.8168$, $N = 0.1631$, and $Q = 0.8075$. (b) Second tone-mapped image using the "Exposure and Gamma" method in Adobe Photoshop, $S = 0.6315$, $N = 0.8657$, and $Q = 0.8744$. (c) Fused image by coefficient averaging in Laplacian pyramid domain, $S = 0.7561$, $N = 0.7409$, and $Q = 0.8955$. (d) Fused image by selecting coefficient of maximal absolute value in Laplacian pyramid domain, $S = 0.7685$, $N = 0.9428$, and $Q = 0.9290$. (e) Fused image by the proposed method, $S = 0.7836$, $N = 0.9970$, and $Q = 0.9413$.

scores correspond to better quality. It can be seen that the fused image is almost always ranked significantly higher than the two source images being fused.

V. Conclusion

We develop an objective model to assess the quality of tone mapped images by combining a multi-scale structural fidelity measure and a statistical naturalness measure. The proposed measure not only provides an overall quality score of an image, but also creates multi-scale quality maps that reflect the structural fidelity variations across scale and space. Our experiments show that TMQI is reasonably correlated with subjective evaluations of image quality. Moreover, we demonstrate the usefulness of TMQI in automatic parameter tuning of tone mapping algorithms and in fusing multiple tone mapped images.

As one of the first attempts on the research topic, our method has several limitations that may be resolved or improved in the future. First, TMQI is designed to evaluate grayscale images only, but most HDR images of natural scenes are captured in color. One simple method to evaluate tone mapped color images is to apply the TMQI to each color channel independently and then combine them. Color fidelity and color naturalness measures may be developed to improve the quality measure.

Second, simple averaging is used in the current pooling method of the structural fidelity map. Advanced pooling method that incorporate visual attention models may be employed to improve the quality prediction performance.

Third, the current statistical naturalness measure is based on intensity statistics only. There is a rich literature on natural image statistics [28] and advanced statistical models (that reflects the structural regularities in space, scale and orientation in natural images) may be included to improve the statistical naturalness measure.

Fourth, using TMQI as a new optimization goal, many existing TMOs may be redesigned to achieve better image quality. Novel TMOs may also be developed by taking advantage of the construction of the proposed quality assessment approach.

Finally, the current method is applied and tested using natural images only. The application scope of HDR images and TMOs is beyond natural images. For example, modern medical imaging devices often capture HDR medical images that need to be tone-mapped before visualization. The TMQI and optimization methods may be adapted to these extended applications.

Acknowledgment

The authors would like to thank M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, for providing us with their subjective test data from Zhejiang University, Hangzhou, China, and Amazon Mechanical Turk.

References

[1] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. San Mateo, CA: Morgan Kaufmann, 2010.

[2] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Tech.*, vol. 21. 2002, pp. 267–276.

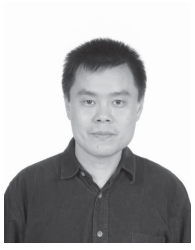
[3] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Trans. Visual. Comput. Graph.*, vol. 3, no. 4, pp. 291–306, Oct.–Dec. 1997.

[4] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 419–426, 2003.

[5] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Tech.*, 2002, pp. 249–256.
[6] F. Drago, W. L. Martens, K. Myszkowski, and H.-P. Seidel, "Perceptual evaluation of tone mapping operators," in *Proc. SIGGRAPH Conf. Sketches Appl.*, 2003, p. 1.
[7] A. J. Kuang, H. Yamaguchi, G. M. Johnson, and M. D. Fairchild, "Testing HDR image rendering algorithms," in *Proc. IS T/SID Color Imag. Conf.*, 2004, pp. 315–320.
[8] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 640–648, 2005.
[9] A. Yoshida, V. Blanz, K. Myszkowski, and H. Seidel, "Perceptual evaluation of tone mapping operators with real-world scenes," *Proc. SPIE, Human Vis. Electron. Imag.*, vol. 5666, pp. 192–203, Jan. 2005.
[10] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Image attributes and quality for evaluation of tone mapping operators," in *Proc. 14th Pacific Conf. Comput. Graph. Appl.*, 2006, pp. 35–44.
[11] M. Barkowsky and P. L. Callet, "On the perceptual similarity of realistic looking tone mapped high dynamic range images," in *Proc. Int. Conf. Image Process.*, 2010, pp. 3245–3248.
[12] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA: Morgan & Claypool Publishers, Mar. 2006.
[13] R. Mantiuk, S. Daly, K. Myszkowski, and S. Seidel, "Predicting visible differences in high dynamic range images—model and its calibration," *Proc. SPIE*, vol. 5666, pp. 204–214, Dec. 2005.
[14] T. O. Aydm, R. Mantiuk, K. Myszkowski, and H. Seidel, "Dynamic range independent image quality assessment," in *Proc. Int. Conf. Comput. Graph. Interact. Tech.*, 2008, pp. 1–69.
[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2003, pp. 1398–1402.
[17] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[18] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29–40, Nov. 2011.
[19] H. Yeganeh and Z. Wang, "Objective assessment of tone mapping algorithms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2477–2480.
[20] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.
[21] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Washington, DC: SPIE, 1999.
[22] J. P. Guilford, *Psychometric Methods*. New York: McGraw-Hill, 1954.
[23] Y. L. Grand, *Light, Colour and Vision*. London, U.K.: Chapman & Hall, 1968.
[24] W. J. Crozier, "On the variability of critical illumination for flicker fusion and intensity discrimination," *J. General Physiol.*, vol. 19, no. 3, pp. 503–522, 1935.
[25] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–536, Jul. 1974.
[26] D. H. Kelly, "Effects of sharp edges on the visibility of sinusoidal gratings," *J. Opt. Soc. Amer.*, vol. 60, no. 1, pp. 98–103, 1970.
[27] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
[28] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.
[29] M. Čadík and P. Slavík, "The naturalness of reproduced high dynamic range images," in *Proc. 9th Int. Conf. Inf. Visual.*, 2005, pp. 920–925.
[30] *Computer Vision Test Images*. (2005) [Online]. Available: http://www-2.cs.cmu.edu/afs/cs/project/cil/www/v-images.html
[31] *UCID - Uncompressed Colour Image Database*. (2004) [Online]. Available: http://www-staff.lboro.ac.uk/~cogs/datasets/UCID/ucid.html
[32] V. Mante, R. Frazor, V. Bonin, W. Geisler, and M. Carandini, "Independence of luminance and contrast in natural scenes and in the early visual system," *Nature Neurosci.*, vol. 8, no. 12, pp. 1690–1697, 2005.
[33] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, 2010.
[34] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 341–357, Jan. 2012.
[35] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, 2002.
[36] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *Proc. Pacific Conf. Comput. Graph. Appl.*, 2007, pp. 382–390.
[37] M. Cadik. (2005). *Evaluation of Tone Mapping Operators* [Online]. Available: http://www.cgg.cvut.cz/members/cadikm/tmo
[38] R. Mantiuk, K. Myszkowski, and H. Seidel, "A perceptual framework for contrast processing of high dynamic range images," in *Proc. 2nd Symp. Appl. Percept. Graph. Visual.*, 2005, pp. 87–94.
[39] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg, "Time-dependent visual adaptation for fast realistic image display," in *Proc. ACM SIGGRAPH Conf. Comput. Graph.*, 2000, pp. 47–54.
[40] *Open Source Community*. (2007) [Online]. Available: http://qtpfsgui.sourceforge.net/index.php
[41] *E. Reinhard's High Dynamic Range Data*. (2009) [Online]. Available: http://www.cs.utah.edu/~reinhard/cdrom/hdr/
[42] *G. Ward's High Dynamic Range Data* (2008) [Online]. Available: http://www.anyhere.com/gward/pixformat/tiffluvimg.html
[43] *P. Debevec's High Dynamic Range Data*. (2010) [Online]. Available: http://www.debevec.org/Research/HDR/
[44] R. S. Blum and Z. Liu, *Multi-Sensor Image Fusion and Its Applications*. New York: Taylor & Francis, 2006.

Hojatollah Yeganeh (S'10) received the B.S. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, and the M.Sc. degree in electronics engineering, specializing in speech processing and speech recognition, from the Amirkabir University of Technology, Tehran, Iran, in 2006 and 2009, respectively. He is currently pursuing the Ph.D. degree with the University of Waterloo, Waterloo, ON, Canada.

His current research interests include image processing, image quality assessment, and biomedical image processing.



Zhou Wang (S'97–A'01–M'02) received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin, Austin, in 2001.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He has authored or co-authored over 100 papers in journals and conferences with over 10 000 citations (Google Scholar). His current research interests include image processing, coding, and quality assessment, computational vision and pattern analysis, multimedia communications, and biomedical signal processing.

Dr. Wang was a recipient of the IEEE Signal Processing Society Best Paper Award in 2009, the ICIP 2008 IBM Best Student Paper Award (as senior author), and the Ontario Early Researcher Award in 2009. He has been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING since 2009 and *Pattern Recognition* since 2006, and a Guest Editor of *Signal, Image and Video Processing* since 2011. He was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2006 to 2010, and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING from 2007 to 2009 and the *EURASIP Journal of Image and Video Processing from* 2009 to 2010.