International Conference on Image Analysis and Recognition (ICIAR11), Burnaby, BC, Canada, June, 2011

Structural Fidelity vs. Naturalness - Objective Assessment of Tone Mapped Images

Hojatollah Yeganeh and Zhou Wang

Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1
hyeganeh@uwaterloo.ca, zhouwang@ieee.org

Abstract. There has been an increasing number of tone mapping algorithms developed in recent years that can convert high dynamic range (HDR) to low dynamic range (LDR) images, so that they can be visualized on standard displays. Nevertheless, good quality evaluation criteria of tone mapped images are still lacking, without which, different tone mapping algorithms cannot be compared and there is no meaningful direction for improvement. Although subjective assessment methods provide useful references, they are expensive and time-consuming, and are difficult to be embedded into optimization frameworks. In this paper, we propose a novel objective assessment method that combines a multi-scale signal fidelity measure inspired by the structural similarity (SSIM) index and a naturalness measure based on statistics on the brightness of natural images. Validations using available subjective data show good correlations between the proposed measure and subjective rankings of LDR images created by existing tone mapping operators.

Keywords: image quality assessment, high dynamic range image, tone mapping, structural similarity, naturalness of images

1 Introduction

The real world scenes exhibit a wide range of luminance variations. The dynamic range could be on the order of 10,000 to 1 from highlights to shadows [18]. High dynamic range (HDR) images allow us to capture greater luminance levels between its brightest and darkest regions than standard or low dynamic range (LDR) images. A common problem that is often encountered in practice is concerned about the visualization of HDR images − most display devices available to us have been designed to accommodate standard LDR images and cannot preserve all information contained in HDR images. In order to visualize HDR images using standard displays, a number of tone mapping algorithms have been proposed that convert HDR to LDR images, for example [15, 11, 8]. It should be noted that due to the dynamic range reduction, tone mapping operators (TMOs) unavoidably cause information loss. So the question is, having multiple TMOs a hand, which TMO faithfully maintains the information in the HDR image, and which TMO produces the most natural-looking good quality LDR image?

Subjective evaluation is the most straightforward method to assess the performance of TMOs. In [7], perceptual evaluations were carried out for six TMOs with regard to similarity and preferences. Seven TMOs were compared in [22] using two architectural interior scene and fourteen subjects were asked to rate basic image attributes as well as naturalness of the LDR images. A more comprehensive subjective experiment was performed in [6], where ten observer were asked to rate LDR images generated by 14 TMOs in terms of brightness, contrast, details and colors, and also to rank the overall quality of the images. These subjective test data are useful references in studying tone mapping algorithms. However, subjective experiments tend to be time-consuming and expensive. In addition, the outcome from these experiments are difficult to be incorporated into the design and optimization of tone mapping algorithms. Moreover, subjective tests may not be able to provide a complete evaluation because subject cannot see all details of HDR images, whose information may be missing from the LDR images and the subjects may not be aware of the existence of the missing details.

The progress on objective assessment of tone mapped images has been quite limited. Typical objective image quality assessment approaches assume that the reference and test images have the same dynamic range [18], and thus are not applicable. A dynamic range independent approach was proposed in [3], where the authors used a visibility model of the human visual system (HVS) to compare pairs of HDR-LDR images and produce quality maps, which reflect the loss of visible features, the amplification of invisible features, and reversal of contrast polarity. These quality maps show good correlations with subjective classifications of image degradation types including blur, sharpening, contrast reversal, and no distortion. However, this method does not provide a single quality score for an entire image, making it impossible to be validated with subjective evaluations of overall image quality.

In this work, we aims to develop an objective quality assessment model for LDR images using their corresponding HDR images as references. Our model is composed of two components − structural fidelity measurement and naturalness assessment. The structural fidelity measure is inspired by the success of the structural similarity (SSIM) index [18], which has been shown to be well correlated with perceived image quality when tested using a number of large-scale subject-rated independent databases [19]. Its performance can be further improved when incorporated into a multi-scale framework [20]. However, SSIM or multi-scale SSIM models cannot be directly applied to compare images with different dynamic ranges. Our method is built upon multi-scale SSIM but is adapted to accommodate contrast comparisons across dynamic ranges. The naturalness assessment component in our approach is based upon brightness statistics of natural images. Although the model is simple, it appears to be useful and especially suited to the problem we are working with, where brightness mapping is an inevitable issue in the design of tone mapping algorithms.

2 Proposed Method

The invisibility of HDR reference image casts big challenges to objective quality assessment of tone mapped images. Because of the reduction of dynamic range, TMOs are deemed not to be able to preserve all information in HDR images, and human observers may not be aware of this. One of the most important factors in assessing TMOs is that how much structural information is preserved after tone mapping. In [21], we presented a novel approach to measure the structural fidelity between HDR and its tone mapped LDR images based on the philosophy of SSIM. However, this does not suffice to provide an overall quality evaluation of tone mapped images because an LDR image that maintains the structural information of the HDR image may not look natural, for example, in our study we observed some LDR images that well maintain the structural information in the HDR images look overly dark. Therefore, we would desire tone mapped images that achieve the best balance between two (sometimes competing) factors – structural fidelity preservation and high naturalness. Our quality assessment model is thus built upon these ingredients.

2.1 Structural Fidelity

Local Structural Fidelity Assessment Our approach is derived from the philosophy behind the design of SSIM, which is based on the belief that the main purpose of human vision is to extract structural information from the visual scene, and thus perceived image distortion should be predictable by a measure of structural information loss. The original local SSIM definition includes a luminance, a contrast and a structure comparison components. Since the local luminance and contrast between HDR and LDR images are meant to be different, it does not make good sense to directly compare local luminance and contrast. Let x and y be two local image patches extracted from the HDR and LDR images respectively. Our local similarity measure is defined as

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

The second term is the structure comparison component as in SSIM, where σ_x, σ_y and σ_{xy} are the local standard deviations and cross correlation between the two patches in HDR and LDR images, respectively, and C_1 and C_2 are positive stabilizing constants. The modified local contrast comparison method is given in the first term, which is developed based on two considerations. First, the contrast difference between HDR and LDR image patches should not be penalized as long as their contrasts are both significant or both insignificant, as opposed to comparing images with the same dynamic range, where SSIM penalizes any change in contrast. Second, the algorithm should penalize the cases that the contrast is significant in one of the image patches, but insignificant in the other. The key issue here is to quantify the significance of local contrast. In order to do this, we pass the local standard deviation through a nonlinear mapping function given by

$$\sigma' = \begin{cases} 0, & \sigma < T_1 \\ \frac{1}{2}\left\{1 + \cos\left[\frac{\pi}{T_2 - T_1}(\sigma - T_2)\right]\right\}, & T_1 \leq \sigma \leq T_2 \\ 1, & T_2 < \sigma\,, \end{cases} \tag{2}$$

where T_1 and T_2 are two threshold values that define the ranges of insignificant and significant contrasts, and a raised cosine function is employed to provide a smooth transition between the two ranges. Note that when two image patches are both significant (σ greater than T_2) or both insignificant (σ smaller than T_1), the first term of Eq. (1) equals 1, and thus the S_{local} measure is fully determined by the structure comparison component in Eq. (1).

Multi-scale Assessment The local S_{local} measure described above is applied to an entire image using a sliding window approach across the image space, resulting in a quality map that indicates the quality variation across space.

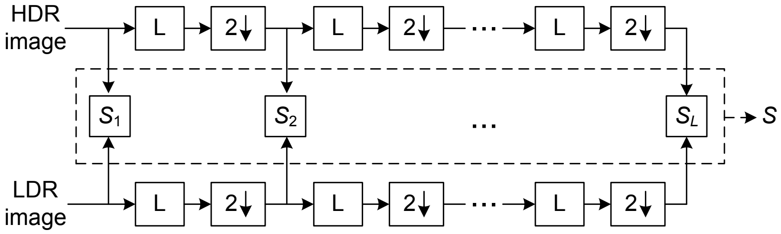


Fig. 1. Multi-scale framework of structural fidelity assessment method.

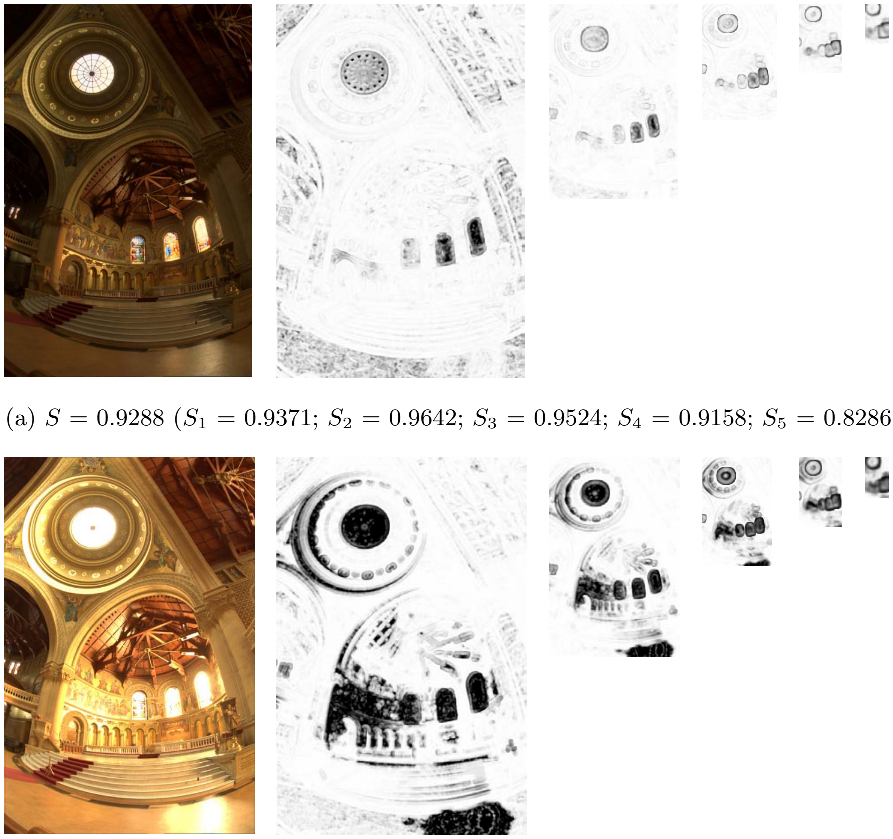
The perceivability of image details also depends on the sampling density of the image signal, the distance from the image to the observer, the display resolution, and the perceptual capability of the observer's visual system. In practice, the subjective evaluation of a given image varies with these parameters. A single-scale method as described in the previous section cannot capture such variations, and a multi-scale method is a convenient way to incorporate HVS features and image details at different resolutions. As in [20], we carry out signal fidelity assessment using a multi-scale structure depicted in Fig. 1, where the images are iteratively low-pass filtered and downsampled, creating an image pyramid structure [4]. The local structural fidelity map is generated at each scale, and the map is then averaged to provide a single score for the scale by

$$S_l = \frac{1}{N_l}\sum_{i=1}^{N_l} S_{\text{local}}(x_i, y_i)\,, \tag{3}$$

where x_i and y_i are the i-th patches in the two images being compared, and N_l is the number of patches in the l-th scale. Fig. 2 shows examples of quality maps computed using the proposed multi-scale approach. Finally, the structural fidelity measures computed at each scale are combined to a multi-scale measure of the overall structural fidelity:

$$S = \prod_{l=1}^{L} S_l^{\beta_l} , \tag{4}$$

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



(a) $S = 0.9288$ ($S_1 = 0.9371$; $S_2 = 0.9642$; $S_3 = 0.9524$; $S_4 = 0.9158$; $S_5 = 0.8286$)

(b) $S = 0.7980$ ($S_1 = 0.8419$; $S_2 = 0.8573$; $S_3 = 0.8330$; $S_4 = 0.7795$; $S_5 = 0.6361$)

Fig. 2. LDR images and their fidelity maps and scores in five scales. The images were created using Adobe Photoshop “Highlight compression” and “Exposure and Gamma” methods (not optimized for quality), respectively. The structural details of the brightest regions are missing in Image (b), but are more visible in Image (a). These are clearly reflected in the quality maps.

There are several parameters in the implementation of the multi-scale structural fidelity model. When computing S_{local}, we set $C_1 = 0.01$, $C_2 = 10$, $T1 = 0.5$, and $T_2 = 4$, respectively. In our test, we find that the overall performance of our quality model is insensitive to these parameters within an order of magnitude, though fine tunings are yet to be performed through carefully designed psychophysical experiment. To create the fidelity map at each scale, we employ a Gaussian sliding window of size 11×11 with standard deviation 1.5. 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 [20]. To assess the quality of color images we first convert them from RGB color space to Yxy space and we apply the proposed structural fidelity measurement on luminance component Y only.

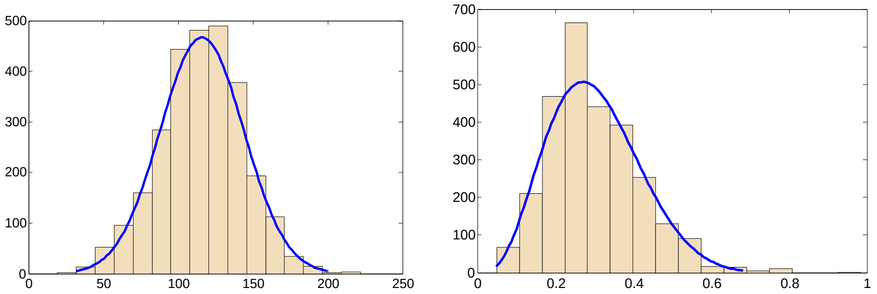
2.2 Naturalness

Tone mapping operators should be designed in a way that not only preserves structural information but also reproduces natural looking images. However, naturalness in general is a very subjective quantity and has not been clearly defined. A large literature has been dedicated to natural image statistics and their connections to biological vision. An excellent review can be found in [16]. Naturalness has also been studied in the context of subjective quality evaluation of tone mapped images. In [5], a subjective experiment was carried out and average correlation coefficients between image naturalness and different image attributes such as brightness, contrast, color reproduction, visibility and reproduction of details, are provided. The results show that among all attributes being tested, brightness and contrast have more correlation with perceived naturalness by subjects. This motivates us to build our naturalness model based on these two attributes. This choice may be oversimplifying in defining the general concept of image naturalness, but it captures 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.

Our method is built upon statistics of good-quality natural images. We gathered almost 3000 8bits/pixel natural images taken from many different scenes. These images are available at [1, 2]. Figure 3 shows the histograms of the means and standard deviations of these images, which are useful measures that reflect the global luminance and contrast of images. We find that these histograms can be well fitted using a Gaussian and a Beta probability density functions, respectively, where the model parameters can be found by regression. The fitting curves are also shown in Fig. 3. Since brightness and contrast can be considered independent quantities in terms of both natural image statistics and biological computation [13], their joint probability density function would be the product of the two. Therefore, we define our naturalness measure as

$$N = \frac{1}{K} P_p \, P_c \,, \tag{5}$$

where K is a normalization factor given by $K = \max\{P_p P_c\}$, such that the naturalness measure is bounded between 0 and 1.



(a)

(b)

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

2.3 Quality Assessment Model

Given a tone mapped LDR image, we now have two available measurements, structural fidelity S and naturalness N, which are given by Eq. (4) and Eq. (5), respectively. These two quantities can be used individually or jointly as a 2D vector that characterizes different aspects of the quality of the LDR image. However, in most applications, users would prefer to have a single quality score of the image. Therefore, an overall quality evaluation that combines both quantities is desirable. In particular, we define the following 3-parameter function to combine the two components

$$Q = aS^{\alpha} + (1-a)N^{\beta} \, , \tag{6}$$

where $0 \leq a \leq 1$ determines the relative weights assigned to the two components, and α and β defines the sensitivities of the two components, respectively. Since both S and N are upper-bounded by 1, this overall quality measure is also upper-bounded by 1. The parameters a, α and β, are left to be determined. In our implementation, they are tuned to best reflect subjective evaluations by utilizing machine learning techniques described next.

Machine Learning Process The parameters in Eq. (6) can be learned from subjective quality evaluation data of tone mapped images. We were provided with subjectively ranked databases from the authors of [17], where the subjects were instructed to look at two LDR images at a time (produced by two different TMOs) and then choose the one with better quality. Two groups of studies have been carried out with such paired comparison approach. The first group of comparisons was conducted at Zhejiang University. 59 naive volunteers were invited to make the paired comparisons and fill the preference matrix. The second comparison was carried out by using Amazon Mechanical Turk, which is an online service for subjective evaluations. Each comparison task was assigned to 150 anonymous subjects. The database includes 6 folders, each of which contains images generated by 5 well-known TMOs, namely adaptive logarithmic mapping [8], bilateral operator [9], uniform rational quantization [10], photoreceptor physiology [15] and exposure fusion [14]. The subjective ranking scores in each folder can then be computed using the preference matrix.

Finding the best parameters in Eq. (6) 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 an iterative method to learn the parameters. At each iteration, one pair of images is randomly selected from the database. If the model produce the correct order, then there is no change to the model parameters; Otherwise, each parameter is updated towards the direction of correcting the ranking error. To maintain the robustness of our approach, we carried out a cross validation process, where we divided the database into 6 folders and chose 5 as training set and the rest for testing. We repeat the same process 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 results. In the end, we fix $a = 0.8037$, $\alpha = 0.3958$ and $\beta = 0.8093$ as our final model parameters.

3 Validation

We used two independent subject-rated databases to test the proposed algorithm. The first is the database from [17] (which has also been used for training the parameters in Eq. (6)). We used leave-one-out cross-validation method described in the previous section to test our model. Table 1 shows the means and standard deviations of Kendall and Spearman rank order correlation coefficients between subjective rankings and our model predictions.

Table 1. Cross validation based on KRCC and SRCC using subjective data from [17]

	KRCC	SRCC
Mean	0.7333	0.8166
Std	0.2065	0.2136

The second database is from [6, 12], where we utilized the overall quality rankings by 10 naive subjects of 14 tone mapped images. KRCC and SRCC between subjective rankings and our structural fidelity, naturalness and overall quality scores are given in Table 2. Fig. 4 shows the scatter plots of the results, where rank numbers 1 and 14 correspond to the best and worst quality images, respectively. It can be observed that the overall quality score generally agrees quite well with subjective rankings and is significantly better than using structural fidelity or naturalness measures alone. It is worth mentioning that the KRCC and SRCC values are even higher than those obtained in the training database, implying good generalization ability.

Table 2. KRCC and SRCC evaluations based on subjective data from [6, 12]

	KRCC	SRCC
Structural Fidelity	0.6154	0.7967
Naturalness	0.4103	0.5606
Overall Quality	0.7692	0.8846

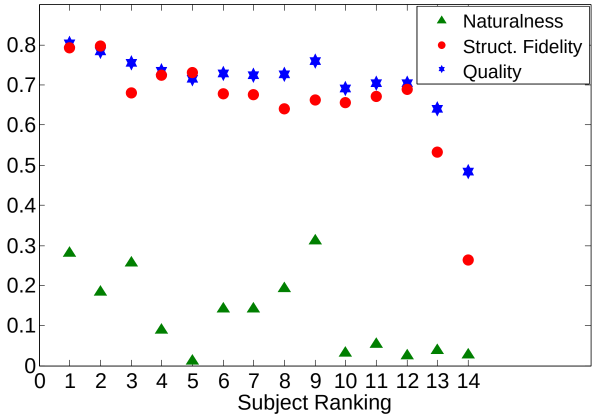


Fig. 4. Comparisons of subjective ranking versus structural fidelity, naturalness and overall quality scores using 14 tone mapped images from [6, 12].

4 Conclusion

In this paper, we proposed an objective method to assess the quality of LDR images created from HDR images by tone mapping algorithms. The proposed approach is based on the combination of two measures, structural fidelity and naturalness. The structural fidelity measure follows the framework of the multi-scale SSIM approach to assess the structural information maintained after tone mapping operations. The naturalness criterion is designed by comparing with luminance statistics taken from natural scenes. Our experiments demonstrate that the proposed measure correlates well with subjective rankings of overall image quality. The proposed algorithm is computationally efficient and provides not only an overall quality score, but also multi-scale fidelity maps that indicate local structural variations across scale and space. As one of the initial attempts in objective assessment of tone-mapped images, the proposed method is quite promising and shows good potentials in the evaluation, design and optimization of tone mapping algorithms.

Acknowledgment

We would like to express our gratitude to the authors of [17] for providing us with their subjective test data at Zhejiang University and Amazon Mechanical Turk. This research was supported in part by Natural Sciences and Engineering Research Council of Canada in the forms of Discovery, Strategic and CRD Grants, and by an Ontario Early Researcher Award, which are gratefully acknowledged.

References

1. http://www-2.cs.cmu.edu/afs/cs/project/cil/www/v-images.html.
2. http://www-staff.lboro.ac.uk/ cogs/datasets/UCID/ucid.html.
3. T. O. Aydm, R. Mantiuk, K. Myszkowski, and H. . Seidel. Dynamic range independent image quality assessment. In *SIGGRAPH'08: International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH*, 2008.
4. P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans. Communications*, 31:532–540, April 1983.
5. Martin Čadík and Pavel Slavík. The naturalness of reproduced high dynamic range images. In *IV '05: Proceedings of the Ninth International Conference on Information Visualisation*, pages 920–925, Washington, DC, USA, 2005. IEEE Computer Society.
6. Martin Čadík, Michael Wimmer, Laszlo Neumann, and Alessandro Artusi. Image attributes and quality for evaluation of tone mapping operators. In *Proceedings of the 14th Pacific Conference on Computer Graphics and Applications*, pages 35–44, Taipei, Taiwan, 2006. National Taiwan University Press.
7. F. Drago, W. L. Martens, K. Myszkowski, and Seidel H. P. Perceptual evaluation of tone mapping operators. *In Proc. Of the SIGGRAPH Conf. Sketches and Applications*, 2003.
8. F. Drago, K. Myszkowski, T. Annen, and N. Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum*, 22(3):419–426, 2003.
9. F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *ACM Transactions on Graphics*, volume 21, pages 257–266, 2002.
10. R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02*, pages 249–256, 2002.
11. G. Ward Larson, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3(4):291–306, 1997.

12. M. Cadik *et al.* Evaluation of tone mapping operators. http://www.cgg.cvut.cz/members/cadikm/tmo.
13. 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 Neuroscience*, 8(12):1690–1697, 2005.
14. T. Mertens, J. Kautz, and F. Van Reeth. Exposure fusion. In *Proceedings - Pacific Conference on Computer Graphics and Applications*, pages 382–390, 2007. Cited By (since 1996): 8.
15. E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. *in Proc. of 29th annual Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH*, 21:267–276, 2002.
16. E. P. Simoncelli and B. A. Olshausen. *Natural image statistics and neural representation*, volume 24 of *Annual Review of Neuroscience.* 2001.
17. M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang. Exposure fusion using a probabilistic model. *IEEE Transactions on Image Processing*, Submitted, 2011.
18. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Proc.*, 13:35–44, 2004.
19. Z. Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *To appear in IEEE Trans. Image Proc.*, 2011.
20. Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for iage quality assessment. *in Proc. of 37th Asilomar Conf. Signals, Systems and Computers*, 2003.
21. H. Yeganeh and Z. Wang. Objective assessment of tone mapping algorithms. In *Proc. IEEE Int. Conf. Image Proc.*, 2010.
22. A. Yoshida, V. Blanz, K. Myszkowski, and H. Seidel. Perceptual evaluation of tone mapping operators with real-world scenes. *Human Vision and Electronic Imagin X, SPIE*, 5666:192–203, 2005.