Presented at: IS&T/SPIE's 16th Annual Symposium on Electronic Imaging
San Jose, CA, Jan. 18-22, 2004
Published in: Human Vision and Electronic Imaging IX, Proc. SPIE, vol. 5292.

Stimulus Synthesis for Efficient Evaluation and Refinement of Perceptual Image Quality Metrics

Zhou Wang and Eero P. Simoncelli

Howard Hughes Medical Institute, Center for Neural Science and Courant Institute of Mathematical Sciences, New York University, New York, NY 10003, USA
zhouwang@ieee.org, eero.simoncelli@nyu.edu

ABSTRACT

We propose a methodology for comparing and refining perceptual image quality metrics based on synthetic images that are optimized to best differentiate two candidate quality metrics. We start from an initial distorted image and iteratively search for the best/worst images in terms of one metric while constraining the value of the other to remain fixed. We then repeat this, reversing the roles of the two metrics. Subjective test on the quality of pairs of these images generated at different initial distortion levels provides a strong indication of the relative strength and weaknesses of the metrics being compared. This methodology also provides an efficient way to further refine the definition of an image quality metric.

1. INTRODUCTION

Objective quality metrics are a fundamental ingredient in the formulation of a variety of image processing problems. In the literature, mean squared error (MSE) has been nearly exclusively employed as the standard metric for algorithm design, parameter optimization and system testing in most image processing applications. However, for applications in which the processed images are meant to be viewed by humans, the quality metric should provide a measure of perceptual image quality. Despite its simplicity and mathematical convenience for optimization purposes, MSE has long been criticized for poorly correlating with perceived image quality (e.g.[1]). A large number of methods for automatically predicting perceptual image quality have been proposed in recent years,[2–4] and extensive studies have been (and continue to be) conducted in order to evaluate and compare these methods (e.g.[2,5]).

The most standard form of evaluation is the direct one: the objective metric is compared to ratings by human subjects on an extensive database of images. Typically, the database contains a variety of reference images along with versions that have been distorted by different amounts. The total number of images in a database of reasonable size is typically in the order of hundreds. Gathering reliable data in such a large-scale subjective experiment is a very time-consuming and expensive task. In order to hold down the number of subjective comparisons that must be measured, the form of distortion is usually highly restricted. For example, the VQEG[2] and LIVE[5] databases include images distorted by lossy block-DCT and wavelet-based compression artifacts, as well as a small set of channel transmission errors. Thus, despite the substantial time involved in collecting psychophysical data in these experiments, there is no guarantee that the test results on these restricted databases provide a sufficient test for a "general-purpose" image quality metric. On the other hand, one cannot hope to simulate and test all possible image distortions.

The purpose of this study is to design subjective experiments that can efficiently evaluate the *relative* strength and weaknesses of different quality metrics with a relatively small number of subjective comparisons. The key idea is to conduct subjective tests on *synthesized* images that best differentiate two candidate quality metrics (rather than a large number of images with known types of distortions). In previous work, the idea of synthesizing images for subjective testing has been employed by the "synthesis-by-analysis" methods of assessing statistical texture models, in which the model is used to generate a texture with statistics matching an original texture, and a human subject then judges the similarity of the two textures.[6–10] These synthesis methods provide a very powerful and efficient means of testing a model (much more so than, for example, classification), and have the added benefit that the resulting images suggest improvements that might be made to the model.[10] In the context of image quality assessment, similar concept has also been used for qualitatively demonstrating the

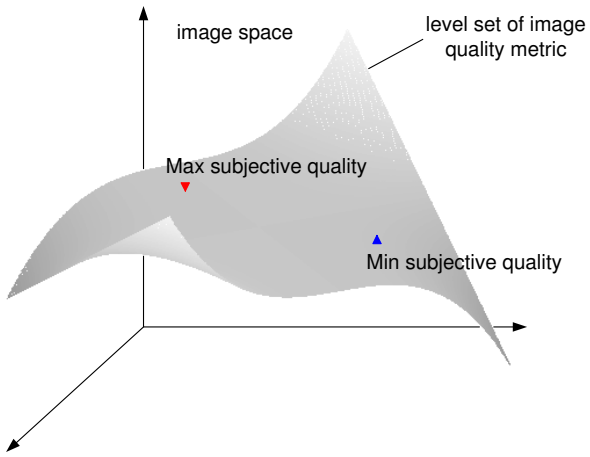

Figure 1. The level set of an image quality metric constitutes a manifold in the image space. Subjective comparison of the best and worst quality images is the strongest test on this manifold.

performance[11–13] and calibrating the parameter settings[14] of image quality metrics. In this paper, we attempt to formulate this approach and propose a systematic methodology that can be used for comparing image quality metrics and may potentially be generalized to a much broader range of psychophysical studies for performance evaluation of computational perceptual models.

2. METHOD

An image quality metric maps each image into a single quality number. The level sets of such a metric define a manifold in the space composed of all possible images, as illustrated in Fig. 1. A complete test of the metric on the level set would require subjective comparisons of all the images on the manifold to see if they have the same perceptual quality. Such a test is impossible due to the enormous number of comparisons needed. If one could somehow obtain the best and worst quality images on the manifold, a comparison of these two images would provide an immediate indication of perceptual adequacy of the metric. Specifically, if the best/worst images appear to be of very different visual quality, then the metric is failing to capture some important aspect of perceptual quality. This is conceptually the strongest test on the level set because the quality of all other images is in-between. However, this test is also not feasible because without seeing all the images on the manifold, one cannot find the best and worst subjective quality images. Instead, we can look for the best/worst case images on the manifold according to a *second* objective quality metric. This second metric is not meant to provide an equivalent replacement for perceptual distortion. Rather, its role is adversarial, in that it is used to probe potential weaknesses of the first metric, and possibly to suggest improvements.

The proposed performance evaluation system has two major components: stimulus synthesis and subjective test. Figure 2 illustrates the framework of the stimulus synthesis system. We denote the two image quality metrics as M_1 and M_2, respectively. Throughout this paper, we assume that all the image quality metrics being compared are full-reference metrics (meaning that a perfect quality reference image is always available for comparison), although the methodology may also be extended to the more general case, where the reference image is not available or only partially available. For each reference image and a given initial distortion level l, the system generates two pairs of synthesized image stimuli. First, the reference image is altered according to the given initial distortion level l (e.g., white noise of a specified variance, σ_l^2, is added) to generate an initial distorted image. Second, the quality of the initial image is calculated using the two given metrics M_1 and M_2, respectively. Third, the system searches for the maximum/minimum quality images in terms of M_2 while

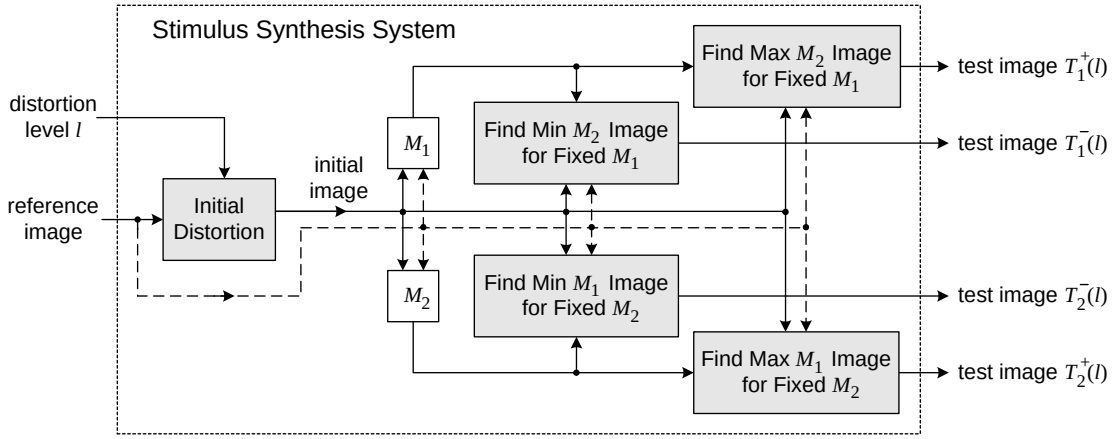

Figure 2. Framework of stimulus synthesis system.

constraining the value of M_1 to remain fixed. The result is a pair of images $T_1^+(l)$ and $T_1^-(l)$ that have the same M_1 value, but potentially very different M_2 values. This procedure is also applied with the roles of the two metrics reversed, to generate the maximum/minimum quality images in terms of M_1 while constraining the value of M_2 to be fixed. The second pair of synthesized images are denoted as $T_2^+(l)$ and $T_2^-(l)$, respectively. This image synthesize procedure is repeated for each of N reference images at L initial distortion levels, resulting in a total of $2NL$ pairs of synthesized image stimuli.

Depending on the specific formulation and complexity of the quality metrics being compared, there may be various methods of finding the maximum/minimum quality image in terms of one of the metrics while constraining the other to be fixed. In this paper, we use a constrained gradient ascent/descent iterative search method, which is illustrated in Fig. 3. In the image space, the initial distorted image is a common point of a level set of M_1 as well as a level set of M_2. As demonstrated in Fig. 3(a), starting from the initial image, we iteratively move along the direction of (or opposite to) the M_2 gradient vector after its component in the direction of the M_1 gradient has been projected out. The iteration continues until a maximum/minimum M_2 image is reached. Fig. 3(b) demonstrates the reverse procedure for searching the maximum/minimum M_1 images along the M_2 level set.

Figure 4 illustrates a single step of the constrained gradient ascent/descent algorithm for optimization of M_2. We represent images as column vectors, in which each entry represents the gray-scale value of one pixel. Denote the reference image $\mathbf{X}$ and the distorted image at the n-th iteration $\mathbf{Y}_n$ (with $\mathbf{Y}_0$ representing the initial distorted image). We compute the gradient of the two quality metrics, evaluated at $\mathbf{Y}_n$:

$$\mathbf{G}_{1,n} = \vec{\nabla}_{\mathbf{Y}} M_1(\mathbf{X}, \mathbf{Y})|_{\mathbf{Y}=\mathbf{Y}_n}\,, \qquad \mathbf{G}_{2,n} = \vec{\nabla}_{\mathbf{Y}} M_2(\mathbf{X}, \mathbf{Y})|_{\mathbf{Y}=\mathbf{Y}_n}\,.$$

We define a modified gradient direction, $\mathbf{G}_n$, by projecting out the component of $\mathbf{G}_{2,n}$ that lies in the direction of $\mathbf{G}_{1,n}$:

$$\mathbf{G}_n = \mathbf{G}_{2,n} - \frac{\mathbf{G}_{2,n}^T \mathbf{G}_{1,n}}{\mathbf{G}_{1,n}^T \mathbf{G}_{1,n}}\,\mathbf{G}_{1,n}\,.$$

A new distorted image is computed by moving in the direction of this vector:

$$\mathbf{Y}'_n = \mathbf{Y}_n + \lambda \mathbf{G}_n\,. \tag{1}$$

Finally, the gradient of M_1 is evaluated at $\mathbf{Y}'_n$, and an appropriate amount of this vector is added in order to guarantee that the new image has the correct value of M_1:

$$\mathbf{Y}_{n+1} = \mathbf{Y}'_n + \nu\,\mathbf{G}'_{1,n}\,, \quad \text{s.t.} \quad M_1(\mathbf{X}, \mathbf{Y}_{n+1}) = M_1(\mathbf{X}, \mathbf{Y}_0)\,. \tag{2}$$

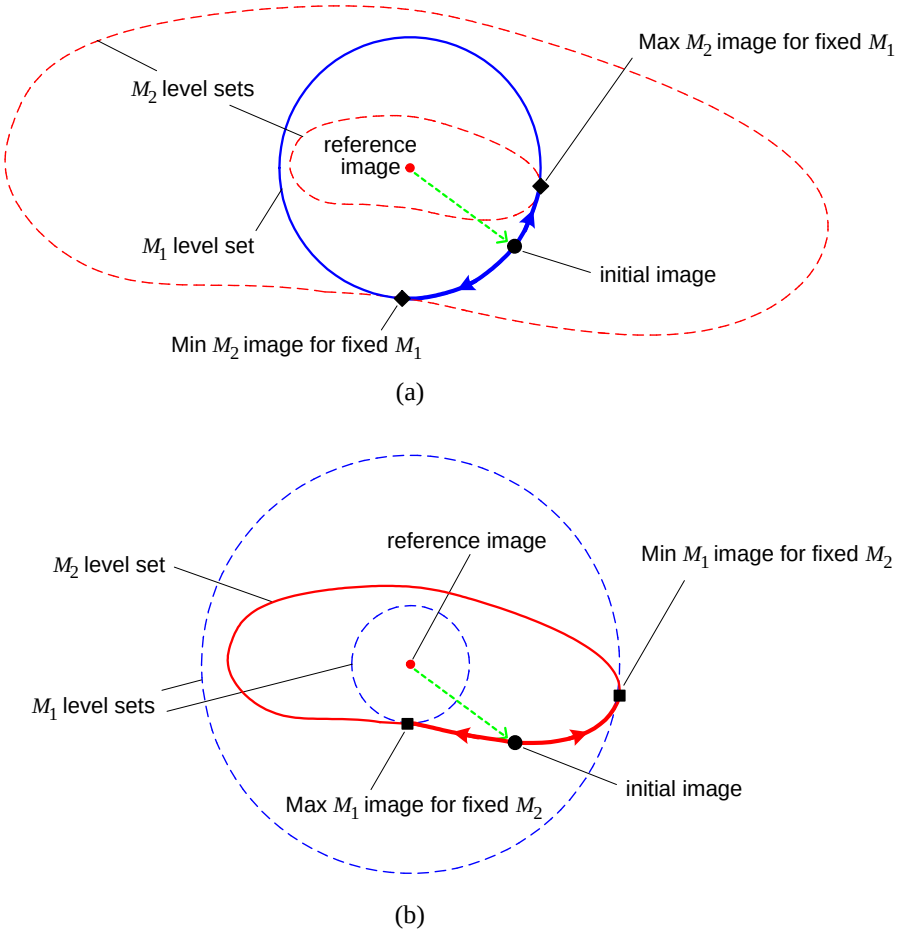

Figure 3. Illustration of constrained gradient ascent/descent iterative procedure for (a) searching the maximum/minimum M_2 image on M_1 level set, and (b) searching the maximum/minimum M_1 image on M_2 level set. This illustration is in 2-D space. In practice, the dimension equals the number of image pixels.

For the case of MSE, the selection of ν is straightforward, but in general it might require a one-dimensional (line) search.

During the iterations, the parameter λ is used to control the speed of convergence and ν must be adjusted dynamically so that the resulting vector does not deviate from the level set of M_1. The iteration continues until it satisfies certain convergence condition, e.g., mean squared change in the distorted image in two consecutive iterations is less than some threshold. If Metric M_2 is differentiable, then this procedure will converge to a local maximum/minimum of M_2. In general, however, to find the global maximum/minimum is difficult (note that the dimension of the search space is equal to the number of pixels in the image), unless the quality metric satisfies certain properties (e.g., convexity or concavity). In practice, there may be some additional constraints that need to be imposed during the iterations. For example, for 8bits/pixel gray-scale images, we may need to limit the pixel values to lie between 0 and 255.

The second major component of the proposed system is pair-wise subjective image quality comparison on the synthesized image stimuli. Subjective discrimination test on the quality of the image pairs $(T_1^+(l),\ T_1^-(l))$ and $(T_2^+(l),T_2^+(l))$ at different distortion level l (as illustrated in Fig. 2) provides a strong means of testing the performance of the quality metrics. For example, if M_1 is a good image quality metric, then the pairs of images with M_1 fixed $(T_1^+(l),\ T_1^-(l))$ should have similar perceptual quality. Furthermore, quantitative measure

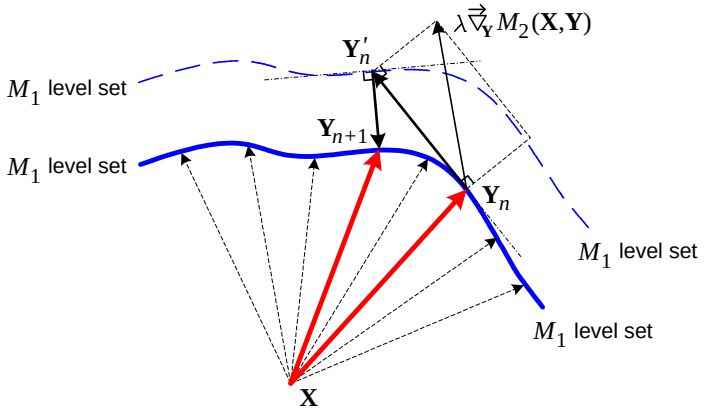

Figure 4. Illustration of the n-th iteration of the gradient ascent/descent search procedure for optimizing M_2 while constraining on the M_1 level set. This illustration is in 2-D space. In practice, the dimension equals the number of image pixels.

of the goodness of the quality metric can be obtained using 2-alternative forced choice (2AFC)[15,16] tests on the subjective discriminability of the quality of these image pairs.

3. DEMONSTRATION

In this section, we use MSE and the recently proposed Structural SIMilarity (SSIM) index approach[13] as two examples of image quality metrics to demonstrate the usage and effectiveness of the proposed performance comparison methodology.

3.1. Definitions of Image Quality Metrics

Given two images $\mathbf{X}$ and $\mathbf{Y}$, the MSE between them can be written in vector format as

$$\mathrm{MSE}(\mathbf{X},\mathbf{Y}) = \frac{1}{N_I}(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})\,, \tag{3}$$

where N_I is the number of pixels in the image.

The SSIM index[13] is usually applied over local image patches and then combined into a quality measure of the whole image. Let $\mathbf{x}$ and $\mathbf{y}$ be column vector representations of two image patches (e.g., 8×8 windows) extracted from the same spatial location from images $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let μ_x, σ_x^2 and σ_{xy} represent the mean of $\mathbf{x}$, the variance of $\mathbf{x}$, and the covariance of $\mathbf{x}$ and $\mathbf{y}$, respectively:

$$\begin{aligned}
\mu_x &= \frac{1}{N_P}(\mathbf{1}^T\cdot\mathbf{x})\,,\\
\sigma_x^2 &= \frac{1}{N_P-1}(\mathbf{x}-\mu_x)^T(\mathbf{x}-\mu_x)\,,\\
\sigma_{xy} &= \frac{1}{N_P-1}(\mathbf{x}-\mu_x)^T(\mathbf{y}-\mu_y)\,,
\end{aligned}$$

where N_P is the number of pixels in the local image patch and $\mathbf{1}$ is a vector with all entries equaling 1. The SSIM index between $\mathbf{x}$ and $\mathbf{y}$ is defined as[13]

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\,\mu_x\,\mu_y+C_1)\,(2\,\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)\,(\sigma_x^2+\sigma_y^2+C_2)}\,, \tag{4}$$

where C_1 and C_2 are small constants given by $C_1 = (K_1 R)^2$ and $C_2 = (K_2 R)^2$, respectively. Here, R is the dynamic range of the pixel values (e.g., $R = 255$ for 8 bits/pixel gray scale images), and $K_1 \ll 1$ and $K_2 \ll 1$ are two scalar constants (specifically, K_1=0.01 and K_2=0.03 in our experiments[13]). It can be shown that the SSIM index achieves its maximum value of 1 if and only if the two image patches $\mathbf{x}$ and $\mathbf{y}$ being compared are exactly the same.

The SSIM indexing algorithm is applied for image quality assessment using a sliding window approach. The window moves pixel-by-pixel across the whole image space. At each step, the SSIM index is calculated within the local window. This will result in an SSIM index map (or a quality map) over the image space. To avoid "blocking artifacts" in the SSIM index map, a smooth windowing approach can be used for local statistics.[13] However, for simplicity, we use an 8×8 square window in this paper. Finally, the SSIM index map is combined using a weighted average to yield an overall SSIM index of the whole image:

$$\mathrm{SSIM}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{N_S} W(\mathbf{x}_i, \mathbf{y}_i) \cdot \mathrm{SSIM}(\mathbf{x}_i, \mathbf{y}_i)}{\sum_{i=1}^{N_S} W(\mathbf{x}_i, \mathbf{y}_i)}, \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i-th sampling sliding windows in images $\mathbf{X}$ and $\mathbf{Y}$, respectively, $W(\mathbf{x}_i, \mathbf{y}_i)$ is the weight given to the i-th sampling window, and N_S is the total number of sampling windows. N_S is generally smaller than the number of image pixels N_I to avoid the sampling window exceed the boundaries of the image. The previous implementations of the SSIM measure[13, 14] corresponds to the case of uniform pooling, where $W(\mathbf{x}, \mathbf{y}) \equiv 1$.

3.2. Results

For given reference images, we create the initial distorted images by adding Gaussian white noise, where the variance σ_l^2 of the noise is determined by the initial distortion level l. Specifically, we let $\sigma_l^2 = 2^l$ for $l = 1, 2, ...,$ and 10, respectively. We then follow the iterative constrained gradient ascent/descent procedure described in Section 2 to search for the images with maximum/minimum SSIM for fixed MSE, as well as the images with minimum/maximum MSE for fixed SSIM. During the iterations, the gradients of MSE and SSIM with respect to the image are calculated explicitly using the methods described in Appendix.

Figure 5 shows the synthesized images for performance comparison of MSE and uniform-pooling SSIM at initial distortion level $l = 10$. Visual inspection of the images indicates that both metrics fail to capture some important aspects of perceptual image quality. In particular, the images in Group (c) and Group (d) have the same MSE with respect to their original images in Group (a). However, Group (c) images have very high quality, but Group (d) images poorly represent many important structures in the original images. On the other hand, although the corresponding images in Group (e) and Group (f) have the same SSIM values with respect to their original images, the distortions in some parts of Group (f) images are extremely annoying, leading to low overall image quality.

Perceptual evaluation of these images immediately reveals clear failures of both quality metrics. Specifically, MSE is blind in distinguishing between structural distortions (Group (d) images) and luminance/contrast changes (Group (c) images), while uniform pooling SSIM does not adequately penalize extreme local distortions (Group (f) images). Such observation also leads us to solutions for the improvement of image quality metrics. In particular, we could use a non-uniformly weighted method in the pooling of local SSIM measurement. As an attempt, here we consider a variance-weighted pooling method, where

$$W(\mathbf{x}, \mathbf{y}) = \sigma_x^2 + \sigma_y^2 + C_2 \,. \tag{6}$$

Figure 6 gives the synthesized images for performance comparison of MSE against SSIM with variance-weighted pooling (weights are given as of Eq. (6)). In this test, SSIM with variance-weighted pooling turns out to perform much better than MSE. On the one hand, for fixed MSE, the maximum SSIM images appear to have significantly better quality than minimum SSIM images at middle and high initial distortion levels. On the other hand, for fixed SSIM, the minimum MSE images are not obviously better than the maximum MSE images at any distortion level, and in some cases they are arguably worse. This comparison remains consistent across a variety of different image types, as is shown in Fig. 7.

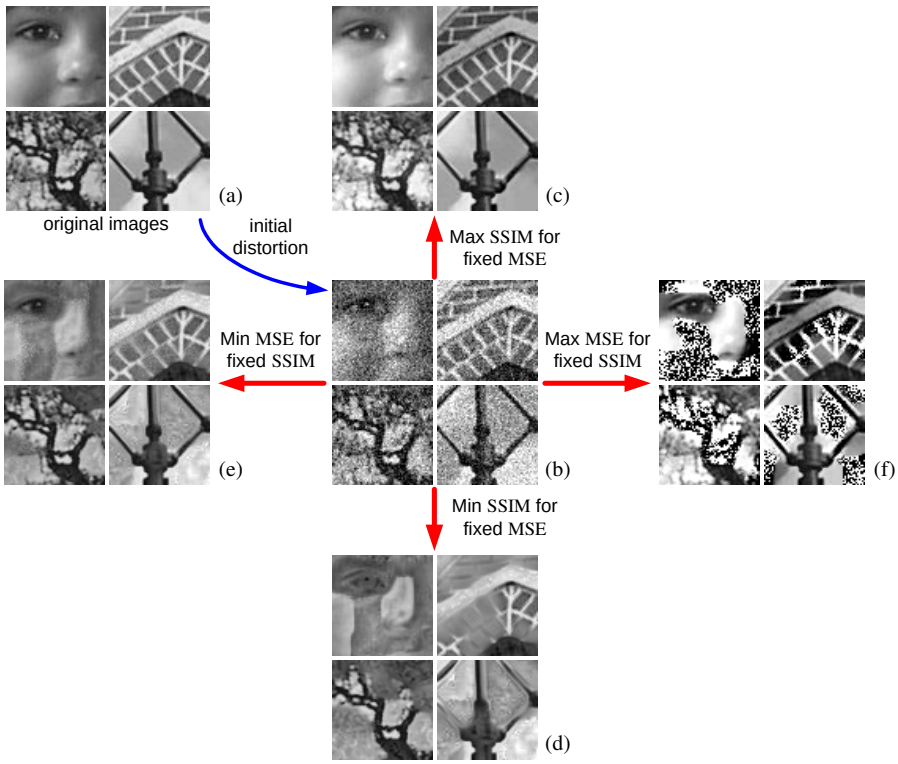

Figure 5. Synthesized images for performance comparisons of MSE and SSIM (uniform pooling: $W(\mathbf{x}, \mathbf{y}) \equiv 1$) at initial distortion level $l = 10$ ($\sigma_l^2 = 1024$). Both metrics fail to capture some important aspects of perceptual image quality.

4. DISCUSSION

We propose a new methodology for performance comparison of perceptual image quality metrics, in which the distorted test images are not predefined, but are optimally synthesized so as to best differentiate the metrics. A systematic way to create such image stimuli is presented and the algorithm is demonstrated using MSE and the SSIM image quality metrics. Subjective image quality comparison with these image stimuli proves to be a very strong and efficient test, and directly reveals the perceptual implications of the quality metrics. It also provides an intuitive and effective way to discover the weaknesses of image quality metrics. This can help us improve a quality metric, and can potentially suggest a means of combining the advantages of a number of weak quality metrics in order to create a much stronger one.

Although the algorithm proposed in this paper is described and demonstrated in the context of general-purpose full-reference image quality assessment, the methodology can be easily adapted to a much broader range of applications. For example, in the scenario of no-reference image quality assessment, the overall stimulus synthesis system shown in Fig. 2 can be applied with minimal modification. Our method may also be restricted to specific-purpose image quality assessment (e.g., quality assessment of compressed images), in which case the search space (level sets in Figs. 3 and 4) for generating the image stimuli is reduced, or equivalently, additional constraints are needed during the iteration procedures described in Section 2. The methodology may also be extended beyond the scope of image quality assessment for performance comparison of competing computational models of a perceptual discrimination quantity.

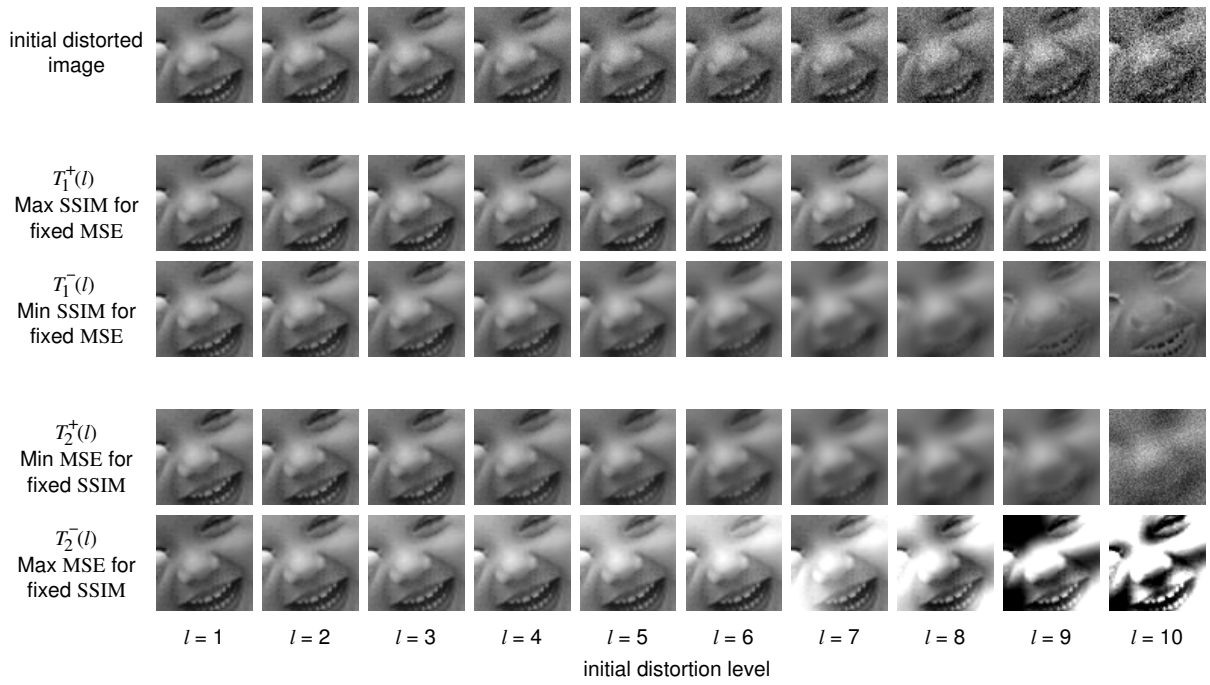

Figure 6. Synthesized images at different initial distortion levels for performance comparisons of MSE and SSIM (variance-weighted pooling as of Eq. (6)). Subjective comparisons are conducted on image pairs ($T_1^+(l)$, $T_1^-(l)$) and ($T_2^+(l)$,$T_2^+(l)$) at different initial distortion levels $l = 1, 2, ...10$.

APPENDIX

In order to use the constrained gradient ascent (descent) algorithm described in Section 2, we need to calculate the gradients of the quality metrics with respect to the distorted image. Here the gradients are represented as column vectors that have the same dimension as the images.

For MSE, it can be easily shown that

$$\vec{\nabla}_{\mathbf{Y}}\mathrm{MSE}(\mathbf{X},\mathbf{Y}) = -\frac{2}{N_I}(\mathbf{X}-\mathbf{Y})\,. \tag{7}$$

For SSIM, taking the derivative of Eq. (5) with respect to $\mathbf{Y}$, we have

$$\begin{aligned}&\vec{\nabla}_{\mathbf{Y}}\mathrm{SSIM}(\mathbf{X},\mathbf{Y})\\ &= \frac{\vec{\nabla}_{\mathbf{Y}}\left[\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\,\mathrm{SSIM}(\mathbf{x}_i,\mathbf{y}_i)\right]\cdot\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i) - \sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\,\mathrm{SSIM}(\mathbf{x}_i,\mathbf{y}_i)\cdot\vec{\nabla}_{\mathbf{Y}}\left[\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\right]}{\left[\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\right]^2}\end{aligned} \tag{8}$$

where

$$\vec{\nabla}_{\mathbf{Y}}\left[\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\right] = \sum_{i=1}^{N_S}\vec{\nabla}_{\mathbf{Y}} W(\mathbf{x}_i,\mathbf{y}_i) = \sum_{i=1}^{N_S}\vec{\nabla}_{\mathbf{y}} W(\mathbf{x},\mathbf{y})|_{\mathbf{x}=\mathbf{x}_i,\mathbf{y}=\mathbf{y}_i} \tag{9}$$

and

$$\vec{\nabla}_{\mathbf{Y}}\left[\sum_{i=1}^{N_S} W(\mathbf{x}_i,\mathbf{y}_i)\,\mathrm{SSIM}(\mathbf{x}_i,\mathbf{y}_i)\right] = \sum_{i=1}^{N_S}\vec{\nabla}_{\mathbf{Y}}\left[W(\mathbf{x}_i,\mathbf{y}_i)\,\mathrm{SSIM}(\mathbf{x}_i,\mathbf{y}_i)\right]$$

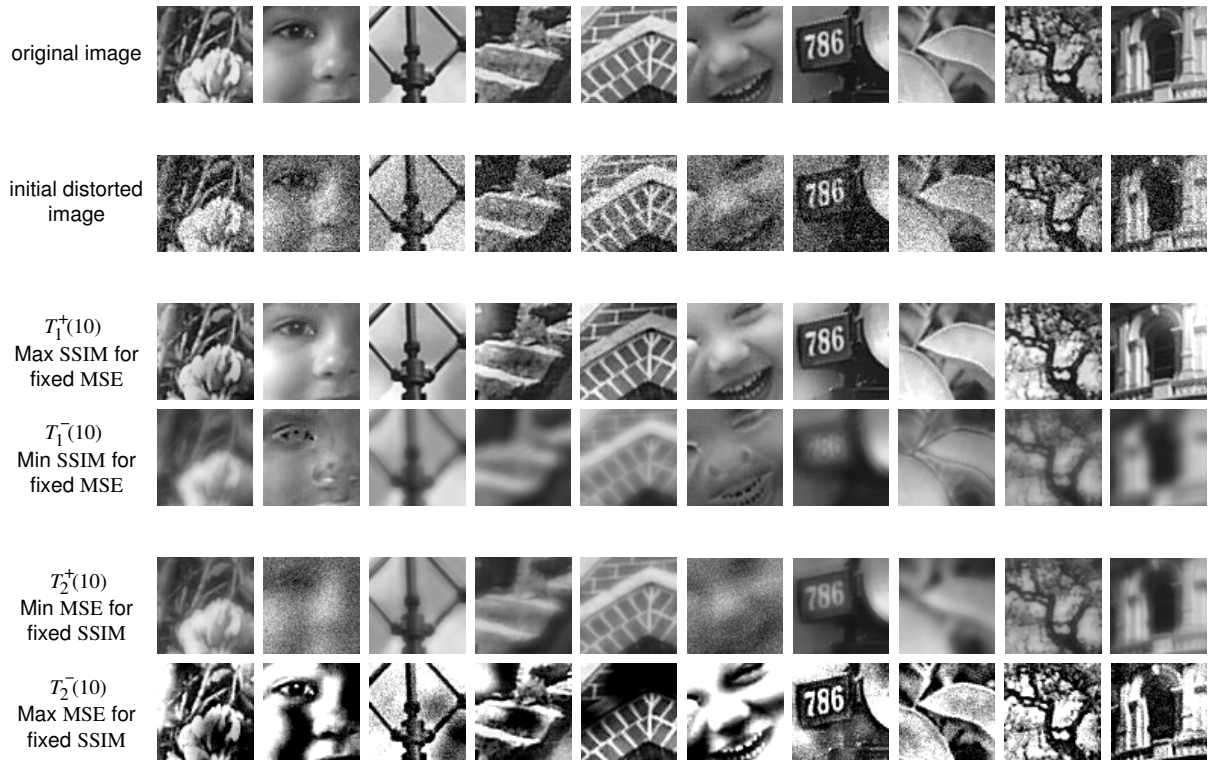

Figure 7. Synthesized images at initial distortion level $l = 10$ for performance comparisons of MSE and SSIM with variance-weighted pooling.

$$= \sum_{i=1}^{N_S} \left[W(\mathbf{x}_i, \mathbf{y}_i) \cdot \vec{\bigtriangledown}_{\mathbf{y}} \mathrm{SSIM}(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{x}_i, \mathbf{y}=\mathbf{y}_i} \right] + \sum_{i=1}^{N_S} \left[\mathrm{SSIM}(\mathbf{x}_i, \mathbf{y}_i) \cdot \vec{\bigtriangledown}_{\mathbf{y}} W(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{x}_i, \mathbf{y}=\mathbf{y}_i} \right] . \tag{10}$$

Thus, the gradient calculation of an entire image is converted into weighted summations of the local gradient measurements. For a local SSIM measure as of Eq. (4), we define

$$\begin{aligned} A_1 &= 2\,\mu_x\,\mu_y + C_1\,, \quad A_2 = 2\,\sigma_{xy} + C_2\,, \\ B_1 &= \mu_x^2 + \mu_y^2 + C_1\,, \quad B_2 = \sigma_x^2 + \sigma_y^2 + C_2\,. \end{aligned}$$

Then it can be shown that

$$\vec{\bigtriangledown}_{\mathbf{y}} \mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{2}{N_P B_1^2 B_2^2} \left[A_1 B_1 \left(B_2\,\mathbf{x} - A_2\,\mathbf{y} \right) + B_1 B_2 \left(A_2 - A_1 \right) \mu_x + A_1 A_2 \left(B_1 - B_2 \right) \mu_y \right] . \tag{11}$$

For the case that $W(\mathbf{x}, \mathbf{y}) \equiv 1$, we have $\vec{\bigtriangledown}_{\mathbf{y}} W(\mathbf{x}, \mathbf{y}) \equiv 0$ in Eq. (9). Therefore,

$$\vec{\bigtriangledown}_{\mathbf{Y}} \mathrm{SSIM}(\mathbf{X}, \mathbf{Y}) = \frac{1}{N_S} \sum_{i=1}^{N_S} \vec{\bigtriangledown}_{\mathbf{y}} \mathrm{SSIM}(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\mathbf{x}_i, \mathbf{y}=\mathbf{y}_i} \,. \tag{12}$$

In the case of variance-weighted average as of Eq. (6), we have

$$\vec{\bigtriangledown}_{\mathbf{y}} W(\mathbf{x}, \mathbf{y}) = 2(\mathbf{y} - \mu_y) \,. \tag{13}$$

Thus, $\vec{\bigtriangledown}_{\mathbf{Y}} \mathrm{SSIM}(\mathbf{X}, \mathbf{Y})$ can be calculated by combining Eqs. (8), (9), (10), (11) and (13).

REFERENCES

1. Z. Wang, "Demo images and free software for 'A Universal Image Quality Index'," `http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html`.
2. VQEG: The Video Quality Experts Group, `http://www.vqeg.org/`.
3. T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.*, A. Bovik, ed., Academic Press, 2000.
4. Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, eds., pp. 1041–1078, CRC Press, Sept. 2003.
5. H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," `http://live.ece.utexas.edu/research/quality/`.
6. O. D. Faugeras and W. K. Pratt, "Decorrelation methods of texture feature extraction," *IEEE Pat. Anal. Mach. Intell.* **2**(4), pp. 323–332, 1980.
7. A. Gagalowicz, "A new method for texture fields synthesis: Some applications to the study of human vision," *IEEE Pat. Anal. Mach. Intell.* **3**(5), pp. 520–533, 1981.
8. D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. ACM SIGGRAPH*, pp. 229–238, Association for Computing Machinery, August 1995.
9. S. Zhu and D. Mumford, "Prior learning and Gibbs reaction-diffusion," *IEEE Pat. Anal. Mach. Intell.* **19**(11), 1997.
10. J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision* **40**, pp. 49–71, December 2000.
11. P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, **2179**, pp. 127–141, 1994.
12. Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters* **9**, pp. 81–84, Mar. 2002.
13. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Image Processing* **13**, Jan. 2004.
14. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, (Asilomar), Nov. 2003.
15. N. Graham, J. G. Robson, and J. Nachmias, "Grating summation in fovea and periphery," *Vision Research* **18**, pp. 815–825, 1978.
16. B. A. Wandell, *Foundations of Vision*, Sinauer Associates, Inc., 1995.