Journal of Electronic Imaging 19(1), 011006 (Jan–Mar 2010)

Most apparent distortion: full-reference image quality assessment and the role of strategy

Eric C. Larson
University of Washington
Department of Electrical Engineering
Seattle, Washington 98195
eclarson@u.washington.edu

Damon M. Chandler
Oklahoma State University
School of Electrical and Computer Engineering
Image Coding and Analysis Lab
Stillwater, Oklahoma 74078

Abstract. *The mainstream approach to image quality assessment has centered around accurately modeling the single most relevant strategy employed by the human visual system (HVS) when judging image quality (e.g., detecting visible differences, and extracting image structure/information). In this work, we suggest that a single strategy may not be sufficient; rather, we advocate that the HVS uses multiple strategies to determine image quality. For images containing near-threshold distortions, the image is most apparent, and thus the HVS attempts to look past the image and look for the distortions (a detection-based strategy). For images containing clearly visible distortions, the distortions are most apparent, and thus the HVS attempts to look past the distortion and look for the image's subject matter (an appearance-based strategy). Here, we present a quality assessment method [most apparent distortion (MAD)], which attempts to explicitly model these two separate strategies. Local luminance and contrast masking are used to estimate detection-based perceived distortion in high-quality images, whereas changes in the local statistics of spatial-frequency components are used to estimate appearance-based perceived distortion in low-quality images. We show that a combination of these two measures can perform well in predicting subjective ratings of image quality. © 2010 SPIE and IS&T.* [DOI: 10.1117/1.3267105]

1 Introduction

The ability to quantify the visual quality of an image in a manner that agrees with human vision is a crucial step for any system that processes consumer images. Over the past several decades, research on this front has given rise to a variety of computational methods of image quality assessment. So-called full-reference quality assessment methods take as input an original image and a distorted version of that image, and yield as output a prediction of the visual quality of the distorted image relative to the original image. The effectiveness of an image quality assessment method is gauged by examining how well the method can predict ground-truth, human-supplied quality ratings obtained via subjective testing.

The earliest methods of full-reference quality assessment were based primarily on the energy of the distortions. Classical examples include mean-squared error (MSE) and peak signal-to-noise ratio (PSNR), which operate based on point-wise differences of digital pixels values. Root-mean-squared (rms) contrast is another example in which the energy of the distortions is measured in the luminance domain.[1] More recently, methods have been developed based on properties of the human visual system (HVS).[2–20] The vast majority of HVS-based methods employ a "perceptual decomposition" that mimics the local spatial-frequency analysis performed in early vision. This decomposition is typically followed by processing stages that take into account near-threshold psychophysical properties such as contrast sensitivity and visual masking.

Another class of methods has recently been proposed that do not explicitly model the stages of vision, but instead operate based on overarching principles of what the HVS is trying to accomplish when viewing a distorted image.[21–24] Overarching principles typically include some form of structural or information extraction, which assumes that a high-quality image is one whose structural content (object boundaries and/or regions of high entropy) most closely matches that of the original image. In Sec. 2, we provide a review of existing approaches to quality assessment.

Despite the clear differences in the way these approaches operate, the vast majority of existing methods share a common thread. Namely, they are rooted in the assumption that when a human determines image quality, the HVS operates via a single strategy. For MSE/PSNR, the assumption is that the strategy employed by the HVS is to gauge the intensity of the distortions. For methods based on near-threshold psychophysics, the assumption is that the strategy employed by the HVS is to process the images via local spatial-frequency decompositions with adjustments for masking, and then to collapse these perceptual decom-

Paper 09070SSPR received May 1, 2009; revised manuscript received Jul. 15, 2009; accepted for publication Jul. 30, 2009; published online Jan. 7, 2010. This paper is a revision of a paper presented at the SPIE conference on Image Quality and System Performance VI, January 2009, San Jose, California. The paper presented there appears (unrefereed) in SPIE Proceedings Vol. 7242.

1017-9909/2010/19(1)/011006/21/$25.00 © 2010 SPIE and IS&T.

(a) Original image | (b) JPEG-2000 compression DMOS = 17.7 | (c) White noise DMOS = 23.5

Fig. 1 When judging the quality of a distorted image containing near-threshold distortions, one tends to rely primarily on visual detection (often using point-by-point comparisons with the original image) in an attempt to locate any visible differences. (a) Close-up of original image Bikes; (b) close-up of image Bikes distorted via JPEG-2000 compression; and (c) close-up of image Bikes distorted by additive Gaussian white noise. DMOS values indicate differential mean opinion scores from the LIVE image database.[25]

positions into a final value of quality. For methods based on overarching principles, the common assumption is that the strategy employed by the HVS is to extract local image structure or use natural image statistics to extract the maximal amount of information, and thus quality is determined based on the extent to which this information can be extracted.

In this work, we advocate an approach to image quality assessment that builds on the strengths of previous approaches, but which operates based on a fundamentally different premise. We assume that the HVS performs multiple strategies when determining quality. Numerous studies have shown the HVS to be a highly adaptive system, with adaptation occurring at multiple levels ranging from single neurons[25] to entire cognitive processes.[26] Thus, even if a human observer is given a fixed, single task of judging image quality, it is reasonable to assume that different strategies might be employed for different conditions (e.g., for different images, different image regions, and/or for different types and amounts of distortion). Here, we present an image quality assessment method that attempts to explicitly model two strategies employed by the HVS: 1. a detection-based strategy for high-quality images containing near-threshold distortions; and 2. an appearance-based strategy for low-quality images containing clearly suprathreshold distortions.

The need to explicitly model these two separate strategies was motivated by our own experiences when simultaneously judging the qualities of several distorted versions of the same original image. We observed that when viewing and judging the quality of each distorted image, the HVS tends to concentrate on different aspects of the images. Specifically, as shown in Fig. 1,[27] some of the distorted images contained just-visible (near-threshold) distortions; these images were consequently judged to be of relatively high quality compared to the original image. For these higher quality images, because the distortions are not readily visible, our visual system seems to employ a detection strategy in an attempt to locate any visible differences. Contrast the images in Fig. 1 with those in Fig. 2. Images shown in Fig. 2 contain clearly visible (suprathreshold) distortions and were consequently judged to be of lower quality. For these lower quality images, the distortions dominate the overall appearance of each image, and thus visual detection is less applicable. Instead, for these latter images, quality is determined based primarily on our ability to rec-

Gaussian blurring (a) DMOS = 59.0 | JPEG-2000 compression (b) DMOS = 67.1 | White noise (c) DMOS = 74.6 | JPEG-2000 compression (d) DMOS = 82.7

Fig. 2 When judging the quality of a distorted image containing clearly visible (suprathreshold) distortions, one tends to rely much less on visual detection and much more on overall image appearance in an attempt to recognize image content in the presence of the dominating distortions. (a) Close-up of image Bikes distorted via Gaussian blurring; (b) and (d) close-up of image Bikes distorted via JPEG-2000 compression; and (c) close-up of image Bikes distorted by additive Gaussian white noise. DMOS values indicate differential mean opinion scores from the LIVE image database.[25]

ognize image content. To summarize, in the high-quality regime, the HVS attempts to look for distortions in the presence of the image, whereas in the low-quality regime, the HVS attempts to look for image content in the presence of the distortions. We argue that these two fundamentally different strategies require two separate computational models.

We have titled our quality assessment method most apparent distortion (MAD) to stress the fact that what is most apparent to the human observer—and thus the strategy employed by the HVS—can change depending on the amount of distortion. MAD operates by using both a detection-based model and an appearance-based model. For detection, we employ a simple spatial-domain model of local visual masking, which takes into account the contrast sensitivity function, and luminance and contrast masking with distortion-type-specific adjustments. Detection-based quality is then estimated based on the mean-squared error (in the lightness domain) between the original and distorted images computed locally for those regions in which the masking model deems the distortions to be visible. For appearance, we employ a model that follows from the texture-analysis literature. The original and distorted images are first decomposed using a log-Gabor filter bank, and the resulting coefficients corresponding to each spatial scale are weighted, with greater weight assigned to coarser scales. Appearance-based quality is then estimated based on the absolute difference between low-level statistics (mean, variance, skewness, and kurtosis) measured for the weighted coefficients of the original and distorted images. Finally, the overall quality of the distorted image is computed by taking a weighted geometric mean of the detection- and appearance-based qualities, where the weight is determined based on the amount of distortion. For highly distorted images, greater weight is given to the appearance-based quality, whereas for images containing near-threshold distortions, greater weight is given to the detection-based quality.

This work is organized as follows. Section 2 summarizes existing methods of full-reference image quality assessment. Section 3 describes the details of the MAD algorithm. Section 4 provides an analysis and discussion of MAD's performance in predicting subjective quality ratings. General conclusions and directions for further research are provided in Sec. 5.

2 Background

Modern methods of image quality assessment can generally be classified as follows: 1. those that operate based on properties of the HVS; 2. those that operate based on measurements of image structure; and 3. those that operate based on other proxy measures of quality. In this section, we provide a brief review of current assessment methods.

2.1 *Methods Based on Properties of the Human Visual System*

Given a distorted image, a human can readily rate the quality of the image relative to the original image and relative to other distored images. Accordingly, a great deal of research in image quality assessment has focused on the use of computational models of the human visual system.[2–17,28–32]

Most HVS-based assessment methods transform the original and distorted images into a "perceptual representation" inspired by both psychophysical and neurophysiological studies of low-level vision.[26,33] The images are typically processed through a set of spatial filters to obtain oriented, spatial-frequency decompositions of the images designed to mimic the decomposition performed in the primary visual cortex. The coefficients of the resulting subbands are then adjusted to take into account variations in visual sensitivity to spatial frequency (contrast sensitivity), and both luminance masking and contrast masking. Finally, the quality of the distorted image is determined based on the extent to which the adjusted subband coefficients of the original image differ from the adjusted subband coefficients of the distorted image. Typically, this final stage is performed by computing point-wise absolute differences between the original and distorted subbands, and then collapsing these differences via an L_p norm (see, e.g., Refs. 5, 6, and 34).

In general, methods of this type perform best as image difference metrics—i.e., they have been designed to determine if changes are visible and accordingly operate best when the distorted images contain artifacts near the threshold of detection. Researchers have previously argued that the underlying visual models need to be extended to take into account higher level properties of human vision.[30,35,36] Unfortunately, although current understanding of low-level (near-threshold) vision is quite mature from a modeling perspective, much less is known about how the HVS operates when the distortions are in the suprathreshold regime in which higher levels of vision are invoked.[33] Nonetheless, recent HVS-based methods have begun to model higher levels of vision.[11,18–20] For example, in Ref. 18 contrast sensitivity and luminance, and contrast masking are augmented by models of suprathreshold contrast perception. In Ref. 20, a wavelet-based model of low-level vision is combined with a model of how the HVS adaptively integrates different spatial frequencies depending on the amount of degradation.

2.2 *Methods Based on Image Structure*

A recent thrust in image quality assessment has focused on measuring degradations in image structure as a proxy for measuring image quality.[19,21,23,37] The central assumption in this approach is that the HVS has evolved to extract structure from the natural environment. Consequently, a higher quality image is one whose structure closely matches that of the original image, whereas a lower quality image exhibits less structural similarity to the original.

Although a precise definition of "image structure" remains an open question, methods of this type have been shown to correlate highly with subjective ratings of quality. In Refs. 21 and 37, Wang *et al.* measure structure based on a spatially localized measure of correlation in pixel values [structural similarity (SSIM)][21] and in wavelet coefficients (MS-SSIM[37]). In Ref. 23, Zhai *et al.* measure structure based on wavelet magnitudes across scales (multiscale edge presentation). In Ref. 19, Carnec, Callet, and Barba combine low-level HVS properties with a measure of structural information, obtained via a stick-growing algorithm and estimates of visual fixation points. In Ref. 38, Yang, Gao, and Pa propose a modified version of MS-SSIM that operates by using the 9/7 wavelet filters. In Ref. 39, Zhang and Mou

combine PSNR with a measure of structure based on differences in wavelet modulus maxima corresponding to low- and high-frequency bands. Structural approaches to quality assessment have also been applied to video[40,41] and wireless applications.[42]

2.3 *Methods Based on Other Measures*

Other measures of image quality have been proposed that operate based on statistical and/or information-theoretic measures. For example, In Ref. 43, Sheikh and Bovik quantify image quality based on natural-scene statistics. VIF operates under the premise that the HVS has evolved based on the statistical properties of the natural environment. Accordingly, the quality of the distorted image can be quantified based on the amount of information it provides about the original. VIF models images as realizations of a mixture of marginal Gaussian densities of wavelet subbands, and quality is then determined based on the mutual information between the subband coefficients of the original and distorted images.

In Ref. 44, Liu and Yang apply supervised learning to derive a measure of image quality based on decision fusion. A training set of images and subjective ratings is used to determine an optimal linear combination of four methods of quality assessment: PSNR, SSIM, VIF, and VSNR. The learning is performed via canonical correlation analysis and images/subjective ratings from the Laboratory for Image and Video Engineering (LIVE) image database[25] (University of Texas at Austin) and the A57 image database.[20] The authors demonstrate that the resulting approach is competitive with VIF.

In Ref. 24, Shnayderman, Gusev, and Eskicioglu measure image quality based on a singular value decomposition (SVD). The Euclidean distance is measured between the singular values of an original image block and the singular values of the corresponding distorted image block; the collection of block-wise distances constitues a local distortion map. An overall scalar value of image quality is computed as the average absolute difference between each block's distance and the median distance over all blocks. Shnayderman, Gusev, and Eskicioglu report that their SVD-based method performs better than SSIM on a suite of test images.

2.4 *Summary of Existing Methods*

Methods that operate based only on the energy of the distortions, such as MSE and PSNR, are attractive due to their mathematical simplicity. However, these methods have also been shown to be relatively poor predictors of visual quality,[45] particularly when comparing images containing different types of distortions. Methods that take into account properties of the human visual system have demonstrated great success at predicting quality for images containing near-threshold distortions. However, these methods generally perform less well for highly distorted images containing suprathreshold distortions unless properties of suprathreshold vision are also taken into account (e.g., Refs. 19 and 20). In contrast, methods based on structural and/or statistical principles have demonstrated success for images containing suprathreshold distortions. However, because these methods lack explicit models of early vision, they generally perform less well on higher quality images containing near-threshold distortions.

In the following section, we describe our approach to image quality assessment, MAD, which builds on the strengths of these existing approaches, but which operates based on the conjecture that the HVS performs multiple strategies when determining quality. We demonstrate that by explicitly modeling two separate strategies (visual detection and visual appearance), and by adaptively combining the outputs of these models based on an estimate of perceived distortion, improved predictions of visual quality can be realized.

3 Algorithm

This section describes the details of the MAD algorithm. First, we describe a method for quantifying perceived distortion in images containing near-threshold distortions (relatively high-quality images); this first method is used to model visual detection. Next, we describe a method for quantifying perceived distortion in images containing suprathreshold distortion (relatively low-quality images); this latter method is used to model image appearance. Finally, we describe a technique used to combine the two perceived distortions into a single estimate of overall perceived distortion.

3.1 *Detection-Based Strategy for High-Quality Images*

When viewing a high-quality image, we argue that the HVS attempts to look for distortions in the presence of the image. To approximate this visual detection strategy, we combine a spatial-domain model of local masking with local mean-squared error measured in the perceived luminance (lightness) domain. An overview of this process is provided in Fig. 3.

As shown in Fig. 3, low-level psychophysical properties such as the spatial contrast sensitivity function, the nonlinear perception of luminance, and luminance and contrast masking are used to compute a map denoting the locations of visible distortions. Next, this visibility map is used to compute a visibility-weighted local MSE map. Finally, the perceived distortion is estimated by collapsing the visibility-weighted local MSE map (via the L_2 norm) into a single scalar value.

3.1.1 *Step 1: compute locations at which the distortions are visible*

Let $\mathbf{I}_{\text{org}}$ denote an n-bit original digital image, and let $\mathbf{I}_{\text{dst}}$ denote a distorted version of the original image, both with digital pixel values in the range $[0, 2^n-1]$ (e.g., $[0-255]$ for 8-bit images). Both images are of size $M \times N$ pixels.

Perceived luminance. To account for the nonlinear relationship between digital pixel values and physical luminances of typical display media, the pixels of the original and distorted images are first converted to luminance images (in units of cd/m^2) via

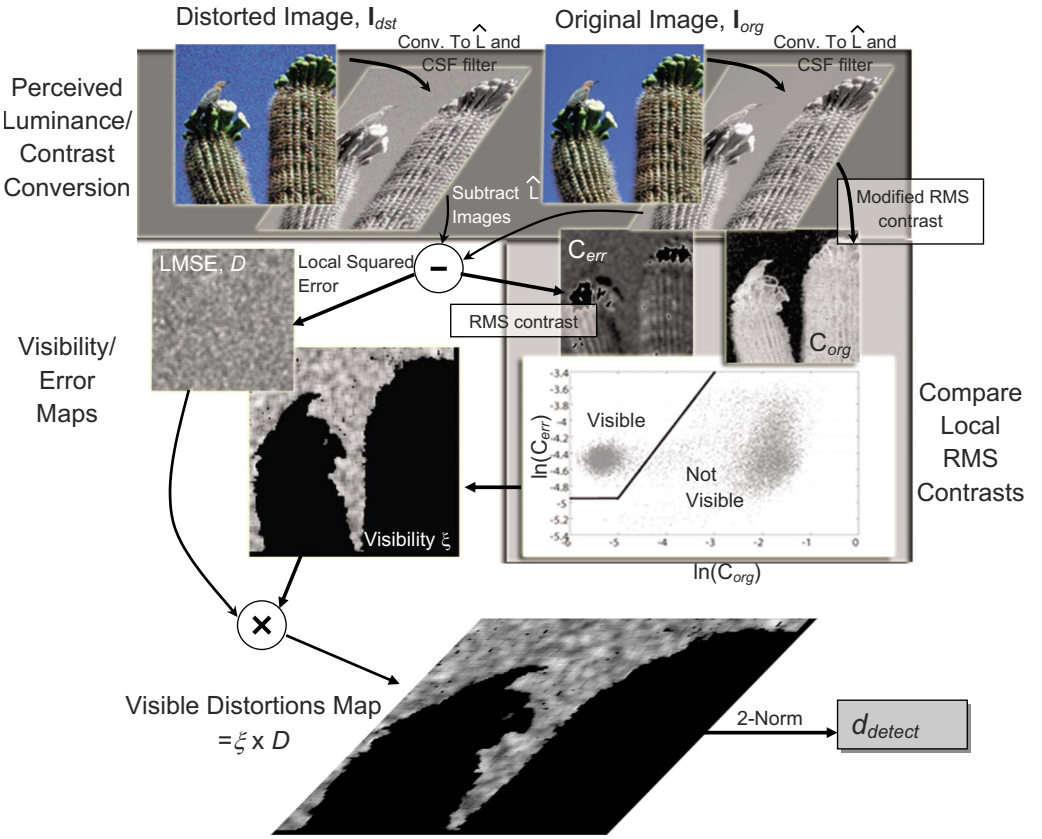

Fig. 3 Evolution of original and distorted image Cactus during the calculation of d_{detect}. Also shown is C_{err} plotted versus C_{org} in log space and the corresponding detection thresholds used to determine visibility.

$$\mathbf{L} = (b + k\mathbf{I})^{\gamma}, \tag{1}$$

where $\mathbf{L}$ denotes the luminance image, and where the parameters b, k, and γ are constants specific to the device on which the images are to be displayed. For 8-bit pixel values and an sRGB display,[46] the values for these parameters are given by $b=0$, $k=0.02874$, and $\gamma=2.2$. Equation (1) is applied to both $\mathbf{I}_{\text{org}}$ and $\mathbf{I}_{\text{dst}}$ to yield $\mathbf{L}_{\text{org}}$ and $\mathbf{L}_{\text{dst}}$, respectively.

Next, we take into account the nonlinear HVS response to luminance by converting luminances into perceived luminances (relative lightness) via

$$\hat{\mathbf{L}} = \sqrt[3]{\mathbf{L}}, \tag{2}$$

where $\hat{\mathbf{L}}$ denotes the relative lightness and is a rough approximation of L^* used in the CIELAB color space. [A cube root is also employed in the computation of the lightness used in the CIELAB color space. However, Eq. (2) lacks the low-luminance nonlinearity and the normalization by the white point commonly used in CIELAB. We account for low-luminance conditions later in Eq. (5). If the white point of the display is known, the lightness used in CIELAB could potentially replace Eq. (2) and the threshold in Eq. (5).] Equation (2) is used for both $\mathbf{L}_{\text{org}}$ and $\mathbf{L}_{\text{dst}}$ to yield the perceived-luminance images $\hat{\mathbf{L}}_{\text{org}}$ and $\hat{\mathbf{L}}_{\text{dst}}$, respectively. The error image is then defined as $\hat{\mathbf{L}}_{\text{err}}=\hat{\mathbf{L}}_{\text{org}}-\hat{\mathbf{L}}_{\text{dst}}$.

Contrast sensitivity function. We account for variations in (near-threshold) sensitivity to spatial frequency by using a model of the contrast sensitivity function (CSF) originally described by Mannos and Sakrison[2] with adjustments specified by Daly[47] (see also Refs. 18 and 48). This CSF, $H(f,\theta)$, is given by

$$H(f,\theta) = \begin{cases} 2.6(0.0192 + \lambda f_\theta)\exp[-(\lambda f_\theta)^{1.1}], & \text{if } f \geq f_{\text{peak}} \text{ c/deg} \\ 0.981 & \text{otherwise} \end{cases}. \tag{3}$$

Here, f denotes the radial spatial frequency in cycles per degree of visual angle (c/deg), $\theta \in [-\pi, \pi]$ denotes the orientation, and $f_\theta = f/[0.15\cos(4\theta)+0.85]$ accounts for the oblique effect (see Appendix A in Sec. 6).

The CSF is applied by filtering both the original image and the error image, where the filtering is performed in the

frequency domain via $\mathbf{I}' = \mathcal{F}^{-1}[\breve{H}(u,v) \times \mathcal{F}[\hat{\mathbf{L}}]]$, where $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ denote the DFT and inverse DFT, respectively. The quantity $\breve{H}(u,v)$ denotes a DFT-based version of $H(f,\theta)$, where u,v are the DFT indices (see Appendix A in Sec. 6). At this point, $\mathbf{I}'_{\text{org}}$ and $\mathbf{I}'_{\text{err}}$ represent the original and error images with values that are linearly proportional to both perceived luminance and perceived contrast. $\mathbf{I}'_{\text{err}}$ can be considered to be the distortions in the image that the HVS could detect if the distortions were viewed against a uniform background (i.e., no masking) rather than being viewed against the image.

Contrast masking. To account for the fact that the presence of an image can reduce the detectability of distortions, we employ a simple spatial-domain measure of contrast masking. First, a local contrast map is computed for the original image by dividing $\mathbf{I}'_{\text{org}}$ into 16×16 blocks (with 75% overlap between neighboring blocks), and then measuring the rms contrast of each block, where rms contrast is measured in the lightness domain. The rms contrast of block p of $\mathbf{I}'_{\text{org}}$ is computed via

$$C_{\text{org}}(p) = \tilde{\sigma}_{\text{org}}(p)/\mu_{\text{org}}(p), \tag{4}$$

where $\mu_{\text{org}}(p)$ denotes the mean of block p of $\mathbf{I}'_{\text{org}}$, and where $\tilde{\sigma}_{\text{org}}(p)$ denotes the minimum of the standard deviations of the four subblocks of p (see Appendix A in Sec. 6).

Whereas $C_{\text{org}}(p)$ is a measure of the local rms contrast in the original image (and thus a measure of the relative masking capability of each region), $C_{\text{org}}(p)$ is independent of the distortions. Accordingly, we next compute a local contrast map for the error image to account for the spatial distribution of the distortions in the distorted image. $\mathbf{I}'_{\text{err}}$ is divided into 16×16 blocks (with 75% overlap between blocks), and then the rms contrast $C_{\text{err}}(p)$ for each block p is computed via

$$C_{\text{err}}(p) = \begin{cases} \sigma_{\text{err}}(p)/\mu_{\text{org}}(p), & \text{if } \mu_{\text{org}}(p) > 0.5 \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\sigma_{\text{err}}(p)$ denotes the standard deviation of block p of $\mathbf{I}'_{\text{err}}$. Here, we employ a lightness threshold of 0.5 to account for the fact that the HVS is relatively insensitive to changes in extremely dark regions.

Finally, we use the two local contrast maps $\{C_{\text{org}}(p)\}$ and $\{C_{\text{err}}(p)\}$ to compute a local distortion visibility map $\{\xi(p)\}$ via

$$\xi(p) = \begin{cases} \ln C_{\text{err}}(p) - \ln C_{\text{org}}(p), & \text{if } \ln C_{\text{err}}(p) > \ln C_{\text{org}}(p) > \delta \\ \ln C_{\text{err}}(p) - \delta, & \text{if } \ln C_{\text{err}}(p) > \delta \geqslant \ln C_{\text{org}}(p) \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Effectively, $\xi(p)$ reflects the amount by which the (log) contrast of the error exceeds the (log) contrast of the original image, if both contrasts are above the threshold ($\delta = -5$). Appendix A in Sec. 6 provides further details of the logic behind Eq. (6).

3.1.2 *Step 2: combine the visibility map with local errors*

After the map of visible locations is created, we use visibility-weighted local MSE measured in the lightness domain to determine perceived distortion. We define the perceived distortion due to visual detection d_{detect} as

$$d_{\text{detect}} = \left\{ \frac{1}{P} \sum_{p} [\xi(p) \times D(p)]^2 \right\}^{1/2}, \tag{7}$$

where the summation is over all blocks and P is the total number of blocks. The quantity $D(p)$ is the local MSE computed for each 16×16 block p via

$$D(p) = \frac{1}{16^2} \sum_{i,j \in N_p} \mathbf{I}'_{\text{err}}(i,j)^2, \tag{8}$$

where N_p is the set of pixels inside block p.

Equation (7) collapses the visibility-weighted local MSE into a single value using the L_2 norm, which is a reasonable approximation of visual summation of distortion in natural images.[49] A value of $d_{\text{detect}} = 0$ indicates that the distortions in the distorted image are not visible. Increasing values of d_{detect} denote increasing perceived distortion and thus decreasing visual quality. However, it is important to note that d_{detect} is not designed to be used in the low-quality regime in which all blocks contain suprathreshold distortion. As discussed in the following section, in the low-quality regime, a different strategy is required.

Figure 4 shows an example of the maps involved in computing d_{detect} for an image distorted via JPEG compression. The original and distorted images are shown in Figs. 4(a) and 4(b), respectively. Notice that the only visible distortions occur in the form of blocking in the background and ringing around the tree's branches. The swarm of worms and the interior of the tree mask the distortion.

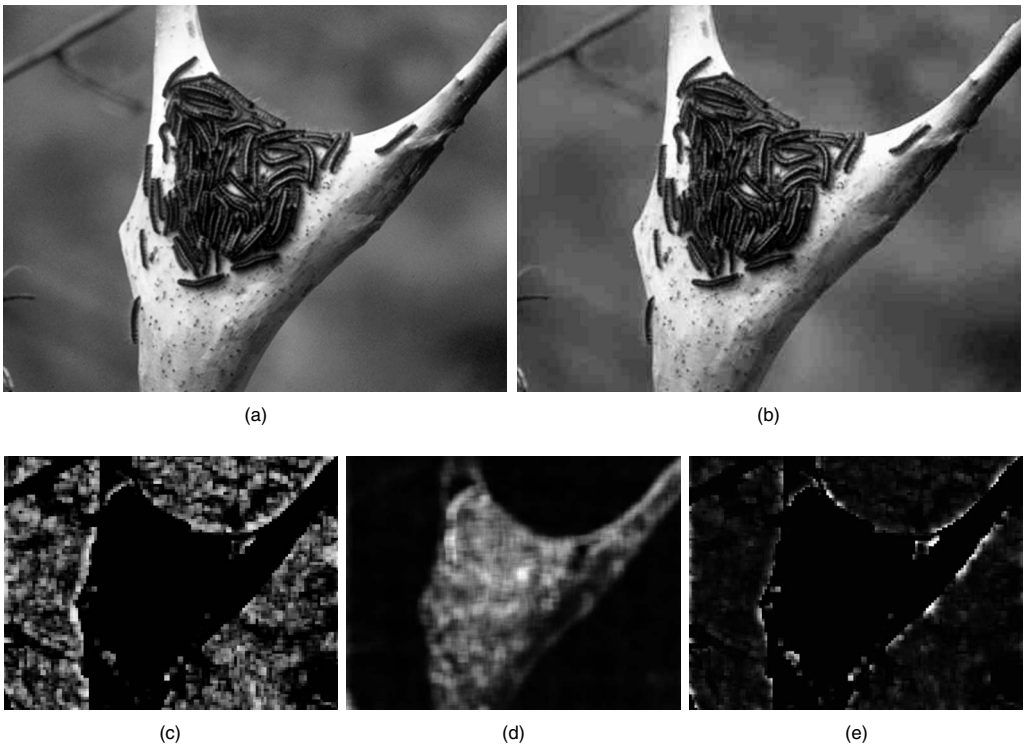

Fig. 4 Perceived distortion in the high-quality regime is determined based largely on masking. Given an original image (a), and a distorted image (b), a map denoting the locations of visible distortions is computed (c), which is then multiplied by local MSE measured in the lightness domain (d). The resulting visibility-weighted local MSE (e) is used to determine d_{detect}. In this image, the only visible distortions occur in the form of blocking in the background and ringing around the branches, whereas the worms and the interior of the tree mask these errors. Because local MSE does not take into account masking, the local MSE map in (d) indicates that the greatest distortions occur in the areas of greatest energy. However, by augmenting local MSE with the distortion visibility map, the final visibility-weighted local MSE map (e) accurately reflects the locations and amounts of visible distortions.

Figures 4(c)–4(e) show the computed visibility map, local MSE map, and visibility-weighted local MSE map $\{\xi(p)\times D(p)\}$, respectively. Notice from Fig. 4(c) that the visibility map does well at capturing the visible artifacts. Notice from Fig. 4(d) that because local MSE does not take into account masking, the local MSE map indicates that the greatest distortions occur in the areas of greatest energy (worms, tree), whereas the distortions in these regions are masked. Finally, notice from Fig. 4(e) that the visibility-weighted local MSE map performs well at indicating both the locations and perceived intensities of the visible distortions. In particular, this latter map captures both the blocking in the background and the ringing around the branches.

Although the spatial-domain model of masking employed here is certainly not a proper model of how masking is effected in the HVS, it provides a reasonable tradeoff between accuracy and computational efficiency. In particular, standard deviation can be approximated efficiently using a fast, separable convolution. Furthermore, although MSE on its own is not always a veridical indicator of perceived distortion, when MSE is computed locally and combined with a visibility map, it can perform well in the high-quality regime. In Sec. 4.5, we demonstrate that, despite its simplicity, this model is quite effective at predicting the perceived distortion of images in the high-quality regime (i.e., for images in which visual detection is the predominant HVS strategy).

3.2 *Appearance-Based Strategy for Low-Quality Images*

In the low-quality regime, we argue that visual masking is of lesser importance to our perception of quality; rather, when the distortions are highly suprathreshold, perceived distortion is better modeled by quantifying the extent to which the distortions degrade the appearance of the image's subject matter. Thus, the HVS switches from a strategy of detecting distortions in the high-quality regime to a strategy of judging image appearance in the low-quality regime.

To model this appearance-based strategy, we employ a method based on local statistics of multiscale log-Gabor filter responses. Models of this type have long been used by the image-processing and computer vision communities to capture the appearance of textures (for example, see Ref. 50). Furthermore, previous researchers have shown that simple cells in the primary visual cortex are well-modeled using log-Gabors,[51,52] and that for textures, changes in the statistics of log-Gabor filter responses are more visually apparent than changes in pixel statistics.[53]

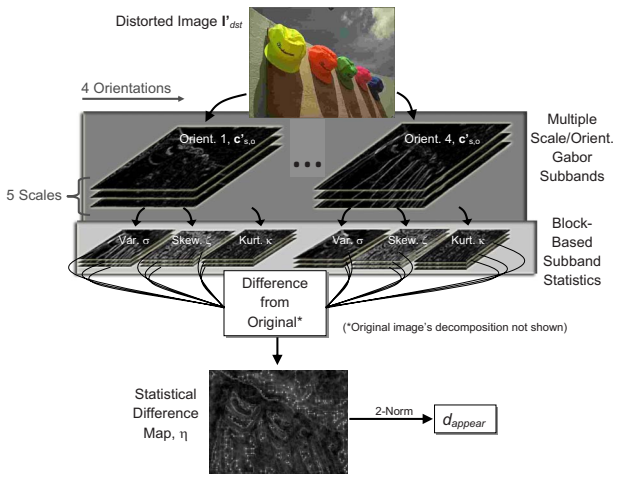

Fig. 5 Evolution of distorted image Caps during the calculation of d_{appear}. The decomposition of the original image is not shown but follows an identical process.

Figure 5 shows an overview of our appearance-based strategy. The decomposition of a distorted image and the subband statistical maps are shown. It is important to note that for this analysis, we deliberately avoid the use of stages designed to model low-level visual detection (CSF and masking). These aspects are captured separately by d_{detect} as described in the previous section. Instead, the method presented here relies solely on changes in log-Gabor statistics in an attempt to capture changes in visual appearance.

3.2.1 *Step 1: apply a log-Gabor decomposition*

The original and distorted images are first decomposed into a set of subbands using a log-Gabor filter bank. The filtering used to obtain the subbands is performed in the frequency domain by computing the inverse DFT of the product of the image's DFT with the following 2-D frequency response:

$$G_{s,o}(r,\theta) = \exp\left[-\frac{(\log r/r_s)^2}{2(\log \sigma_s/r_s)^2}\right] \times \exp\left[-\frac{(\theta-\mu_o)^2}{2\sigma_o^2}\right], \quad (9)$$

where $G_{s,o}$ is the filter denoted by spatial scale index s and orientation index o. The parameter $r=\sqrt{[(u/M/2)^2+(u/N/2)^2]}$ is the normalized radial frequency and $\theta=\arctan(v/u)$ is the orientation. The parameter r_s denotes the normalized center frequency of the scale, and σ_s/r_s determines the bandwidth. The parameter μ_o is the center orientation of the filter and σ_o is the angular spread. Because the log-Gabor filters have been shown to approximate cortical responses in the primary visual cortex,[52] the parameters r_s, σ_s, μ_o, and σ_o can be selected to match corresponding estimates obtained from the mammalian visual system (see Appendix B in Sec. 7).

The log-Gabor decomposition is performed for both the original image and the distorted image via multiplication in the frequency domain. Let $\{\acute{\mathbf{c}}_{s,o}\}$ denote the set of log-Gabor subbands computed for either the original or distorted image, where each subband $\acute{\mathbf{c}}_{s,o} \in \mathbb{R}^{M\times N}$ is the same size as the images. The log-Gabor decomposition is computed using five scales $s=1,\ldots,5$, and four orientations $o=1,\ldots,4$, thus yielding 20 subbands per image. This decomposition is applied to both the original image $\mathbf{I}_{\text{org}}$ and the distorted image $\mathbf{I}_{\text{dst}}$ to yield the sets of subbands $\{\acute{\mathbf{c}}_{s,o}^{\text{org}}\}$ and $\{\acute{\mathbf{c}}_{s,o}^{\text{dst}}\}$, respectively.

3.2.2 *Step 2: compare subband statistics*

After computing $\{\acute{\mathbf{c}}_{s,o}^{\text{org}}\}$ and $\{\acute{\mathbf{c}}_{s,o}^{\text{dst}}\}$, the local subband statistics of the original image are compared with the corresponding local subband statistics of the distorted image to define a local statistical difference map $\{\eta(p)\}$. Specifically, for each 16×16 block (with 75% overlap between blocks), the difference in standard deviation, skewness, and kurtosis of the block's corresponding subband coefficients is computed via

$$\eta(p) = \sum_{s=1}^{5}\sum_{o=1}^{4} w_s[|\sigma_{s,o}^{\text{org}}(p) - \sigma_{s,o}^{\text{dst}}(p)| + 2|\mathrm{s}_{s,o}^{\text{org}}(p) - \mathrm{s}_{s,o}^{\text{dst}}(p)| + |\kappa_{s,o}^{\text{org}}(p) - \kappa_{s,o}^{\text{dst}(p)}|], \quad (10)$$

where $\sigma_{s,o}(p)$, $\mathrm{s}_{s,o}(p)$, and $\kappa_{s,o}(p)$ denote, respectively, the standard deviation, skewness, and kurtosis of the 16×16 subband coefficients corresponding to scale s and orientation o, and corresponding in location to block p. The scale-specific weights $w_s=0.5$, 0.75, 1, 5, and 6 (for the finest to coarsest scales, respectively) are used to account for the HVS's preference for coarse scales over fine scales. Appendix B in Sec. 7 provides additional details regarding Eq. (10).

The final scalar value of perceived distortion in the low-quality regime is given by

$$d_{\text{appear}} = \left[\frac{1}{P}\sum_{p}\eta(p)^2\right]^{1/2}, \quad (11)$$

where the summation is over all blocks and P is the total number of blocks. A value of $d_{\text{appear}}=0$ denotes no perceived distortion, and increasing values of d_{appear} denote increasing perceived distortion and thus decreasing visual quality.

Figure 6 shows the resulting statistical difference map $\{\eta(p)\}$ used in computing d_{appear} for an image Caps from the LIVE image database distorted via JPEG compression. The original and distorted images are shown in Figs. 6(a) and 6(b), respectively. Notice that the most disturbing artifacts manifest as differences in the appearances of the sky and the shadows of the caps. As shown in Fig. 6(c), the statistical difference map succeeds at capturing these changes in visual appearance. It is important to note that we are not saying the HVS performs statistical comparisons, but that these statistics can serve to approximate what the HVS defines as appearance.

3.3 *Adaptively Combining the Two Strategies*

The previous sections presented two measures of perceived distortion: d_{detect} designed for high-quality images containing near-threshold distortions, and d_{appear} designed for low-quality images containing clearly suprathreshold distortions. In this section, we describe how these two measures

(a) (b) (c)

Fig. 6 Perceived distortion in the low-quality regime is determined based largely on changes in the appearance of the image's subject matter. Given an original image (a), and a distorted image (b), a map of local changes in log-Gabor filter-response statistics is computed (c). This statistical difference map is used to determine d_{appear}. Note that the difference map is based on statistics from multiple scales, so distortions are not completely localized.

are combined to yield an overall measure of perceived distortion, applicable for all images, based on how apparently distorted the image appears.

High-quality images should obtain their rating mostly from d_{detect} and low-quality images from d_{appear}. We hypothesize that in the transition between high- and low-quality assessment, the HVS uses a mixture of strategies. This hypothesis makes sense for images in which some regions may appear to be of high quality, while other regions may contain clearly suprathreshold distortions. Images compressed with JPEG and JPEG-2000, for example, fall into this category. Figure 7 shows an example. In comparison to the original image shown in Fig. 7(a), the distorted image shown in Fig. 7(b) contains a mixture of what one would consider high-quality and low-quality regions. The left sidewall of the bridge and the mountain in the background exhibit statistical appearance changes (severe blurring), while the high-contrast trees and brighter areas of the sky largely mask the distortions. In addition, the middle slats of the bridge look somewhat normal, whereas the more distant slats appear clearly degraded.

To capture the interacting strategies, we propose using a weighted geometric mean of d_{detect} and d_{appear}, given by

$$\text{MAD} = (d_{\text{detect}})^{\alpha}(d_{\text{appear}})^{1-\alpha}, \tag{12}$$

where $\text{MAD} \in [0,\infty]$ denotes the overall perceived distortion. The weight $\alpha \in [0,1]$ is chosen based on the (predicted) overall level of distortion. For low levels of distortion, MAD should obtain its value mostly from d_{detect} (i.e., α should approach a value of 1). For high levels of distortion, MAD should obtain its value mostly from d_{appear} (α should approach a value of 0).

Although the optimal technique of selecting α remains an area of future research, we have found that selecting α based on d_{detect} can yield good performance. Here, α is computed via

$$\alpha = \frac{1}{1+\beta_1(d_{\text{detect}})^{\beta_2}}, \tag{13}$$

where β_1 and β_2 are free parameters. For the A57 database,[54] the optimal values for these parameters are $\beta_1 = 0.467$ and $\beta_2 = 0.130$.

3.4 *Summary of Most Apparent Distortion*

To summarize, given an original image $\mathbf{I}_{\text{org}}$, and a distorted version of the original $\mathbf{I}_{\text{dst}}$, MAD uses several stages to arrive at a final prediction of perceived distortion. Each stage is summarized next and shown in Figs. 3 and 5.

Calculate detection model d_{detect}

Stage 1: Compute locations of visible distortions.

1. Convert $\mathbf{I}_{\text{org}}$ and $\mathbf{I}_{\text{dst}}$ to perceived luminance $\hat{\mathbf{L}}_{\text{org}}$, $\hat{\mathbf{L}}_{\text{dst}}$.
2. Compute errors $\hat{\mathbf{L}}_{\text{err}} = \hat{\mathbf{L}}_{\text{org}} - \hat{\mathbf{L}}_{\text{dst}}$.
3. Apply the CSF via Eq. (3) to get $\mathbf{I}'_{\text{org}}$ and $\mathbf{I}'_{\text{err}}$.
4. Compute the rms contrast images C_{org} and C_{err} via Eqs. (4) and (5).
5. Use log rms contrast differences to determine the visibility map via Eq. (6).

Stage 2: Combine visibility with local error image.

1. Compute the local MSE map via Eq. (8).
2. Combine the visibility map and local MSE map via pixel-by-pixel multiplication.
3. Collapse into a single quantity d_{detect} via Eq. (7).

Calculate appearance model d_{appear}

Stage 3: Decompose $\mathbf{I}_{\text{org}}$ and $\mathbf{I}_{\text{dst}}$ into log-Gabor subbands.

1. Compute each subband $\acute{\mathbf{c}}_{s,o}$ at scale index s and orientation index o for the original image and distorted image.

(a) (b)

Fig. 7 Image containing regions of both high quality (trees and nearer slats of the bridge) and low quality (the sidewall and distant slats of the bridge, and the mountain in the background). (a) Original image Bridge. (b) Distorted version of Bridge containing JPEG-2000 compression artifacts with both high-quality and low-quality regions.

Stage 4: Calculate second-, third-, and fourth-order statistics of each subband.

1. Compute local standard deviation $\sigma_{s,o}$, local skewness $\xi_{s,o}$, and local kurtosis $\kappa_{s,o}$, for each subband of the original and distorted images.

2. Take weighted sum via Eq. (10) to get the aggregate statistical difference map.

3. Collapse into a single quantity d_{appear} via Eq. (11).

Calculate adaptation

Stage 5: Calculate adaptive blending by using d_{detect}.

1. Compute blending parameter α via Eq. (13).

2. Calculate final prediction via geometric mean of the detection and appearance model outputs, $MAD = (d_{\text{detect}})^{\alpha}(d_{\text{appear}})^{1-\alpha}$.

In the following section, we analyze MAD's ability to predict subjective ratings of image quality. We show that using the detection and appearance models from before adaptively can be a powerful tool for assessing the relative quality of a distorted image.

4 Results

In this section, the performance of MAD is analyzed in terms of its ability to predict subjective ratings of image quality. To assess its predictive performance, MAD was applied to four databases of subjective image quality: 1. the TID database,[55] 2. the LIVE database,[56] 3. the Toyama database,[57] and 4. a new database of subjective quality released by the authors entitled the Categorical Subjective Image Quality (CSIQ) database.[58,59]

The TID database is a collaborative European effort. The database contains 25 original images, 68 distorted versions of each original image, and subjective ratings for pair-wise image comparisons from 654 different observers from three different countries. TID contains 17 different types of distortions and four different levels of distortion intensity for each original image. This results in 1700 total images in the database. We refer the reader to Ref. 55 for a discussion of the selected distortion types. The overall ratings of TID are presented as mean opinion scores (MOS).

It is of note that the TID database was collected in an uncontrolled manner. Lighting conditions, screen size, monitor type, and color gamma varied between trials. However, the pair-wise comparison task is considerably easier for subjects, and has inherently less variation for images with highly apparent distortions. But, the quality ratings of near-threshold distortions on TID should be taken with a degree of skepticism. Variations in monitors and viewing distances can drastically change the visibility of near-threshold distortions. Even so, TID reports that the variations between observers is smaller than any other database to date. In this respect, TID shows an affinity to capturing how a quality assessment algorithm would perform in a real-world deployment, where variations between monitors and viewing distances are impossible to control.

The LIVE database contains 29 original images, 26 to 29 distorted versions of each original image, and subjective ratings of quality for each distorted image [differential mean opinion scores (DMOS) values]. The distortions used in LIVE are: Gaussian blurring, additive white noise, JPEG compression, JPEG-2000 compression, and simulated data packet loss of transmitted JPEG-2000-compressed images. LIVE contains a total of 779 distorted images.

The Toyama database from Japan contains 14 original images and 168 distorted versions of the originals. The subjective scores were collected using a calibrated CRT monitor in fixed viewing conditions. The database contains two types of distortions, JPEG and JPEG-2000 compressed images. The subjective ratings were collected using a single stimulus absolute scaling. The overall ratings of Toyama are presented in the form of MOS.

The CSIQ database is a new database released by the authors. It consists of 30 original images distorted using six different types of distortions at four to five different levels of distortion. The distortions used in CSIQ are: JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, additive white Gaussian noise, and Gaussian blurring. This results in 866 distorted versions of the original images. CSIQ images are subjectively rated based on a linear displacement of the images across four calibrated LCD monitors placed side by side with equal viewing distance to the observer. All of the distorted versions of an original image were viewed simultaneously on the monitor array and placed in relation to one another according to overall quality. Across image ratings are realigned according to a separate, but identical, experiment in which observers place subsets of all the images linearly in space. The database contains 5000 subjective ratings from 25 different observers, and ratings are reported in the form of DMOS.

4.1 *Performance Measures*

Before evaluating the performance of a particular quality assessment method on a particular database, it is customary to apply a logistic transform to the predicted ratings to bring the predictions on the same scale as the DMOS/MOS values, and to attempt to obtain a linear relationship between the predictions and opinion scores. The logistic transform recommended by the Video Quality Experts Group[60] is a four-parameter sigmoid given by

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp\frac{x - \tau_3}{\tau_4}} + \tau_2. \tag{14}$$

The parameters τ_1, τ_2, τ_3, and τ_4 are chosen to minimize the MSE between the predicted values and the opinion scores. The logistic transform is monotonic and was chosen mainly for its ability to facilitate the use of various performance measures.

The most common measures of predictive performance are the Pearson correlation coefficient (CC) and Spearman rank-order correlation coefficient (SROCC). CC and SROCC are measures of how well an algorithm's predictions correlate with the raw opinion scores, and how well an algorithm predicts the relative ordering of the distorted images, respectively.

There are also two measures of performance more specific to image quality prediction: the outlier ratio[61] R_{out} and the outlier distance d_{out}. These two measures attempt to account for the inherent variation in human subjective ratings of quality. For instance, if the perceived rating of a particular image has large variation between observers, then the average rating is not necessarily a good indication of what should be predicted. Instead, some leeway should

Table 1 Performances of MAD and other quality assessment algorithms on images from the TID, LIVE, Toyama, and CSIQ databases. The best performances are bolded. Italic entries are statistically the same as the best performer for the particular database.

		PSNR	SSIM	MS-SSIM	NQM	VSN	VIF	MAD
CC	TID	0.5355	0.6520	**0.8390**	0.6103	0.6820	0.8055	*0.8306*
	LIVE	0.8707	0.9378	0.9330	0.9096	0.9233	0.9595	**0.9683**
	Toyama	0.6353	0.7970	*0.8924*	*0.8917*	0.8708	**0.9136**	*0.8951*
	CSIQ	0.7998	0.8149	0.8972	0.7418	0.8002	0.9252	**0.9502**
	Average	0.7103	0.8004	0.8904	0.7883	0.8191	0.9009	**0.9110**
SROCC	TID	0.5245	0.6448	**0.8528**	0.6243	0.7046	0.7496	*0.8340*
	LIVE	0.8763	0.9473	0.9437	0.9051	0.9278	0.9633	**0.9675**
	Toyama	0.6126	0.7864	*0.8864*	*0.8886*	0.8610	**0.9080**	*0.8908*
	CSIQ	0.8056	0.8367	0.9137	0.7401	0.8105	0.9192	**0.9466**
	Average	0.7048	0.8038	0.8992	0.7895	0.8260	0.8850	**0.9097**
OR	LIVE	68.16%	59.18%	61.87%	63.80%	58.79%	54.56%	**41.46%**
	Toyama	22.02%	14.29%	*8.33%*	*6.55%*	9.52%	**5.36%**	*7.14%*
	CSIQ	34.30%	33.49%	24.48%	37.30%	31.06%	22.63%	**18.01%**
	Average	41.49%	35.65%	31.56%	35.88%	33.13%	27.52%	**22.21%**
OD	LIVE	4943.3	2814.1	2960.0	3616.8	3246.8	1890.4	**1369.8**
	Toyama	12.9526	6.6174	*1.8967*	*1.6174*	3.6299	**1.3249**	*1.7753*
	CSIQ	3178.0	2896.2	1528.4	4351.9	3325.3	1218.2	**626.2**

be given around the opinion scores associated with the variability of observers. This variability is normally quantified using the standard deviation of all subjective ratings for a particular image σ_{subj}. With this in mind, the outlier ratio is defined as[61]

$$R_{\text{out}} = \frac{N_{\text{false}}}{N_{\text{total}}}, \tag{15}$$

where N_{false} is the number of predictions outside two standard deviations $2\sigma_{\text{subj}}$ of the DMOS or MOS, and N_{total} is the total number of predicted ratings. The range of $2\sigma_{\text{subj}}$ was chosen because it contains 95% of all the subjective quality scores for a given image.

In addition to knowing if a predicted rating is an outlier, it is also informative to know how far outside of the error bars ($\pm 2\sigma_{\text{subj}}$) the outlier falls. To quantify this, we propose a new measure, termed the outlier distance, which is the distance from an outlier to the closest error bar. The outlier distance d_{out} is defined as

$$d_{\text{out}} = \sum_{x \in X_{\text{false}}} \min\left|f(x) - [OS(x) + 2\sigma_{\text{subj}}(x)]\right|, \quad \left|f(x) - [OS(x) - 2\sigma_{\text{subj}}(x)]\right|, \tag{16}$$

where $OS(x)$ is the DMOS or MOS rating of image x, $f(x)$ is the predicted DMOS/MOS rating as defined in Eq. (14), and X_{false} is the set of all predicted ratings outside $2\sigma_{\text{subj}}$. Note that because d_{out} is a sum of the MOS or DMOS, it is dependent on the dynamic range of the database, and therefore cannot be used to compare across databases, only within.

For comparison, we selected several image quality assessment algorithms for which code is readily available. We compare SSIM,[21] MS-SSIM,[62] NQM,[18] VSNR,[54] and VIF[43] to the performance of MAD. All of the algorithms use grayscale versions of the original and distorted images. To ready the images for comparison, they were first converted to grayscale according to $\mathbf{I} = 0.2989 \times \mathbf{R} + 0.5870 \times \mathbf{G} + 0.1140 \times \mathbf{B}$, where $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{B}$ are the red, green, and blue color components of the image. Additionally, each algorithm has slightly different settings associated with its implementation. The default implementation provided in

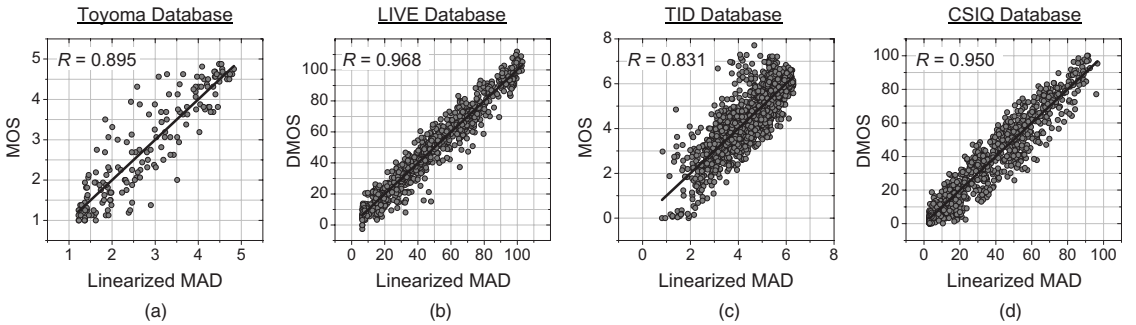

Fig. 8 The overall logistically transformed fits for MAD on the Toyama, LIVE, TID, and CSIQ image databases. The average σ_{subj} is 0.70, 15.8, 0.79, and 9.33 for each database, respectively. All the databases have differing dynamic ranges for MOS and DMOS as seen from the differing *y* axes.

Ref. 63 was chosen for each quality assessment algorithm, which is based on the settings suggested by the authors of the individual algorithms. Settings were held constant across databases.

MAD is implemented assuming a maximum spatial frequency of 16 c/deg. As previously mentioned, we assume $b=0$, $k=0.02874$, and $\gamma=2.2$ in Eq. (1).

4.2 *Overall Performance*

Table 1 shows the performance of MAD and other quality assessment algorithms on the entire set of images from the TID, LIVE, Toyama, and CSIQ databases using CC, SROCC, R_{out}, and d_{out}. R_{out} and d_{out} are not calculated on TID, as the standard deviations between subjects have not yet been released for the database. Also shown in Table 1 is the average performance across databases for CC, SROCC, and R_{out}. Bold entries denote the algorithm with the absolute best performance measure in each database. Italicized entries denote algorithm(s) that are statistically the same as the best-performing algorithm based on an F-test. Notice that MAD is either the best performer in each category or has the same statistical performance as the best-performing algorithm. It also has the best average performance across databases.

For images from the TID database, MAD is tied for the best-performing algorithm. MS-SSIM is the best-performing algorithm (followed closely by MAD, which has the same performance statistically). VIF has a comparatively high linear CC, but this correlation drops considerably when the images are rank ordered.

For images from the LIVE image database, MAD demonstrates the best predictive performance over all algorithms with respect to all four performance measures (CC, ROCC, R_{out}, and d_{out}). In particular, MAD demonstrates considerably lower R_{out} and d_{out} on LIVE than any of the other algorithms.

The Toyama database shows the least selectivity. The best statistical performing algorithms are MAD, MS-SSIM, NQM, and VIF. MAD has the second highest CC and SROCC. VIF shows the best performance, closely followed by NQM, MAD, and MS-SSIM. The limited number of images in the Toyama database makes it difficult to make a statistically significant comparison.

The results on CSIQ are similar to those on LIVE, with MAD demonstrating the best performance in all categories. In fact, MAD vastly outperforms the other algorithms on CSIQ. For instance, d_{out} for MAD is twice as low as the next best-performing algorithm. We attribute this to the even clustering of near- and suprathreshold distortions present in CSIQ. For instance, 1% of the images are subjectively rated within 1.8% of perfect quality, and 14% of the images are rated within 5% of perfect quality. Algorithms must be highly selective in regards to near-threshold distortions to perform well on CSIQ. MAD achieves this.

Scatterplots of DMOS versus MAD can be seen in Fig. 8 for Toyama, LIVE, TID, and CSIQ. For TID, notice that the residuals appear to be heteroscedastic (non-Gaussian residuals), indicating that linear CC should not be the primary means of comparison. In that respect, the considerably higher linear CC achieved by an algorithm could, in part, be attributed to the homoscedasticity of its residuals. To further this comparison, we compute a goodness of fit measure for Gaussianity. This is included in the next discussion on statistical significance.

Figure 8 also shows a fair number of outliers on the LIVE database and some at high quality on TID. These outliers arise largely from nonuniform distortions in the databases (e.g., packet loss). MAD consistently predicts that these images are of lower quality than subjectively rated. For uniform distortion, this is largely a nonissue. This issue is further explored in Sec. 4.4.

In summary, MAD is the best-performing algorithm or tied for best-performing algorithm on each individual database. When MAD is the tied for best-performing algorithm (i.e., on Toyama or TID), it is statistically the same as the best performer. When MAD is the best performer (i.e., on LIVE or CSIQ), all other algorithms have statistically worse performance. When looking at overall performance across databases, MAD has the best average performance.

4.3 *Statistical Significance*

To establish statistical significance, we compare each algorithm's residuals (errors in predictions). If the residuals are sufficiently Gaussian, an F-test can be applied to determine the probability that the residuals are drawn from two different distributions and are thus statistically different; smaller residual variance denotes a better prediction. Note that if the residuals are not Gaussian, the test for significance is considerably more difficult and often inconclusive. One formal test of Gaussianity is the Jarque-Bera (JB) statistic.[64] If the JB statistic of a dataset is sufficiently

Table 2 Statistical significance relationships (ratio of residual variances) between the algorithms on all the databases. A value <1 denotes that the algorithm in the row has smaller residuals than the algorithm in the column. Bold entries are statistically significant with confidence greater than 95%. A value >1 denotes larger residuals. Italicized entries indicate the algorithm in the row is statistically worse than the algorithm(s) in the column. An * appears next to the statistically best-performing algorithms in the database.

		PSNR	SSIM	MS-SSIM	NQM	VSNR	VIP	MAD
TID	PSNR	—	*1.2406*	*2.4088*	*1.1365*	*1.3334*	*2.0306*	*2.2994*
	SSIM	**0.8060**	—	*1.9416*	0.9161	1.0748	*1.6868*	*1.8535*
	MS-SSIM*	**0.4151**	**0.5150**	—	**0.4718**	**0.5535**	**0.8430**	0.9546
	NQM	**0.8799**	1.0916	*2.1194*	—	*1.1732*	*1.7867*	*2.0232*
	VSNR	**0.7500**	0.9304	*1.8065*	**0.8524**	—	*1.5229*	*1.7245*
	VIF	**0.4925**	**0.6110**	*1.1862*	**0.5597**	**0.6566**	—	*1.1324*
	MAD*	**0.4349**	**0.5395**	1.0475	**0.4943**	**0.5799**	**0.8831**	—
	JBSTAT	20.1	98.9	31.6	282.6	559.2	283.1	438.3
LIVE	PSNR	—	*2.0052*	*1.8680*	*1.4012*	*1.6394*	*3.0448*	*3.8744*
	SSIM	**0.4987**	—	0.9316	**0.6988**	**0.8176**	*1.5185*	*1.9822*
	MS-SSIM	**0.5353**	1.0734	—	**0.7501**	0.8776	*1.6300*	*2.0741*
	NQM	**0.7137**	*1.4310*	*1.3331*	—	1.1700	*2.1730*	*2.7650*
	VSNR	**0.6100**	*1.2231*	*1.1394*	**0.8547**	—	*1.8573*	*2.3633*
	VIF	**0.3284**	**0.6586**	**0.6135**	**0.4602**	**0.5384**	—	*1.2724*
	MAD*	**0.2581**	**0.5176**	**0.4821**	**0.3617**	**0.4231**	**0.7859**	—
	JBSTAT	11.8	*2.6*	*7.1*	180.0	20.0	*4.8*	123.5
Toyama	PSNR	—	*1.6852*	*2.9284*	*2.9284*	*2.4672*	*3.6057*	*3.0002*
	SSIM	**0.6116**	—	*1.7909*	*1.7909*	*1.5088*	*2.2051*	*1.8348*
	MS-SSIM*	**0.3415**	**0.5584**	—	1.0000	0.8425	1.2313	1.0245
	NQM*	**0.3415**	**0.5584**	1.0000	—	0.8425	1.2313	1.0245
	VSNR	**0.4053**	**0.6628**	1.1869	1.1869	—	*1.4615*	1.2160
	VIF*	**0.2773**	**0.4535**	0.8122	0.8122	**0.6842**	—	0.8321
	MAD*	**0.3333**	**0.5450**	0.9761	0.9761	0.8223	1.2018	—
	JBSTAT	*8.0*	*2.9*	*0.0*	*2.8*	*18.1*	*2.9*	13.0
CSIQ	JPSNR	—	1.0730	*1.8486*	**0.8012**	1.0018	*2.5020*	*3.7087*
	SSIM	0.9320	—	*1.7229*	**0.7467**	0.9337	*2.8318*	*3.4564*
	MS-SSIM	**0.5410**	**0.5804**	—	**0.4334**	**0.5419**	*1.3534*	*2.0062*
	NQM	*1.2481*	*1.8392*	*2.8073*	—	*1.2504*	*3.1228*	*4.6289*
	VSNR	**0.9982**	1.0710	*1.8452*	**0.7997**	—	*2.4974*	*3.7020*
	VIF	**0.3997**	**0.4289**	**0.7389**	**0.3202**	**0.4004**	—	*1.4823*
	MAD*	**0.2696**	**0.2893**	**0.4985**	**0.2160**	**0.2701**	**0.6746**	—
	JBSTAT	*2.2*	38.6	26.9	43.9	45.6	32.2	17.4

small, one can assume that the data follow a Gaussian distribution. Larger values of the JB statistic denote larger deviations from Gaussianity.

Table 2 shows the summary for overall statistical performance of each algorithm on TID, LIVE, Toyama, and CSIQ. Each entry is the ratio of the residual variance of the algorithm in the row to the algorithm in the column. An F-test was performed between each pair of algorithms. Bold entries denote that the algorithm in the row has statistically smaller residuals than the algorithm in the column with confidence greater than 95%. Italicized entries denote that the algorithm in the row has statistically greater residual variances with the same confidence. Plain text entries denote that there is statistically no difference between the residuals of the two predictions. Starred algorithms in column 2 indicate that no other algorithm performs statistically better on the particular database.

Table 2 also contains the JB statistic measure of Gaussianity. Italicized JB entries denote that the residuals can be deemed Gaussian with 99% confidence. A large JB statistic indicates an increased departure from Gaussian (usually due to the presence of outliers).

On TID, MAD and MS-SSIM are statistically the best-performing algorithms. It is notable that none of the algorithm residuals can be deemed Gaussian with high confidence. For instance, MAD has one of the highest JB statistics due to a large number of outliers at high quality (see Fig. 8). It is surprising, then, that MAD stays competitive with MS-SSIM, which has considerably more Gaussian residuals. That is to say, the variance of MAD is inflated by the existence of outliers, while the variance of MS-SSIM arises from a more even, but wider, clustering across quality ranges.

On LIVE, MAD is statistically the best-performing algorithm. Again, MAD is also deemed as highly non-Gaussian as denoted by the JB statistic. The high JB value is attributable to several outliers (see Fig. 8 in the LIVE database group) that inflate the variance of MAD, similar to the high-quality outliers on TID.

On CSIQ, MAD is statistically the best-performing algorithm. Notice that MAD has slightly more Gaussian residuals than every algorithm except PSNR. While none of the algorithms (except PSNR) can be considered Gaussian, the JB statistics between algorithms are much more consistent. There are no algorithms that are wildly non-Gaussian as seen in other databases. Because of the similarity in Gaussianity for the algorithms, CSIQ provides the most unbiased application of the F-test. Under these conditions, MAD performs well.

On Toyama, statistical significance is highly nonselective. MAD, MS-SSIM, NQM, and VIF have statistically the same performance. Even so, it is of note that MAD performs comparatively well. Toyama seems to show the least preference toward MAD, which is a contraindication from the other database analyses. Perhaps Toyama shows the least preference because it contains the least number of distortion types. This is investigated in the next section.

In summary, MAD was shown to have significantly smaller residuals on the CSIQ and LIVE databases. For Toyama and TID, the residuals of MAD were shown to be drawn from the same distribution as the best-performing algorithm with 95% confidence. For Toyama and TID, there was no algorithm that consistently ranked better than MAD. TID showed mild preference to MS-SSIM, and Toyama showed mild preference to VIF.

4.4 *Performance on Individual Distortion Types*

Table 3 shows the SROCC of MAD and other quality assessment algorithms on the subsets of images from the TID, LIVE, Toyama, and CSIQ databases. The subsets are selected based on the distortion type. Bold entries denote the algorithm with the best overall performance for each distortion on each database. Italicized entries denote the second-best performer. This type of comparison is useful for applications where the distortion type is known beforehand. Table 3, last row, also shows the number of times that the SROCC was above 0.95 (i.e., highly correlated with MOS or DMOS).

When the distortion is known beforehand, the bolded entry from the table denotes the algorithm that might be most appropriate to use. On average, when the source of distortion is held constant, MS-SSIM and VIF show the best performances. MAD performs well on photographic distortions and compression distortions (blur, JPEG, etc.). MAD also performs well on denoised images.

Additionally, notice that MAD has significant trouble with images containing nonuniform distortion, such as packet loss errors, randomly placed solid blocks, and impulse noise. This same phenomenon was observed when looking at the outliers on LIVE and TID. We believe that this is an indication that d_{detect} and d_{appear} could be combined differently. For instance, it is possible to combine the masking and statistical maps and then perform error pooling, rather than error pooling each individual map and then combining the scalar values. In this way, the combination of strategies would be spatially dependent and may be more resilient to nonuniform distortions.

MAD also demonstrates a high number of times that its SROCC is above 0.95, meaning that it could be used effectively when the distortions in an images result from a single factor. VIF shows the most high correlations, with 13 image sets where it is highly correlated. When the distortions come from multiple sources or unknown sources, MAD is still the best overall choice.

4.5 *Performance at Different Distortion Levels*

It is informative to examine how the separate strategies of MAD perform on images containing only near-threshold distortions versus images containing highly visible distortions (i.e., using just d_{detect} and d_{appear} on subsets containing near-threshold distortions or highly suprathreshold distortions).

Table 4 (top) shows the SROCC of various quality assessment algorithms on near-threshold distorted images from the CSIQ, LIVE, and TID databases (top fifth of the quality range of the database). Table 4 (bottom) shows the performances on highly distorted images from the same databases (bottom fifth of the quality range). Toyama is excluded from this comparison because the top and bottom fifth of the quality ranges contain a limited number of images (10 to 20 total images). Also listed in Table 4 are results from the various other quality assessment algorithms. Algorithms in bold are statistically tied for best performance based on an F-test. Notice that indeed d_{detect} per-

Table 3 SROCC of MAD and other quality assessment algorithms on multiple databases using single types of distortion. Bold entries are the best performers in the database for the particular type of distortion. Italicized entries are the second-best performers. The last row shows the number of times the SROCC was above 0.95.

		PSNR	SSIM	MS-SSIM	NQM	VSNR	VIF	MD	Images
Blur	CSIQ	0.929	0.924	*0.972*	0.958	0.945	**0.975**	0.966	150
	LIVE	0.781	0.951	*0.959*	0.859	0.942	**0.973**	0.899	145
	TID	0.868	0.937	**0.961**	0.885	0.933	*0.955*	0.914	100
Awgn	CSIQ	0.936	0.925	0.947	0.939	0.924	*0.957*	**0.960**	150
	LIVE	*0.985*	0.969	0.973	**0.986**	0.979	*0.985*	0.971	145
	TID	**0.911**	0.825	0.809	0.768	0.773	*0.880*	0.863	100
Jpeg	CSIQ	0.888	0922	0.962	0.953	0.9003	**0.970**	*0.966*	150
	LIVE	0.881	0.975	*0.979*	0.957	0.966	**0.984**	0.949	175
	TOY	0.284	0.627	0.835	0.889	0.797	**0.907**	*0.895*	84
	TID	0.901	0.897	*0.935*	0.907	0.917	0.917	**0.941**	100
Jpeg2000	CSIQ	0.936	0.921	*0.969*	0.963	0.948	0.967	**0.977**	150
	LIVE	0.895	0.961	*0.965*	0.944	0.956	**0.969**	0.938	169
	TOY	0.860	0.915	0.947	0.904	0.925	**0.956**	*0.955*	84
	TID	0.830	0.877	**0.974**	0.953	0.952	0.971	*0.972*	100
Contrast	CSIQ	0.862	0.740	**0.952**	*0.948*	0.869	0.936	0.917	116
	TID	0.613	0.629	0.640	*0.727*	0.424	**0.819**	0.492	100
1/fnoise	CSIQ	0.934	0.894	0.933	0.911	0.908	*0.951*	**0.954**	150
Jpeg2000	LIVE	0.893	0.955	*0.930*	0.800	0.905	**0.965**	0.883	145
pack. loss	TID	0.777	0.847	**0.852**	0.726	0.791	*0.851*	0.840	100
Jpeg pack. loss	TID	0.766	0.819	**0.874**	0.737	0.805	*0.858*	0.851	100
Quantity	TID	**0.870**	0.804	*0.854*	0.821	0.827	0.796	0.850	100
Denoised	TID	0.938	0.926	**0.957**	*0.945*	0.929	0.919	*0.945*	100
Awgncolor	TID	**0.907**	0.825	0.806	0.749	0.779	*0.878*	0.839	100
Corrnoise	TID	**0.923**	0.849	0.820	0.772	0.766	0.870	*0.898*	100
Masknoise	TID	*0.849*	0.818	0.816	0.707	0.729	**0.870**	0.736	100
Hifreqnoise	TID	**0.932**	0.858	0.868	0.901	0.881	*0.907*	0.897	100
Impulse	TID	**0.918**	0.759	0.687	0.762	0.647	*0.883*	0.512	100
Pattern	TID	0.593	0.678	0.734	0.680	0.572	*0.761*	**0.838**	100
Block	TID	0.585	**0.891**	0.762	0.235	0.193	*0.832*	0.161	100
Mean shift	TID	0.697	**0.757**	*0.737*	0.525	0.372	0.513	0.589	100
#>0.95		1	5	11	6	4	13	8	NA

Table 4 SROCC for d_{detect} and d_{appear} and other algorithms on images from the multiple databases containing mostly near-threshold distortions (DMOS or MOS in the top fifth of the quality range of the database) and images that are highly distorted (DMOS or MOS in the bottom fifth of the quality range of the database). Bold entries have the best performance, statistically.

		SSIM	MS-SSIM	NQM	VSNR	VIF	d_{detect}	d_{appear}	Images	
High	CSIQ	0.576	**0.717**	0.430	0.563	0.649	**0.684**	0.471	316	DMOS<20
quality	LIVE	0.532	0.532	0.375	0.634	0.557	**0.696**	0.182	193	DMOS<25
image	TID	0.318	0.216	**0.473**	0.338	**0.508**	0.334	0.258	227	MOS>6
Low	CSIQ	0.306	0.576	0.555	0.452	**0.782**	0.561	**0.752**	177	DMOS>65
quality	LIVE	0.558	0.598	0.620	0.513	0.445	0.479	**0.786**	112	DMOS>80
images	TID	0.214	0.491	0.232	0.194	**0.631**	0.128	**0.622**	125	MOS<2.5

forms well on near-threshold distorted images and performs poorly on highly distorted images as designed. Similarly as designed, the opposite relationship is seen for d_{appear}. It performs well on highly distorted images and performs poorly when the distortions are near-threshold. These results strengthen the argument that the HVS changes strategies based on how distorted the images appear.

For LIVE and CSIQ at near-threshold, d_{detect} is the best-performing algorithm or tied for best-performing algorithm, as we would predict. This is not the case for TID. No algorithm performs well when the distortions are mostly near-threshold. As mentioned previously, we expected all algorithms to perform poorly on near-threshold distortions in TID because of the wide variation in viewing conditions during subjective testing. This variability makes reliable calculations of masking impossible, which explains why d_{detect} may be performing poorly. NQM and VIF appear to be the most resilient to these variations, with SROCC's of 0.47 and 0.51. Though they are the best performers at this distortion range on TID, the correlations are quite small and the overall relationship to quality is still weak.

For highly distorted images, d_{appear} is the best performer or statistically equal to the best performer on all the databases. VIF also does well on CSIQ and TID, but not on LIVE. As one might predict, d_{appear} is the most consistent performer in this distortion range. These results further support the use of separate strategies for near- and suprathreshold distortions.

Additionally, it is interesting to quantify how the dual strategy contributes to the value of MAD. Table 5 shows the SROCC and linear CC (after logistic transformation) of 1. d_{detect} versus MAD, 2. d_{appear} versus MAD, and 3. d_{appear} versus d_{detect}. On Toyama d_{appear} is more correlated with the value of MAD than d_{detect}. The opposite relationship is seen on TID, LIVE, and CSIQ. Moreover, both d_{detect} and d_{appear} are always highly correlated with MAD, indicating it receives significant influence from both strategies. The values of d_{detect} and d_{appear} are less correlated with each other, indicating that each strategy indeed provides some information independent of the other. However, it is reasonable to assume that d_{detect} is slightly more influential than d_{appear}, as Toyama has a limited number of images and is the only database to show preference to d_{appear}.

Also shown in Table 5 is the mean and standard deviation of α on each of the databases. For the most part, α has a mean of about 0.45 and 0.1 standard deviation. Each strategy receives a varying weight depending on the visibility of the distortions (i.e., α does not stay within a limited range in the databases). Toyama shows the least variability, with alpha having a standard deviation of 0.0414. It is unclear why α is restricted on Toyama. Perhaps more images or more distortions would provide a wider range of variability.

Table 5 SROCC/linear CC, respectively, for d_{detect} versus MAD, d_{appear} versus MAD, and d_{appear} versus d_{detect} on multiple database. The mean and standard deviation of α on each database is also shown.

	d_{appear} versus MAD	d_{detect} versus MAD	d_{appear} versus d_{detect}	mean α	std α
TID	0.8206/0.9111	0.9558/0.9895	0.6481/0.8784	0.394	0.0721
TOY	0.9912/0.9752	0.7402/8520	0.6552/0.9537	0.6493	0.0414
LIVE	0.9523/0.5439	0.9588/0.9787	0.8522/0.6073	0.4216	0.1207
CSIQ	0.9406/0.9050	0.9410/0.9903	0.8079/0.8866	0.4714	0.1735

4.6 *Limitations and Future Work*

The performance of MAD is not without its downsides. In particular, the efficiency, masking implementation, and appearance model are in the first iterations of design. MAD is meant to advocate that by using simple assumptions about dual strategies, one can achieve excellent subjective quality prediction. There are still a number of directions that can be improved on in MAD.

The appearance model d_{appear} has a significant computation time and somewhat high memory footprint. A typical Matlab implementation with a 512×512 image can take up to 50 s on a 2.0-GHz AMD Athlon X2 processor. The significant computation time can be attributed to the log-Gabor image decomposition and the subsequent local statistics calculation on each subband. This bottleneck of the implementation requires the use of three nested loops. However, only a single loop traverses based on the image size. Thus the implementation grows at $O(N)$, where N denotes the number of pixels in the image.

In addition, each log-Gabor subband must be saved in memory while the statistics are calculated. At any given time, d_{appear} may have the subband statistics of the reference and distorted image loaded in memory, and the log-Gabor filter bank decomposition for a single subband in memory. This results in three additional double precision maps of size N pixels each. Thus the memory footprint is three times the size of the input image (converted to double precision).

The computation of d_{detect} has a relatively low computational complexity and memory footprint. A typical Matlab implementation with a 512×512 image takes about 0.8 s on a 2.0-GHz AMD Athlon X2 processor. It requires at most two double precision maps of size N pixels at any time during processing.

Aside from complexity, both d_{detect} and d_{appear} could be implemented in completely different manners. The current version of MAD uses a masking model with which it is difficult to separate the effects of different types of masking in an informative way. For instance, spatial frequency and Weber's law are heuristically incorporated using the mean and standard deviation, but not accounted for optimally (i.e., using real experimental observations). Additionally, contrast detection thresholds and contrast (or pattern) masking are heuristically incorporated (again, using the mean and standard deviation). From this perspective, it is arguable that d_{detect} provides only a reasonable masking approximation, not an optimal model of masking.

Furthermore, the appearance model used in MAD is only an approximation of what the HVS might be doing. It is unclear why log-Gabor statistics capture overall appearance, and if the information could be achieved in a more efficient, direct manner. We highly encourage other interpretations of what the HVS might be capturing in this quality regime.

As mentioned previously, MAD has trouble predicting the quality of images with nonuniform distortions such as packet loss and impulse noise. We are currently looking at making Eq. (12) spatially dependent. Instead of a single, scalar α blending parameter, there would be an entire α-map. In this way, one α-map could be ascertained from MAD, which quantifies the portions of an image where the HVS employs strategies that are either mostly detection based or mostly appearance based. This type of mapping may be useful for applications involving not only compression, but also unequal error protection in images.

We are also looking at extending MAD to account for color artifacts. In its current state, MAD is blind to color-only distortions and masking effects related to hue. These types of extensions to MAD could provide significant improvements. Video quality assessment could also benefit from a dual strategy decomposition like MAD. It is unclear how the HVS adapts to different quality regimes in video, and MAD could be a good starting model. We note, however, that the computational complexity must be reduced for a video implementation of MAD, and are actively working to further reduce the run time and memory footprint of MAD.

5 Conclusions

This work presents a new method of image quality assessment that operates under the premise that the HVS performs two distinct strategies when assessing image quality. For high-quality images, because the distortions are not readily visible, our visual system seems to employ a detection strategy in an attempt to locate any visible differences. For low-quality images, because the distortions tend to dominate the image's overall appearance, visual detection is less applicable; rather, quality is determined based primarily on our ability to recognize image content. Thus, in the high-quality regime, the HVS attempts to look for distortions in the presence of the image; whereas in the low-quality regime, the HVS attempts to look for image content in the presence of the distortions. We argue that these two fundamentally different strategies require two separate computational models.

Accordingly, two separate computational measures of perceived distortion are presented. The first measure, designed for high-quality images, assesses perceived distortion by taking into account contrast sensitivity, and local luminance and contrast masking. The perceived distortion is computed via a visibility-weighted local error measurement computed in the lightness domain (L^*). The second measure, designed for low-quality images, assesses perceived distortion based on the changes in local statistics between the subbands of the original image and the subbands of the distorted image. An overall measure of perceived distortion is computed via a weighted geometric mean of the high and low-quality measures, where the weight is determined based on the estimated level of distortion.

The proposed MAD measure is shown to perform well on images from the TID image database,[55] LIVE image database,[25] Toyama image database,[57] and the CSIQ image database.[58] Statistically significant improvements in predicting subjective ratings are achieved in comparison to a variety of existing algorithms. Some notable limitations of MAD include its relatively high computational complexity and memory footprint attributable to the log-Gabor decomposition required for the low-quality measure. In addition, MAD is blind to color-only distortion and has not yet been tested over a range of viewing distances. We are currently in the process of refining the masking model and the log-Gabor decomposition to better take into account viewing distance. We are also exploring extensions of this work for color images and video quality assessment.

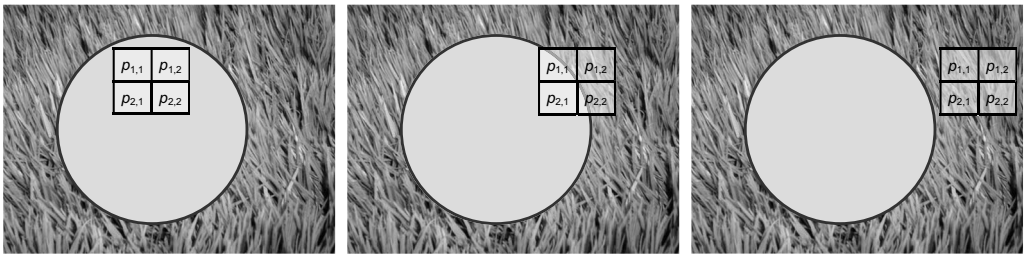

Fig. 9 Example of measuring the local (block-based) contrast in an image containing a low-contrast object (here, a solid disk) placed on a high-contrast background, as is often found in natural images. In (a), the contrast of the block is low. In (c), the contrast of the block is high. In (b), the contrast of the block would normally be higher than the contrast measured in (a), despite the fact that this object border is not effective at masking.[68] Here, we overcome this limitation by taking the effective local contrast to be the minimum contrast of all four subblocks. Thus in (b), since subblock $p_{2,1}$ has the lowest contrast, the effective local contrast is measured to be just as low as that measured in (a).

Appendix 1: Details of the Detection-Based Strategy

This section provides details of the contrast sensitivity function and contrast masking model used in the computation of d_{detect}.

A.1 Contrast Sensitivity Function

Equation (3) in Sec. 3.1.1 specifies the CSF used in the computation of d_{detect}. The quantity f_θ in this equation is given by $f_\theta = f/[0.15\cos(4\theta)+0.85]$, which represents an orientation-based modification of f that effects an approximately −3-dB attenuation in $H(f,\theta)$ along the diagonal orientations to account for decreased contrast sensitivity along diagonal orientations (the oblique effect, see Ref. 5).

The CSF is further adjusted as described in Ref. 5 to have a lowpass profile by explicitly setting frequencies below f_{peak} to 0.981, which is the maximum value of $H(f,\theta)$ as determined by λ. In Refs. 5, 18, and 48, $\lambda=0.114$, resulting in a peak at a frequency of $f_{\text{peak}}\approx 8$ c/deg (where the peak is measured before forcing the lowpass profile). Here, we have used $\lambda=0.228$, resulting in $f_{\text{peak}}\approx 4$c/deg, which is within the range of 1 to 6 c/deg typically reported for the CSF (see, e.g., Refs. 26 and 65).

The CSF in Eq. (3) is specified in terms of f and θ. These quantities can be computed from discrete Fourier transform (DFT) indices $u\in[-M/2,M/2]$ and $v\in[-N/2,N/2]$ via

$$f=\left[\left(\frac{u}{M/2}\right)^2+\left(\frac{v}{N/2}\right)^2\right]^{1/2}\frac{\rho\vartheta\tan\left(\frac{\pi}{180}\right)}{2}\ \text{c/deg}, \tag{17}$$

$$\theta=\arctan\left(\frac{v}{u}\right), \tag{18}$$

where ρ is the display resolution in pixels per unit distance (e.g., pixels/inch), and ϑ is the viewing distance in those same distance units (e.g., inches). [The quantity $\rho\vartheta\tan(\pi/180)$ is the display visual resolution[66] in units of pixels per degree of visual angle. Dividing this quantity by two yields the maximum number of cycles per degree (maximum spatial frequency) in the horizontal or vertical direction. Scaling this latter quantity by $\sqrt{[(u/M/2)^2+(v/N/2)^2]}$ yields the particular spatial frequency f in c/deg.] Thus, given ρ and ϑ, we use the prior conversions to compute a DFT-based CSF $\breve{H}(u,v)$ from $H(f,\theta)$ via

$$\breve{H}(u,v)=H\left\{\left[\left(\frac{u}{M/2}\right)^2+\left(\frac{v}{N/2}\right)^2\right]^{1/2}\frac{\rho v\tan\left(\frac{\pi}{180}\right)}{2},\arctan\left(\frac{v}{u}\right)\right\}. \tag{19}$$

A.2 Effective Local Root Mean Square Contrast in the Original Image

The contrast masking model described in Sec. 3.1.1 uses a modified standard deviation to compute the local rms contrast of the original image. Specifically, $\tilde{\sigma}_{\text{org}}(p)$ in Eq. (4) denotes the modified standard deviation of block p of I'_{org} given by

$$\tilde{\sigma}_{\text{org}}(p)=\min[\sigma(p_{1,1}),\sigma(p_{1,2}),\sigma(p_{2,1}),\sigma(p_{2,2})]. \tag{20}$$

The quantities $\sigma(p_{1,1})$, $\sigma(p_{1,2})$, $\sigma(p_{2,1})$, and $\sigma(p_{2,2})$ correspond to the standard deviations of the four subblocks of block p as illustrated in Fig. 9.

This modified standard deviation is used to compensate for the fact that edges and object boundaries are not effective at masking, despite the fact that these regions exhibit high contrast.[67] As shown in Fig. 9,[68] by defining the effective contrast based on the minimum contrast of each block's four subblocks, the resulting contrast around edges/object borders is measured to be relatively low.

A.3 Mapping Contrast Ratios to Visibility

To determine the locations at which the distortions in the distorted image are visible, Eq. (6) is used to compare C_{err} with C_{org} to yield the visibility map $\{\xi(p)\}$ as described in Sec. 3.1.1.

The first statement in Eq. (6) handles the case in which the (log) contrast of the error is greater than the (log) contrast of the image in block p, and both of these contrasts are

above a minimum (log) threshold contrast ($\delta=-5$, chosen empirically by visually inspecting a number of generated masking maps and distorted images). See Fig. 3 for a plot of C_{err} versus C_{org}. In this case, $\xi(p)$ is set to the amount by which the (log) contrast of the error exceeds the (log) contrast of the image.

The second statement in Eq. (6) handles the case in which the image's (log) contrast in block p is at or below the threshold (δ), but the (log) contrast of the error in p is above the threshold. In this case, $\xi(p)$ is set to the amount by which the (log) contrast of the error exceeds the threshold. This second case is designed to avoid disproportionately high values of $\xi(p)$ due to blocks with low image contrast.

The final statement handles the case in which both the image and error exhibit a contrast at or below the threshold, in which case $\xi(p)$ is set to zero. Progressively greater values of $\xi(p)>0$ denote locations at which the error is progressively more visible. A demonstration of each stage of the masking model is provided in an online supplement to this work located at http://vision.okstate.edu/mad/.

Appendix 2: Details of the Appearance-Based Strategy

In this section, we provide details regarding the log-Gabor decomposition and the comparison of local subband statistics used in the computation of d_{appear}.

B.1 Log-Gabor Decomposition

The log-Gabor decomposition is specified by Eq. (9) in Sec. 3.2.1. This decomposition is computed at five scales $s=1,\ldots,5$, using values of $r_s=2/3s$ and $\sigma_s/r_s=1.1$. These values result in an approximately 1.5 octave spacing of the center frequencies of the bands, with each band spanning an approximately 1.5 octave bandwidth. Four orientations $o=1,\ldots,4$ are computed using $\mu_o=(o-1)\pi/4$ corresponding to 0, 45, 90, and 135 deg. The angular spread σ_o is fixed at $\pi/6$. These values were chosen empirically to minimize overlap between bands, but still uniformly cover most of the frequency spectrum as in the mammalian visual system.[51]

Note that the frequency responses of the filters defined by Eq. (9) are nonzero only for positive frequencies due to the one-sided orientation component of the response. This deliberate lack of conjugate symmetry allows the computation of two spatial-domain convolutions via a single DFT/inverse DFT. Specifically, each subband is computed via

$$\acute{\mathbf{c}}_{s,o}=\mathcal{F}^{-1}[\breve{\mathbf{G}}_{s,o}(u,v)\times\mathcal{F}[\mathbf{I}]], \tag{21}$$

where $\breve{G}_{s,o}(u,v)=G_{s,o}\{\sqrt{[(u/M/2)^2+(v/N/2)^2]},\arctan(v/u)\}$, and where $\mathbf{I}$ denotes the original or distorted image. This equation is equivalent to convolving $\mathbf{I}$ in the spatial domain with both even and odd-symmetric log-Gabor filter kernels. The inverse DFT yields complex-valued subband coefficients ($\mathbf{c}_{s,o}\in\mathbb{C}^{M\times N}$) in which the real part of the coefficients correspond to the even-symmetric filter outputs, and the imaginary part of the coefficients correspond to the odd-symmetric filter outputs.

For each subband, we collapse the real and imaginary components of each coefficient (even and odd filter output) into a single magnitude via

$$\acute{\mathbf{c}}_{s,o}=\sqrt{\Re\{\mathbf{c}_{s,o}\}^2+\Im\{\mathbf{c}_{s,o}\}^2}, \tag{22}$$

where $\acute{\mathbf{c}}_{s,o}\in\mathbb{R}^{M\times N}$. The log-Gabor decomposition thus yields 20 subbands (5 scales$\times$4 orientations) containing only positive coefficients. This decomposition is applied to both the original image $\mathbf{I}_{\text{org}}$ and the distorted image $\mathbf{I}_{\text{dst}}$ to yield the sets of subbands $\{\acute{\mathbf{c}}^{\text{org}}_{s,o}\}$ and $\{\acute{\mathbf{c}}^{\text{dst}}_{s,o}\}$, respectively.

B.2 Comparing Subband Statistics

To capture appearance-based changes, the local subband coefficient statistics of the original image are compared with those of the distorted image. This comparison is specified by Eq. (10) in Sec. 3.2.2.

The skewness difference in Eq. (10) is multiplied by a factor of 2 to bring it on approximately the same scale as the σ and κ differences. The scale-specific weights w_s are used to account for the HVS's preference for coarse scales over fine scales (Ref. 69). The values chosen for w_s for the finest to coarsest scales are 0.5, 0.75, 1, 5, and 6, which were selected to yield the best performance on the A57 database.[54] Using these weights, the finest scale contributes approximately 3.8% to $\eta(p)$, and coarser scales contribute 5.6, 7.5, 37.7, and 45.2%, respectively. (The proper values of w_s remain an area of future research. Adjustment of w_s only marginally affects the performance of the appearance-based model.)

The statistics of the log-Gabor filter outputs have been widely used to define visual appearance and texture. Specifically, changes in standard deviation, skewness, and kurtosis have been shown to be good indications of discriminable texture appearance.[53] A change in the standard deviation of log-Gabor subband coefficients means that the outputs of certain log-Gabor filters change in intensity. When these changes are computed on a block-by-block basis, we are also able to approximately locate where in the image the filter outputs change. For instance, blurring dramatically changes the histogram of the log-Gabor outputs. 1. At high frequency, the histogram of filter outputs becomes more peaked with smaller standard deviation. 2. Additionally, if the outputs were skewed to one side before blurring, the histogram is likely to be more symmetric afterward. Compression artifacts would have different but measurable changes in the log-Gabor statistics. These changes in local subband statistics can approximate the perceived distortion of local image structure.

References

1. B. Moulden, F. A. A. Kingdom, and L. F. Gatley, "The standard deviation of luminance as a metric for contrast in random-dot images," *Perception* **19**, 79–101 (1990).
2. J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of image," *IEEE Trans. Inf. Theory* **20**, 525–535 (1974).
3. F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.* **30**(7), 1679–1692 (1982).
4. N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.* **33**(6), 551–557 (1985).
5. S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206 (1993).
6. P. C. Teo and D. J. Heeger, "Perceptual image distortion," *Proc. SPIE* **2179**, 127–141 (1994).
7. S. J. P. Westen, R. L. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel HVS model," *Intl. Conf. Acoustics, Speech, Signal Process.* 4, 2351–2354 (1995).
8. J. Lubin, "A visual discrimination model for imaging system design

and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, Ed., pp. 245–283, World Scientific, New York (1995).
9. C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image representation and quality assessment applications," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Process.*, pp. 2291–2294 (May 1996).
10. Y. Lai and C. J. Kuo, "Image quality measurement using the haar wavelet," *Proc. SPIE* **3169**, 127 (1997).
11. M. Miyahara, K. Kotani, and V. R. Algazi, "Objective picture quality scale (PQS) for image coding," *IEEE Trans. Commun.* **46**(9), 1215–1226 (1998).
12. W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," *Proc. IEEE Intl. Conf. Image Process.*, 3, 414–418 (1998).
13. S. Winkler, "A perceptual distortion metric for digital color images," *Proc. IEEE Intl. Conf. Image Process.*, 3, 399–403 (1998).
14. A. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Process.* **8**, 717–730 (May 1999).
15. S. Winkler, "Visual quality assessment using a contrast gain control model," in *IEEE Signal Process. Soc. Workshop Multimedia Signal Process.*, pp. 527–532 (Sep. 1999).
16. J. Lubin, M. Brill, A.. De Vries, and O. Finard, "Method and apparatus for assessing the visibility of differences between two image sequences," U.S. Patent 5,974,159 (1999).
17. "Jndmetrix technology," Sarnoff Corp., see http://www.sarnoff.com/.
18. N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.* **9**, 636–650 (2000).
19. M. Carnec, P. L. Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *ICIP 2003*, **2**, 185–188 (2003).
20. See http://foulard.ece.cornell.edu/VSNR.html.
21. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, 600–612 (2004).
22. H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.* **14**(12), 2117–2128 (2005).
23. G. Zhai, W. Zhang, X. Yang, and Y. Xu, "Image quality assessment metrics based on multi-scale edge presentation," *IEEE Workshop Signal Process. Syst. Design Implement.*, pp. 331–336 (2005).
24. A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Trans. Image Process.* **15**(2), 422–429 (2006).
25. R. L. DeValois and K. K. DeValois, *Spatial Vision*, Oxford University Press, Oxford, UK (1990).
26. J. Schulkin, *Cognitive Adaptation*, 1st ed., Cambridge University Press, Boston, MA (2008).
27. H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at live," see http://live.ece.utexas.edu/research/quality/.
28. C. C. Taylor, Z. Pizlo, J. P. Allebach, and C. A. Bouman, "Image quality assessment with a gabor pyramid model of the human visual system," *Proc. SPIE* **3016**(1), 58–69 (1997).
29. P. LeCallet, A. Saadane, and D. Barba, "Frequency and spatial pooling of visual differences for still image quality assessment," *Proc. SPIE* **3959**, 595–603 (2000).
30. T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," *Proc. ICIP*, pp. 237–240 (1996).
31. M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.* **70**, 177–200 (1998).
32. A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Which semi-local visual masking model forwavelet based image quality metric?" in *ICIP*, pp. 1180–1183 (2008).
33. N. Graham, *Visual Pattern Analyzers*, Oxford University Press, New York (1989).
34. A. B. Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A* **14**, 2378–2390 (1997).
35. W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in jpeg 2000," *Signal Process. Image Commun.* **17**, 85–104 (2001).
36. D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.* **14**(4), 397–410 (2005).
37. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst. Computers*, Vol. 2, p. 1398 (Nov. 2003).
38. C. L. Yang, W. R. Gao, and L. M. Po, "Discrete wavelet transform-based structural similarity for image quality assessment," in *ICIP*, pp. 377–380 (2008).
39. M. Zhang and X. Mou, "A psychovisual image quality metric based on multi-scale structure similarity," in *ICIP*, pp. 381–384 (2008).
40. Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.* **19**(2), pp. 121–132 (2004).
41. S. Ye, K. Su, and C. Xiao, "Video quality assessment based on edge structural similarity," *Image Signal Process. Congress* **3**, 445–448 (2008).
42. U. Engelke and H. J. Zepernick, "Pareto optimalweighting of structural impairments for wireless imaging quality assessment," in *ICIP*, pp. 373–376 (2008).
43. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–444 (2006).
44. M. Liu and X. Yang, "A new image quality approach based on decision fusion," *Fuzzy Syst. Knowledge Discovery, 4th Intl. Conf.*, 4, 10–14 (2008).
45. B. Girod, "What's wrong with mean-squared error?" in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 207–220 (1993).
46. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, "A standard default color space for the internet-srgb," see http://www.w3.org/Graphics/Color/sRGB (1996).
47. S. J. Daly, "Subroutine for the generation of a human visual contrast sensitivity function," Eastman Kodak Tech. Report 233203y (1987).
48. T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," *IEEE Intl. Conf. Acoustics, Speech, Signal Process.*, pp. 301–304 (1993).
49. D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Am. A* **20**, pp. 1167–1180 (July 2003).
50. A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recogn.* **24**, 1167–1186 (May 1991).
51. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
52. B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by v1?" *Vision Res.* **37**, 3311–3325 (1997).
53. F. A. A. Kingdom, A. Hayes, and D. J. Field, "Sensitivity to contrast histogram differences in synthetic wavelet-textures," *Vision Res.* **41**, 585–598 (1995).
54. D. M. Chandler and S. S. Hemai, "Vsnr: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.* **16**(9), 2284–2298 (2007).
55. N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in *4th Intl. Workshop Video Process. Quality Metrics Consumer Electron.*, **6**, p. 6 (Jan. 2009).
56. H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," see http://live.ece.utexas.edu/research/quality/.
57. Y. Horita, K. Shibata, and Z. M. Parvez Saddad, "Subjective quality assessment toyama database," see http://mict.eng.u-toyama.ac.jp/mict/ (2008).
58. See http://vision.okstate.edu/csiq/.
59. E. C. Larson and D. M. Chandler, "Most apparent distortion: a dual strategy for full-reference image quality assessment," *Proc. SPIE* **7242**, 72420S (2009).
60. VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii," see http://www.vqeg.org (Aug. 2003).
61. H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**, 1349–1364 (2006).
62. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *37th IEEE Asilomar Conf. Signals, Syst. Computers*, Vol. 2, pp. 1398 (2003).
63. M. D. Gaubatz, D. M. Rouse, and S. S. Hemami, "MeTriX MuX," http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.
64. A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ. Lett.* **6**, 255–259 (1980).
65. E. Peli, L. E. Arend, G. M. Young, and R. B. Goldstein, "Contrast sensitivity to patch stimuli: effects of spatial bandwidth and temporal presentation," *Spatial Vis.* **7**, 1–14 (1993).
66. A. B. Watson, G. Y. Tangand, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.* **6**, 1164–1175 (1997).
67. D. M. Chandler, M. D. Gaubatz, and S. S. Hemami, "A patch based structural masking model with an application to compression," in *EURASIP J. Image Video Process.* **2009**, 649316 (2009).
68. S. S. Hemami, D. M. Chandler, B. C. Chern, and J. A. Moses, "Suprathreshold visual psychophysics and structure-based visual masking," in *Proc. SPIE* **6077**, 60770O (2006).
69. D. M. Chandler and S. S. Hemami, "Suprathreshold image compression based on contrast allocation and global precedence," *Proc. SPIE* **5007**, 73–86 (2003).

Eric C. Larson received his BS and MS in electrical engineering from Oklahoma State University, Stillwater, where he specialized in image processing and perception, in 2006 and 2008, respectively. His main research area is the analysis of psychovisual signal processing in computer systems, especially pertaining to human-computer interaction. He is also active in the areas of supported signal processing and machine learning for healthcare and environmental applications. He is active in signal processing education, and is a member of IEEE and HKN. He is currently pursuing his PhD in electrical and computer engineering at the University of Washington in Seattle.

Damon M. Chandler received the BS degree in biomedical engineering from The Johns Hopkins University, Baltimore, Maryland, in 1998; and the MEng, MS, and PhD degrees in electrical engineering from Cornell University, Ithaca, New York, in 2000, 2003, and 2005, respectively. From 2005 to 2006, he was a postdoctoral research associate in the Department of Psychology at Cornell. He is currently an assistant professor in the School of Electrical and Computer Engineering at Oklahoma State University, where he heads the image coding and analysis laboratory. His research interests include image processing, data compression, computational vision, natural scene statistics, and visual perception.