VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images

Damon M. Chandler, *Member, IEEE*, and Sheila S. Hemami, *Senior Member, IEEE*

***Abstract*—This paper presents an efficient metric for quantifying the visual fidelity of natural images based on near-threshold and suprathreshold properties of human vision. The proposed metric, the *visual signal-to-noise ratio* (VSNR), operates via a two-stage approach. In the first stage, contrast thresholds for detection of distortions in the presence of natural images are computed via wavelet-based models of visual masking and visual summation in order to determine whether the distortions in the distorted image are visible. If the distortions are below the threshold of detection, the distorted image is deemed to be of perfect visual fidelity ($\text{VSNR} = \infty$) and no further analysis is required. If the distortions are suprathreshold, a second stage is applied which operates based on the low-level visual property of perceived contrast, and the mid-level visual property of global precedence. These two properties are modeled as Euclidean distances in distortion-contrast space of a multiscale wavelet decomposition, and VSNR is computed based on a simple linear sum of these distances. The proposed VSNR metric is generally competitive with current metrics of visual fidelity; it is efficient both in terms of its low computational complexity and in terms of its low memory requirements; and it operates based on physical luminances and visual angle (rather than on digital pixel values and pixel-based dimensions) to accommodate different viewing conditions.**

***Index Terms*—Contrast, distortion, human visual system (HVS), image fidelity, image quality, noise, visual fidelity, wavelet.**

I. Introduction

THE rapid proliferation of digital imaging and communications technologies has given rise to a growing number of applications which yield images for end-use by humans. In many cases, the end-user receives a distorted version of the original digital image (e.g., due to lossy compression, digital watermarking, packet loss), and it is, therefore, necessary to quantify the visual impact of this distortion. To this end, visual fidelity metrics have been developed in an attempt to accurately and efficiently quantify the fidelity of a distorted image relative to the original image in a manner that agrees with subjective judgments made by humans.

Current approaches to quantifying the visual fidelity of a distorted image have primarily been based either on purely bottom-up properties of vision or on specific premises of what the human visual system attempts to accomplish when viewing a distorted image. These approaches can roughly be divided into the following paradigms:

1) *mathematically convenient metrics*, which operate based only on the intensity of the distortions, e.g., mean-squared error (MSE), peak signal-to-noise ratio (PSNR);
2) *metrics based on near-threshold psychophysics*, which typically employ a frequency-based decomposition, and which take into account the visual detectability of the distortions by using contrast detection thresholds (e.g., weighted MSE) and/or by also accounting for elevations in these thresholds due to masking effects imposed by the images (e.g., activity-based measures [1]–[3]);
3) *metrics based on overarching principles* such as structural or information extraction, which quantify visual fidelity based on the premise that a high-quality image is one whose structural content, such as object boundaries and/or regions of high entropy, most closely matches that of the original image [4]–[6] (see also [7]).

It is well-known that MSE/PSNR can be a poor predictor of visual fidelity [8]; rather, a veridical measure must take into account properties of the human visual system (HVS). Furthermore, in the context of image compression, several studies have argued that the visual detectability of the distortions does not always correlate with subjective fidelity ratings of images which contain suprathreshold (visible) distortions [9]–[11]. For example, an image containing clearly visible white noise can be judged to be subjectively more pleasing than an image containing, e.g., just-noticeable JPEG blocking artifacts. Consequently, the applicability of metrics based on near-threshold psychophysics for images containing suprathreshold distortions remains unclear. Although metrics based on overarching principles have demonstrated success at predicting visual fidelity for a variety of suprathreshold distortions, because these methods largely ignore properties of low-level vision, they generally cannot determine if the distortions are below the threshold of visual detection, nor is there a clear guideline for adjusting these metrics to account for different viewing conditions.

This paper presents a wavelet-based visual signal-to-noise ratio (VSNR) for quantifying the visual fidelity of distorted images based on recent psychophysical findings reported by the authors involving both near-threshold and suprathreshold distortions [12]–[15]. The proposed metric operates by using both low-level and mid-level properties of human vision: low-level HVS properties of contrast sensitivity and visual masking are first used via a wavelet-based model to determine if the distortions are below the threshold of visual detection. If

Manuscript received July 10, 2006; revised April 10, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ercan E. Kuruoglu.

D. M. Chandler is with Oklahoma State University, Stillwater, OK 74078 USA (e-mail: damon.chandler@okstate.edu).

S. S. Hemami is with Cornell University, Ithaca, NY 14853 USA (e-mail: hemami@ece.cornell.edu).

Digital Object Identifier 10.1109/TIP.2007.901820

the distortions are suprathreshold, the low-level HVS property of perceived contrast and the mid-level HVS property of global precedence (i.e., the visual system's preference for integrating edges in a coarse-to-fine-scale fashion) are taken into account as an alternative measure of structural degradation. The proposed metric thus estimates visual fidelity by computing: 1) contrast thresholds for detection of the distortions, 2) a measure of the perceived contrast of the distortions, and 3) a measure of the degree to which the distortions disrupt global precedence and, therefore, degrade the image's structure.

This paper is organized as follows. Section II reviews previous approaches to quantifying visual fidelity. Section III presents an interpretation of recent psychophysical results [13], [16] in the context of applying the results toward quantifying visual fidelity. Section IV describes the VSNR metric. The performance of the VSNR metric is evaluated, and limitations and extensions of the metric are discussed in Section V. General conclusions are provided in Section VI.

II. Previous Approaches to Quantifying Image Fidelity

This section reviews previous approaches to quantifying the visual fidelity of distorted images. The majority of these approaches can roughly be classified into the following paradigms: 1) mathematically convenient metrics, 2) metrics based on near-threshold psychophysics, and 3) metrics based on overarching principles.

A. Mathematically Convenient Metrics

Mean-squared error (MSE), and its closely related variant PSNR, have found widespread use due to their mathematical convenience. Metrics of this type generally operate based solely on the energy of the distortions. Specifically, let $\mathbf{I}$ denote an n-bit original digital image, and let $\hat{\mathbf{I}}$ 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). The distortions $\mathbf{E}$ are given by $\mathbf{E} = \hat{\mathbf{I}} - \mathbf{I}$ with digital pixel values in the range $[-2^n + 1, 2^n - 1]$. The MSE between $\mathbf{I}$ and $\hat{\mathbf{I}}$ is given by

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} E_i^2 = \frac{\|\mathbf{E}\|^2}{N} \tag{1}$$

and the PSNR (in decibels) is given by

$$\begin{aligned}\mathrm{PSNR} &= 10\log_{10}\left(\frac{(2^n-1)^2}{\mathrm{MSE}}\right)\\ &= 20\log_{10}(2^n-1) - 20\log_{10}\left(\frac{\|\mathbf{E}\|}{\sqrt{N}}\right)\end{aligned} \tag{2}$$

where $\|\cdot\|$ denotes the L_2-norm, E_i denotes the i^{th} digital pixel value of $\mathbf{E}$, and N denotes the total number of pixels. Thus, MSE is proportional to the energy of $\mathbf{E}$, and PSNR is proportional to the difference between $\log(2^n - 1)$ and the log energy of $\mathbf{E}$.

Despite its mathematical simplicity, MSE/PSNR oftentimes correlates poorly with subjective ratings of fidelity in natural images [8]. Part of the reason these metrics fail is the fact that they operate based on the *digital* values of $\mathbf{E}$ rather than on the *physical* luminances of the distortions ultimately emitted from the device on which the images are displayed. To overcome this limitation, a common practice in vision science is to first model the display device's relationship between digital pixel value P and physical luminance L given in $\mathrm{cd/m^2}$ via

$$L(P) = (b + kP)^{\gamma} \tag{3}$$

where b represents the black-level offset, k the pixel-value-to-voltage scaling factor, and γ the gamma of the display monitor [17]. The visibility of the distortions is then characterized based on the root-mean-squared (RMS) contrast [18]–[21] of the distortions, $C(\mathbf{E})$, which is given by

$$C(\mathbf{E}) = \frac{1}{\mu_{\mathbf{L}(\mathbf{I})}}\left(\frac{1}{N}\sum_{i=1}^{N}\left[L(E_i + \mu_{\mathbf{I}}) - \mu_{\mathbf{L}(\mathbf{E}+\mu_{\mathbf{I}})}\right]^2\right)^{1/2} \tag{4}$$

where $\mu_{\mathbf{I}} = (1/N)\sum_{i=1}^{N} I_i$ and $\mu_{\mathbf{L}(\mathbf{I})} = (1/N)\sum_{i=1}^{N} L(I_i)$ denote the average pixel value and average luminance of $\mathbf{I}$, respectively; and where $\mu_{\mathbf{L}(\mathbf{E}+\mu_{\mathbf{I}})} = (1/N)\sum_{i=1}^{N} L(E_i + \mu_{\mathbf{I}})$ denotes the average luminance of the mean-offset distortions $\mathbf{E} + \mu_{\mathbf{I}}$. The quantities $L(I_i) = (b + kI_i)^{\gamma}$ and $L(E_i + \mu_{\mathbf{I}}) = (b + k[E_i + \mu_{\mathbf{I}}])^{\gamma}$ correspond to the luminance of the i^{th} pixel of the image and the mean-offset distortions, respectively.

Thus, the RMS contrast of the distortions is the standard deviation of the luminances of $\mathbf{E} + \mu_{\mathbf{I}}$ normalized by the average luminance of $\mathbf{I}$. This normalization by $\mu_{\mathbf{L}(\mathbf{I})}$ attempts to account for Weber's Law, wherein the distortions are more difficult to see in brighter regions of an image as compared to in darker regions.

However, although RMS contrast accounts for the pixel-value-to-luminance response characteristics of the display device and for visual sensitivity to luminance, neither MSE/PSNR nor RMS contrast takes into account visual sensitivity to *contrast* and how contrast sensitivity is affected by the *image*. Thus, RMS contrast on its own is not a good predictor of the visual detectability of distortions contained in real images. These issues are discussed further in the following section.

B. Metrics Based on Near-Threshold Psychophysics

Perhaps the most widely used property of human vision for predicting visual fidelity is *contrast sensitivity*. Contrast sensitivity to a visual target (e.g., to the distortions in a distorted image) is defined as the inverse of the physical contrast of the target when the target is at the threshold of visual detection. Thus, in order for a human to visually detect a target, the contrast of the target must be greater than a certain *contrast detection threshold*, and the inverse of this threshold is contrast sensitivity. Whereas contrast sensitivity is traditionally measured by using targets consisting of sine-wave gratings, in the context of visual fidelity one is interested in contrast sensitivity to the *distortions* $\mathbf{E}$ when $\mathbf{E}$ is presented against the image (mask) $\mathbf{I}$ which is experiencing the distortions. If the contrast of the distortions is below the corresponding contrast detection threshold, the distortions are not visible, and, therefore, the distorted image $\hat{\mathbf{I}}$ is of perfect visual fidelity.

Numerous researchers have shown that contrast sensitivity to a target depends both on the spatial frequency of the target and on properties of the mask (image) on which the visual target is displayed [22]–[24]. Variations in sensitivity as a function of

spatial frequency is summarized via the well-known contrast sensitivity function (CSF) which has traditionally shown that sensitivity peaks between 1–6 cycles per degree of visual angle [13], [23]–[26]. Variations in sensitivity as a function of the mask upon which a target is displayed is also well known, and is typically termed (contrast or pattern) *masking*. In particular, we have shown in [12] and [13] that natural images impose both image-selective and spatial-frequency-selective masking of targets consisting of wavelet distortions. Visual angle, contrast sensitivity, and masking are discussed further in Section III.

A class of visual fidelity metrics has been developed which operates based on these low-level, near-threshold properties of vision (contrast sensitivity and masking) [1]–[3], [27]–[39] (see also [9] and [40]). Typically, the original and distorted images are processed through a set of subband filters to obtain oriented, spatial frequency decompositions of the images; this stage is designed to mimic the cortical decomposition performed by the HVS. The CSF is taken into account either by applying a CSF-shaped prefilter to the images prior to the subband decomposition, or by appropriately adjusting the relative gains of the subband filters based on their passbands. Masking is typically taken into account by means of a model which considers not only the characteristics of the target, but also the characteristics of the mask (image) on which the target is displayed [41]. The human visual responses to the original and distorted images are computed based on the magnitudes of these adjusted subband coefficients, and the distortions are deemed visible if these mimicked visual responses to the original and distorted images are sufficiently different from each other.

Visual fidelity metrics which operate based on near-threshold psychophysics are very effective at determining whether or not the distortions are visible and, therefore, whether or not the original and distorted images are visually distinguishable. However, because the underlying visual models have traditionally been developed and refined to fit contrast *detection* thresholds, the applicability of these models for clearly visible, suprathreshold distortions remains unclear. In the context of image compression, we have previously shown that detection thresholds, even after adjustments for masking, are of limited utility for low-rate compression in which the distortions are highly suprathreshold; rather, one must consider the perceived contrast of the distortions and the effects of the distortions on an observer's ability to visually process the image's edges (*global precedence*; see Section III-B) [11], [14].

C. Metrics Based on Overarching Principles

Recently, a different class of visual fidelity metrics has been developed which does not use low-level properties of vision, but instead operates based on overarching hypotheses of what the human visual system attempts to achieve when shown a distorted image [4]–[7]. The Structural SIMilarity (SSIM) metric of Wang *et al.* [4] operates based on the notion that the HVS has evolved to extract structural information from natural images, and, therefore, a high-quality image is one whose structure most closely matches that of the original. To this end, the SSIM metric employs a modified measure of spatial correlation between the pixels of the original and distorted images to quantify the extent to which the image's structure has been distorted. The SSIM metric has been shown to correlate well with subjective ratings of images containing a variety of suprathreshold distortions, and extensions of the metric have also been applied to video [42].

More recently, Sheikh *et al.* [5] proposed an information-theoretic approach to quantifying visual fidelity by means of an Information Fidelity Criterion (IFC) derived based on natural-scene statistics. The IFC metric operates under the premise that the HVS has evolved based on the statistical properties of the natural environment, and, therefore, given an original and distorted image, the visual fidelity of the distorted image can be quantified based on the amount of information it provides about the original. The IFC metric models images as realizations of a mixture of marginal Gaussian densities chosen for wavelet subband coefficients, and visual fidelity is quantified based on the mutual information between the coefficients of the original and distorted images. A more recent extension of the metric, Visual Information Fidelity (VIF) [43], which begins to incorporate properties of vision, has also been proposed by Sheikh *et al.*. As with the SSIM metric, the IFC and VIF metrics have been shown to perform well for a variety of suprathreshold distortions.

Metrics based on these overarching principles are particularly attractive due to their mathematical foundations which facilitates analysis and optimization. However, because these metrics do not consider the detectability of the distortions, the applicability of these metrics for determining whether or not a distorted image is of perfect visual fidelity remains unclear. Moreover, because these metrics largely ignore viewing conditions, in particular, viewing distance and the pixel-value-to-luminance response characteristics of typical display devices, it is not clear how the results of these metrics should be adjusted to account for different viewing conditions (see [44] for an extension of SSIM which attempts to account for viewing distance).

D. Summary and Other Approaches

In summary, numerous metrics have been developed to quantify the visual fidelity of distorted images. Mathematically convenient metrics such as MSE/PSNR, and luminance-based extensions of these metrics such as RMS contrast, do not generally correlate well with subjective judgments of fidelity. Metrics which take into account low-level, near-threshold properties of vision have proved useful for quantifying the detectability of the distortions, but the correct way to extend these metrics for use with clearly visible, suprathreshold distortions remains an open question [9]. Metrics based on overarching principles have shown great success for a variety of suprathreshold distortions; however, because these metrics largely ignore low-level properties of vision and viewing conditions, their applicability for images containing near-threshold distortions and/or images viewed under various viewing conditions remains unclear.

Some researchers have considered both low-level and higher-level properties of human vision toward quantifying visual fidelity. Classical experiments in multidimensional scaling have attempted to relate overarching properties to lower-level perception [45], [46]. The work of Karunasekera and Kingsbury [47] incorporates luminance and contrast masking into a measure of visual sensitivity to blocking (edge) artifacts. Similarly,

Miyahara *et al.* [34] combine a model of low-level vision with techniques for quantifying a variety of distortion factors (e.g., blocking, error correlation). The work of Osberger *et al.* [35] combines a model of low-level vision with an algorithm for predicting the most visually important spatial locations in images. The work of Damera-Venkata *et al.* [48] incorporates contrast sensitivity and luminance and contrast masking, and also considers suprathreshold contrast perception by using discrimination thresholds (as opposed to detection thresholds) measured for sine-wave gratings.[1] More recently, Carnec *et al.* [50] have proposed a metric which combines low-level HVS properties with a measure of structural information obtained via a stick-growing algorithm and estimates of visual fixation points. The work of Carnec *et al.* has been reported to correlate well with subjective ratings (see also [51]); however, the complexities of its algorithms impose hefty computational and memory requirements, even for moderately sized images.

In this paper, we propose a wavelet-based metric which combines the mathematical convenience of RMS contrast with both low-level and mid-level properties of vision by utilizing recent psychophysical results performed by the authors designed specifically to quantify the visual detectability and visual perception of distortions in natural images. The proposed metric, the VSNR, operates by modeling: 1) the average masking effects which natural images impose on the detectability of distortions, and 2) the perceived contrast of suprathreshold distortions, and 3) an alternative measure of structural degradation based on the mid-level visual property of global precedence (see Section III-B). The following section interprets the results of these psychophysical experiments in the context of applying the results toward quantifying visual fidelity. The resulting VSNR metric is described in Section IV.

III. Visual Detection and Perception of Distortion in Natural Images

This section reviews and interprets several key findings from recent psychophysical experiments performed by the authors to quantify the perception of distortions in natural images. We focus on three factors: 1) the detectability of distortions in natural images [12], [13], 2) the perceived contrast of suprathreshold distortions in natural images [14], and 3) the effects of suprathreshold distortions on the perceived fidelity of natural images [12], [14], [15]. Note that, because our psychophysical experiments utilized wavelet distortions generated by using the separable 9/7 filters [52], the models presented in this section rely on an octave-bandwidth wavelet decomposition of the image. Accordingly, the following discussion should not be considered a complete account of visual perception of distortions in images; rather, we present simplified models designed for computational efficiency.

Throughout this section, the term *spatial frequency* is used to describe the radial frequency content of visual stimuli expressed in units of cycles per degree of visual angle (cycles/degree). The visual angle provides a convenient means of specifying the size of a stimulus as it appears on the retina; 1 degree of visual angle covers approximately 290 μm of the retina [53]. Visual stimuli are also commonly characterized based on their orientation and spatial location; however, in this paper, we do not consider the separate effects of different orientations nor locations on visual fidelity.

[1]Whereas a contrast detection threshold specifies the minimum contrast required to visually detect a target, a contrast discrimination threshold specifies the minimum difference in contrast required to visually distinguish between two visual targets; see [49].

A. Visual Detection of Distortions

Visual Detectability and Masking of Distortions: The ability of a human observer to detect a visual target depends, among other factors, on the spatial frequency of the target. When displayed against a solid gray background, contrast sensitivity to sine-wave gratings is typically greatest (and, thus, the contrast detection threshold is lowest) for spatial frequencies between 4–6 cycles per degree of visual angle [23], [24]. Contrast sensitivity to Gaussian-modulated sine-waves and wavelets typically peaks around center frequencies of 1–3 cycles/degree [13], [25], [26]. Beyond these frequencies, sensitivity is reduced; thus, lower frequency and higher frequency targets are generally more difficult to detect. Contrast sensitivity is also known to vary with the background (mask) upon which targets are displayed. In general, it is more difficult to detect a target presented against a high-contrast mask than it is to detect a target presented against a lower-contrast mask [22]. In addition, variations in a mask's spatial pattern also affects sensitivity [41] (see [24] and [54] for a description of other parameters which affect contrast sensitivity; these parameters are not considered in this paper).

In the context of image fidelity, we are interested in the detectability of the distortions and the masking effects images impose upon this detectability [12], [13], [30]. Let $\mathbf{I}$ and $\hat{\mathbf{I}}$ denote, respectively, an original image and a distorted version of that image; and, let $\mathbf{E} \equiv \hat{\mathbf{I}} - \mathbf{I}$ denote the distortions contained within $\hat{\mathbf{I}}$. To determine whether the distortions are visible, contrast detection thresholds are computed as a function of both spatial frequency and the contrast of the image. Note that because our psychophysical experiments utilized wavelet distortions generated by using the separable 9/7 filters [52], the following equations rely on an octave-bandwidth wavelet decomposition of the image.

Step 1: Compute Threshold Contrast SNRs: Let $\mathbf{I}_f$ and $\mathbf{E}_f$ denote, respectively, the image and distortion content within an octave band of spatial frequencies centered at frequency f (here, f denotes the radial distance from zero-frequency in cycles/degree). The *contrast SNR* at f, $CSNR_f$, is defined as

$$CSNR_f(\mathbf{I}, \mathbf{E}) \equiv \frac{C(\mathbf{I}_f)}{C(\mathbf{E}_f)} \tag{5}$$

where $C(\mathbf{I}_f)$ and $C(\mathbf{E}_f)$ denote the respective RMS contrasts of these components.

In a previous study [12], we have shown that for detection of wavelet distortions in natural images, average contrast SNRs at the threshold of detection are modeled as

$$CSNR_f^{thr} = a_0 f^{a_2 \ln(f) + a_1} \tag{6}$$

where $CSNR_f^{thr}$ denotes the threshold contrast SNR for octave-bandwidth distortions centered at spatial frequency f and

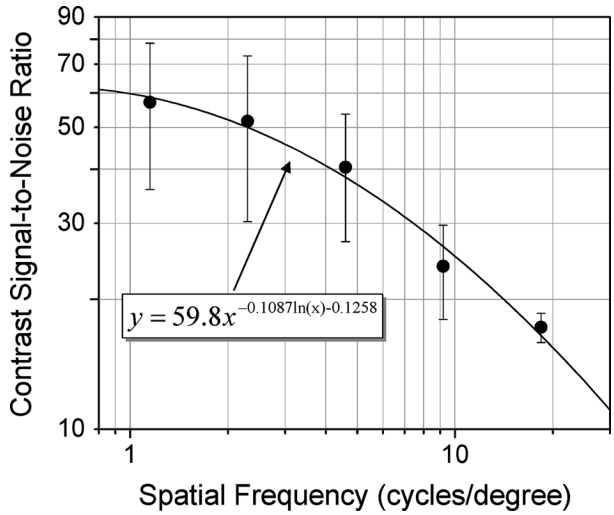

Fig. 1. Average threshold contrast SNRs from [12] for detection of wavelet distortions centered at spatial frequencies of 1.15, 2.3, 4.6, 9.2, and 18.4 cycles/degree in the presence of natural images. The horizontal axis denotes the center spatial frequency of the octave-bandwidth wavelet distortions; the vertical axis denotes contrast SNR. Data points denote threshold contrast SNRs averaged over the 14 natural images used in [12] and adjusted according to a linear model of visual summation (see [13]); error bars denote standard deviations of the means over images. The solid line represents (6).

presented against a typical natural image. This function represents a parabola in log-log coordinates whose offset and shape are defined by the parameters $a_0 = 59.8$, $a_1 = -0.1258$, and $a_2 = -0.1087$. These values were obtained based on the results for 14 natural images in [12], and were further adjusted to account for visual summation [13]. Fig. 1 depicts the average summation-adjusted threshold contrast SNRs from [12] along with (6); the error bars denote standard deviations of the means.

Step 2: Compute Contrast Detection Thresholds: The threshold contrast SNRs are used to compute corresponding contrast detection thresholds. Specifically, when $\mathbf{E}_f$ is at the threshold of detection when presented against an image mask $\mathbf{I}$, the contrast of $\mathbf{E}_f$, $C(\mathbf{E}_f) = CT(\mathbf{E}_f|\mathbf{I})$, where $CT(\mathbf{E}_f|\mathbf{I})$ denotes the contrast detection threshold of $\mathbf{E}_f$ computed based on the corresponding $CSNR_f^{thr}$ via

$$CT(\mathbf{E}_f|\mathbf{I}) = \frac{C(\mathbf{I}_f)}{CSNR_f^{thr}} = \frac{C(\mathbf{I}_f)}{a_0 f^{a_2 \ln(f)+a_1}}. \qquad (7)$$

Thus, to determine whether the distortions are visible within each octave band of frequencies, the actual contrast of the distortions, $C(\mathbf{E}_f)$, is compared with the corresponding contrast detection threshold $CT(\mathbf{E}_f|\mathbf{I})$. This process is described further in Section IV-B.

Note that the results depicted in Fig. 1 represent thresholds which have been averaged across images. Although this averaging operation allows (7) to implicitly account for masking induced by natural images and, therefore, provides a computationally efficient means of assessing the visibility of the distortions (see Section IV-B), the use of average thresholds can lead to an overestimate of contrast detection threshold for images (or regions of images) that are not particularly effective at masking the distortions; [55]–[57] describe techniques for determining RMS contrast thresholds on a per-image basis. In addition, (7) was developed based on psychophysical experiments using octave-bandwidth wavelet subband quantization distortions (using the separable 9/7 filters). Although it is generally accepted that the HVS analyzes visual stimuli via nearly octave-bandwidth spatial-frequency channels, the application of (7) to other types of distortion (e.g., geometric distortion) remains an open question.

B. Visual Perception of Suprathreshold Distortions

Perceived Contrast of Suprathreshold Distortions: Contrast detection thresholds and the corresponding CSF are crucial for determining whether or not the distortions in a distorted image are visible. However, the utility of these thresholds for images containing suprathreshold distortions in which the distortions are, by definition, visible, remains an area of debate [14], [58]–[60]. Although considerably less is known about suprathreshold vision, several researchers have investigated the perceived contrast of suprathreshold targets; i.e., how much contrast a spatial target *appears* to have. In particular, contrast-matching experiments have traditionally revealed that the perceived contrast of a suprathreshold target depends much less on the spatial frequency of the target than what is predicted by the CSF, a finding termed *contrast constancy* [58], [60]. Using a similar contrast-matching paradigm in the presence of natural images, we have also found contrast constancy for targets consisting of octave-bandwidth wavelet distortions [14]. These findings suggest that when distortions are visible, the perceived contrast of the distortions can be approximated by the physical contrast of the distortions; i.e., in terms of perceived contrast, there is little to no spatial-frequency dependence, and masking effects are, to a first approximation, negligible [14].

Global Precedence and the Effects of Distortions on Fidelity: In addition to the insights into low-level vision provided by these findings on perceived contrast, several researchers have argued that when the task is to judge the visual fidelity of a distorted image (as opposed to judging just the contrast of the distortions), humans tend to consider the effects of the distortions on the image's edges [4], [14], and, therefore, one must also consider higher-level properties of vision. Here, we take into account the mid-level property of *global precedence*, which states that the HVS visually integrates an image's edges in a coarse-to-fine-scale (global-to-local) fashion [61]–[63]. Physiological evidence for global precedence has recently been observed in secondary visual cortex (V2) of macaque when free-viewing natural scenes [64]; Willmore *et al.* found that V2 cells integrate activity across spatial frequency in an effort to enhance the representation of edges. In the context of image compression, we have advocated that the contrasts of suprathreshold quantization distortions should be proportioned across spatial frequency so as to preserve global precedence; specifically, because edges are visually integrated in a coarse-to-fine-scale order, the visual fidelity of an image can be maintained by preserving coarse scales at the expense of fine scales [11], [12], [14]. More recently, using a psychophysical scaling paradigm, we have shown that distortions which disrupt global precedence have much more of an impact on the visual fidelity of images than distortions which are spatially uncorrelated with the images [15].

Model of Global Precedence: The findings in [12] and [64] provide insights into global precedence; namely, these studies suggest that for suprathreshold distortions, the contrasts of distortions should be proportioned across spatial frequency so as to preserve global precedence and thereby attempt to maintain visual fidelity. Unfortunately, there are no data which explicitly specify a precise, computational model of global precedence. Here, we propose a descriptive model based on contrast SNRs; this is an improved version of our previous model [11] developed in the context of image compression which operated based only on distortion contrast.

Let $CSNR_f^*(\mathbf{E})$ denote the global-precedence-preserving contrast SNR for an octave band of spatial frequencies centered at frequency f. We then assume the following.

1) Based on contrast detection thresholds, for low-contrast distortions $CSNR_f^*(\mathbf{E})$ should approach $CSNR_f^{thr}$.
2) Based on global precedence, for increasingly suprathreshold distortions, $CSNR_f^*(\mathbf{E})$ corresponding to coarse scales (low f) should be increasingly greater than $CSNR_f^*(\mathbf{E})$ corresponding to fine scales (high f).

Under these assumptions, we propose the following model for $CSNR_f^*(\mathbf{E})$, which represents one possible relation which meets these two criteria, and which provides a mathematically simple transition from $CSNR_f^{thr}$ to $CSNR_f^*(\mathbf{E})$ as the distortion contrast is increased

$$CSNR_f^*(\mathbf{E}) = b_0(\mathbf{E}) f^{b_2(\mathbf{E})\ln(f)+b_1(\mathbf{E})} \tag{8}$$

where the parameters $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ are given by

$$\begin{aligned} b_0(\mathbf{E}) &= -a_0 v(\mathbf{E}) + a_0 \\ b_1(\mathbf{E}) &= (1.0 - a_1) v(\mathbf{E}) + a_1 \\ b_2(\mathbf{E}) &= (-1.0 - a_2) v(\mathbf{E}) + a_2 \end{aligned}$$

where $v(\mathbf{E}) \in [0,1]$ represents an index of visibility chosen such that the total RMS contrast of the distortions is $C(\mathbf{E})$ [see Appendix A for details on computing $v(\mathbf{E})$ given $C(\mathbf{E})$].

Together, the parameters $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ serve to adapt the shape of the CSNR curve [based on $v(\mathbf{E})$] so as to satisfy the two assumptions in the above list. The model presented here is based qualitatively on the results of [12]; namely, as the distortions become increasingly suprathreshold, coarser scales should have progressively greater SNRs than finer scales. For simplicity, $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ were chosen to be linear functions of $v(\mathbf{E})$; $b_0(\mathbf{E})$ and $b_2(\mathbf{E})$ decrease with $v(\mathbf{E})$, and b_1 increases with $v(\mathbf{E})$. Specifically, the parameter $b_0(\mathbf{E}) \in [0, a_0]$ effects a vertical shift in the CSNR curve [see Fig. 2(a)]. At $v(\mathbf{E}) = 0$, $b_0(\mathbf{E}) = a_0$, which corresponds to the same vertical shift as that used for $CSNR_f^{thr}$. As $v(\mathbf{E}) \to 1$, $b_0(\mathbf{E}) \to 0$, resulting in a downward vertical shift, which is equivalent to scaling the CSNR values (and, thus, scaling contrast thresholds) equally across the frequency range. Consequently, the parameter $b_0(\mathbf{E})$ alone does not take into account global precedence. The parameter $b_2(\mathbf{E}) \in [-1, a_2]$ effects a downward concave curvature in the CSNR curve [see Fig. 2(b)]. At $v(\mathbf{E}) = 0$, $b_2(\mathbf{E}) = a_2$, which corresponds to the same curvature as that used for $CSNR_f^{thr}$. As $v(\mathbf{E}) \to 1$, $b_2(\mathbf{E}) \to -1$, resulting in progressively downward concave curvature and, thus, progressively lower CSNR values for $f > 1.15$. The parameter $b_1(\mathbf{E}) \in [a_1, 1]$ was selected to regulate this curvature imposed by $b_2(\mathbf{E})$. At $v(\mathbf{E}) = 0$, $b_1(\mathbf{E}) = a_1$, which corresponds to the same curve as that used for $CSNR_f^{thr}$. As $v(\mathbf{E}) \to 1$, $b_1(\mathbf{E}) \to 1$, resulting in the proposed global-precedence-preserving CSNR curves depicted in Fig. 2(c). Note that these latter curves [in Fig. 2(c)] represent a compromise between decreasing all CSNR values equally [as in Fig. 2(a)] and effectively decreasing only the CSNR values for $f > 1.15$ [as in Fig. 2(b); $f = 1.15$ was the lowest spatial frequency tested in [12]]. We have found these proposed global-precedence-preserving CSNR curves to be particularly effective in image compression; in Section V, we demonstrate that this model is also effective for fidelity assessment.

Notice from Fig. 2(c) that as $v(\mathbf{E})$ is increased, i.e., as the distortions become increasingly suprathreshold, lower frequencies (coarser scales) have increasingly greater contrast SNRs than higher frequencies (finer scales) in order to preserve the visual integration of edges in a coarse-to-fine-scale fashion. Thus, together, the parameters $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ adapt the shape of the CSNR curve based on the total contrast of the distortions: for low-contrast distortions $CSNR_f^*(\mathbf{E}) \to CSNR_f^{thr}$; in particular, (6) is a special case of (8) for $v(\mathbf{E}) = 0$ (Assumption 1 in the above list). As $v(\mathbf{E}) \to 1$, $CSNR_f^*(\mathbf{E})$ corresponding to coarse scales is increasingly greater than $CSNR_f^*(\mathbf{E})$ corresponding to fine scales (Assumption 2 in the above list).

It is important to note that we are in no way claiming that (8) is a complete account of global precedence; rather, it represents a first step toward incorporating global precedence into a wavelet-based metric of visual fidelity. In particular, the parameters $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ represent one possible way to adapt the distribution of $CSNR_f^*$ across spatial frequency; the equations reported here were chosen based on their mathematical simplicity [linear functions in $v(\mathbf{E})$, and a parabola in log-frequency for (6)]. However, as we demonstrate in Section V, our simplified model performs quite well for a variety of commonly encountered distortions. As additional insight into global precedence and other mid- to higher-level properties of vision are made available, we expect more accurate models to be developed.

Also note that whereas $CSNR_f(\mathbf{I}, \mathbf{E})$ given by (5) specifies the *actual* contrast SNRs in the distorted image, $CSNR_f^*(\mathbf{E})$ given by (8) specifies the contrast SNRs selected in an attempt to preserve global precedence under the constraint that the distortions are at contrast $C(\mathbf{E})$. Thus, for a given distortion contrast $C(\mathbf{E})$, (8) specifies one technique of distributing the contrast of the distortions across spatial frequency in an attempt to preserve global precedence and thereby attempt to maintain visual fidelity within the context of the simplified model presented here. (We again stress that there are many other properties of vision which must be taken into account to truly maximize visual fidelity).

The following section describes an algorithm which estimates visual fidelity based on the extent to which the actual contrast SNRs deviate from these proposed, global-precedence-preserving contrast SNRs.

IV. VSNR: Visual SNR

Based on the psychophysical results presented in the previous section, this section presents a metric, the VSNR, which quan-

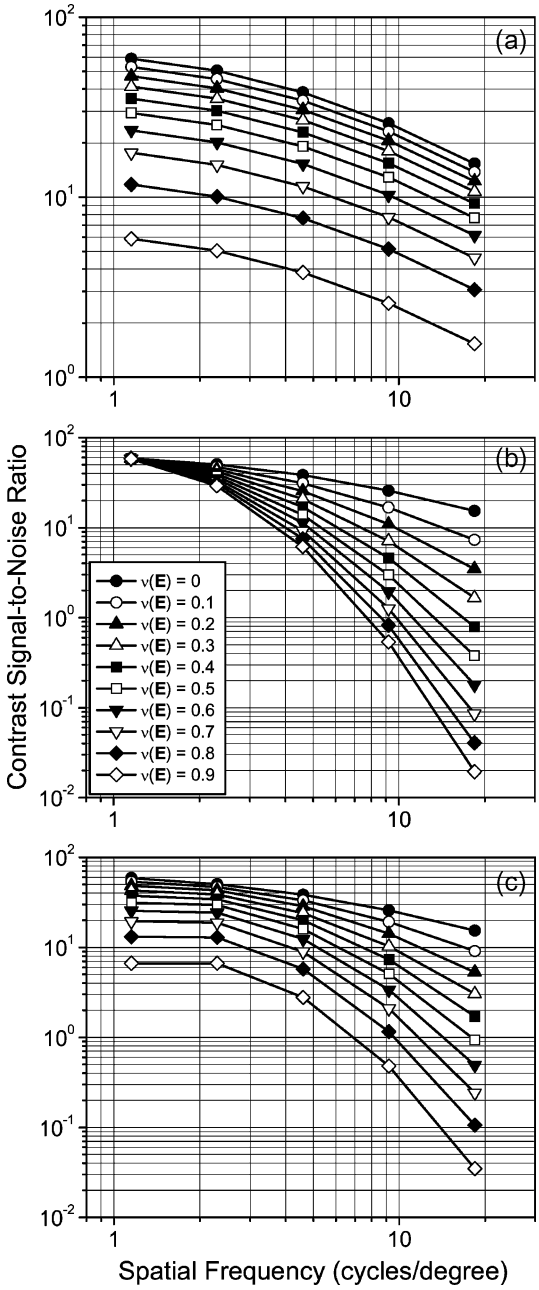

Fig. 2. Effects of the parameters $b_0(\mathbf{E})$, $b_1(\mathbf{E})$, and $b_2(\mathbf{E})$ on contrast SNRs computed via (8) for five octave frequency bands centered at spatial frequencies of 1.15, 2.3, 4.6, 9.2, and 18.4 cycles/degree. (a) Contrast SNRs obtained by varying only $b_0(\mathbf{E}) \in [0, a_0]$ (b_1, b_2 have been fixed at a_1, a_2, respectively), which results only in a vertical shift and, thus, does not take into account global precedence. (b) Contrast SNRs obtained by varying only $b_2(\mathbf{E}) \in [-1, a_2]$ (b_0, b_1 have been fixed at a_0, a_1, respectively), which results in progressively lower CSNR values for $f > 1.15$. (c) Proposed global-precedence-preserving contrast SNRs obtained by varying $b_0(\mathbf{E}) \in [0, a_0]$, $b_2(\mathbf{E}) \in [-1, a_2]$, and also $b_1(\mathbf{E}) \in [a_1, 1]$, which represents a compromise between the two extremes depicted in (a) and (b). In each graph, the horizontal axis denotes the center spatial frequency f, and the vertical axis denotes the corresponding contrast SNR. Each curve corresponds to contrast SNRs computed for a given visibility index $v(\mathbf{E})$ [see the legend in (b), which applies to all three graphs]. Note that for image compression applications, $C(\mathbf{E}_f) \leq C(\mathbf{I}_f)$, and, thus, $CSNR_f \geq 1$.

tifies the visual fidelity of distorted images. Given the original and distorted images $\mathbf{I}$ and $\hat{\mathbf{I}}$, the proposed metric, which operates for both near-threshold and suprathreshold distortions, estimates visual fidelity via two stages. In the first stage, contrast detection thresholds are computed as described in Section III-A. If the distortions are below the threshold of detection, the distorted image is deemed to be of perfect visual fidelity (VSNR $= \infty$), and then the algorithm terminates. If the distortions are suprathreshold, a second stage is applied which estimates visual fidelity based on a measure of perceived contrast and a measure of the extent to which the distortions disrupt global precedence.

Note that the proposed metric is limited by the fact that it does not take into account the spatial localization of distortion. Furthermore, the metric operates only on the luminance of the distortion and is, therefore, blind to chrominance distortion. See Section V-C for further limitations and possible extensions of the metric.

A. Preprocessing: Perceptual Decomposition and Viewing Conditions

To provide an approximation of the cortical decomposition performed by the HVS, an M-level separable discrete wavelet transform (DWT) using the 9/7 filters is performed on both $\mathbf{I}$ and $\mathbf{E}$ to obtain two sets of $3M + 1$ subbands: $\{\mathbf{s_I}\}$ and $\{\mathbf{s_E}\}$. Although the DWT is certainly not an ideal model of the decomposition performed by the HVS (cf [65]), the computational efficiency afforded by the DWT makes it particularly attractive for image analysis. Typically, $M = 5$ levels of decomposition provides a sufficient approximation; however, fewer levels may be used for smaller images. Note that the use of a separable wavelet transform limits the ability of our model to distinguish between image/distortion components oriented at 45° and those oriented at 135°.

Viewing conditions are taken into account by modeling the pixel-value-to-luminance response characteristics of the display device, and by considering the distance from which and the spatial resolution of the device on which the images are to be viewed. Thus, the algorithm requires the parameters b, k, and γ [see (3)], the spatial resolution of the display device, and the intended viewing distance. A vector of octave-spaced frequencies $\mathbf{f} = [f_1, f_2, \ldots, f_M]$, in cycles/degree, is then computed based on the viewing distance and the resolution of the display via

$$f_m = 2^{-m} r v \tan\left(\frac{\pi}{180}\right) \tag{9}$$

$m = 1, 2, \ldots, M$, where r denotes the resolution of the display in pixels per unit distance (e.g., pixels/in), and v is the viewing distance expressed in the corresponding units of distance (e.g., inches); see [26] and [54].

B. Stage 1: Assess the Detectability of the Distortions

To determine whether the distortions in $\hat{\mathbf{I}}$ are visible, contrast thresholds for detection of the distortions within each band centered at f_m are computed as described in Section III-A. These contrast thresholds are then compared with the actual contrasts of the distortions within the band centered at f_m in the distorted image. Specifically, the following steps are performed.

1) For each f_m in $\mathbf{f}$, compute the contrast detection threshold $CT(\mathbf{E}_{f_m}|\mathbf{I})$ via (7).

2) For each f_m in $\mathbf{f}$, measure the actual contrast of the distortions $C(\mathbf{E}_{f_m})$ via the approximation

$$C(\mathbf{E}_{f_m}) \approx \frac{k\gamma}{2^m \mu_{\mathbf{L(I)}}(b+k\mu_{\mathbf{I}})^{1-\gamma}} \times \sqrt{\sigma^2\left[s_{\mathbf{E}(m,LH)}\right] + \sigma^2\left[s_{\mathbf{E}(m,HL)}\right] + \sigma^2\left[s_{\mathbf{E}(m,HH)}\right]} \quad (10)$$

where $\sigma[s_{\mathbf{E}(m,\theta)}]$ denotes the standard deviation of the subband of $\mathbf{E}$ at the m^{th} level of decomposition with orientation $\theta = LH$, HL, or HH (see Appendix of [11] for derivation).[2]

Due to the separability of the DWT filters, the HH band can be considered to correspond to spatial frequency content centered at $\sqrt{2}f_m$. However, for simplicity, we have included the HH band in (10) which our tests have revealed to yield results similar to those obtained by considering the HH band to be centered at the higher spatial frequency. [This rationale also applies to (13), presented later in this paper, which is used to measure the contrast of the image $\mathbf{I}_{f_m}$]. Also note that, because a major goal of our metric is computational efficiency, (10) combines $\sigma^2[s_{\mathbf{E}(m,\theta)}]$ for all orientations ($\theta = LH$, HL, HH) at decomposition level m. Although there is no general tenet regarding how one should combine the results across orientation for natural-image stimuli, our previous results in [13], [16] using distortion of LH and HL subbands suggest that visual summation across orientation is reasonably well-modeled, to a first approximation, as specified by (10) when distortions are presented against natural images.

Upon completion of these steps, if $C(\mathbf{E}_{f_m}) < CT(\mathbf{E}_{f_m}|\mathbf{I})$, $\forall f_m \in \mathbf{f}$, the distortions are below the threshold of visual detection, $\hat{\mathbf{I}}$ is visually indistinguishable from $\mathbf{I}$, and, therefore, $\hat{\mathbf{I}}$ is deemed to be of perfect visual fidelity. In this case, VSNR $= \infty$, and no further analysis is required.

If the distortions are suprathreshold, a second stage (described next) is applied to compute VSNR based on the perceived contrast of the distortions and the extent to which the distortions disrupt global precedence.

C. Stage 2: Compute the Visual SNR

To compute the finite VSNR when the distortions in $\hat{\mathbf{I}}$ are suprathreshold, we consider the low-level HVS property of perceived contrast, and the mid-level HVS property of global precedence described in Section III-B. These properties are take into account as follows.

1) Let d_{pc} denote a measure of the perceived contrast of the distortions. As mentioned in Section III-B, for suprathreshold distortions, perceived contrast is relatively invariant with spatial frequency (contrast constancy). Accordingly, we approximate the perceived contrast of the distortions by the total RMS distortion contrast; thus

$$d_{pc} = C(\mathbf{E}) \quad (11)$$

where $C(\mathbf{E})$ is measured via (4).

2) Let d_{gp} denote a measure of the extent to which global precedence has been disrupted. To quantify this effect, the contrast of the distortions within each band centered at f_m is compared with the corresponding global-precedence-preserving contrast, $C^*(\mathbf{E}_{f_m})$, computed via

$$C^*(\mathbf{E}_{f_m}) = \frac{C(\mathbf{I}_{f_m})}{CSNR^*_{f_m}(\mathbf{E})} \quad (12)$$

where $CSNR^*_{f_m}(\mathbf{E})$ is the global-precedence-preserving contrast SNR given by (8); and where $C(\mathbf{I}_{f_m})$ is approximated via

$$C(\mathbf{I}_{f_m}) \approx \frac{k\gamma}{2^m \mu_{\mathbf{L(I)}}(b+k\mu_{\mathbf{I}})^{1-\gamma}} \times \sqrt{\sigma^2\left[s_{\mathbf{I}(m,LH)}\right] + \sigma^2\left[s_{\mathbf{I}(m,HL)}\right] + \sigma^2\left[s_{\mathbf{I}(m,HH)}\right]} \quad (13)$$

where $\sigma[s_{\mathbf{I}(m,\theta)}]$ denotes the standard deviation of the subband of $\mathbf{I}$ at the m^{th} level of decomposition with orientation $\theta = LH$, HL, or HH (see Appendix of [11] for derivation).[2] The quantity d_{gp} is then computed via

$$d_{gp} = \left(\sum_{m=1}^{M}\left[C^*(\mathbf{E}_{f_m}) - C(\mathbf{E}_{f_m})\right]^2\right)^{1/2} \quad (14)$$

where $C(\mathbf{E}_{f_m})$ denotes the actual contrast of the distortions within the band centered at f_m computed in Stage 1.

Geometrically, the quantities d_{pc} and d_{gp} can be interpreted as follows: Let $\mathbf{C}(\mathbf{E_f}) = [C(\mathbf{E}_{f_1}), C(\mathbf{E}_{f_2}), \cdots, C(\mathbf{E}_{f_M})]$ denote a vector of actual distortion contrasts, and let $\mathbf{C}^*(\mathbf{E_f}) = [C^*(\mathbf{E}_{f_1}), C^*(\mathbf{E}_{f_2}), \cdots, C^*(\mathbf{E}_{f_M})]$ denote a vector of global-precedence-preserving distortion contrasts. Thus, $\mathbf{C}(\mathbf{E_f})$ and $\mathbf{C}^*(\mathbf{E_f})$ can be viewed as points in an M-dimensional space in which each axis corresponds to distortion contrast in the band centered at f_m, $m = 1, 2, \ldots, M$. The quantity d_{pc}, which is given by the total RMS contrast of the distortions, is approximately the distance from the origin.[3] The quantity d_{gp} is the distance between the actual contrasts of the distorted image and the global-precedence-preserving contrasts computed for the same total RMS distortion contrast.

This geometric interpretation is illustrated in Fig. 3 for $M = 2$ dimensions. The horizontal axis corresponds to the distortion contrast in f_1, and the vertical axis corresponds to the distortion contrast in f_2. The points $\mathbf{C}(\mathbf{E_f})$ and $\mathbf{C}^*(\mathbf{E_f})$ are denoted by open and closed circles, respectively. Note that both of these points lie on the contour of constant contrast indicated by the dotted line; i.e., they are equidistant from the origin as defined by (8). The quantity d_{pc} is approximately this distance from the origin. The quantity d_{gp} is the distance between the two points; thus, in general, $d_{gp} \in [0, \sqrt{2}d_{pc}]$. Note that MSE is a pixel-value-based analogue of d_{pc}.

[2]Note that one may alternatively compute $C(\mathbf{E}_{f_m})$ via (4); however, such an approach requires performing an inverse DWT to obtain a spatial-domain representation of $\mathbf{E}_{f_m}$. In practice, the approximation in (10) performs quite well. It constitutes a percent relative error ($|actual - approximated|/actual \times 100\%$) of approximately 0.05% [11].

[3]The total RMS contrast of the distortions $C(\mathbf{E})$ is equivalent to the L_2-norm of the vector $\mathbf{C}(\mathbf{E_f})$ if the display device has $\gamma = 1$ and if the contrast of the distortions in the LL subband is zero. In practice, $C(\mathbf{E})$ is very closely approximated by $\|\mathbf{C}(\mathbf{E_f})\|$ over the range of values of b, k, and γ found in most display monitors

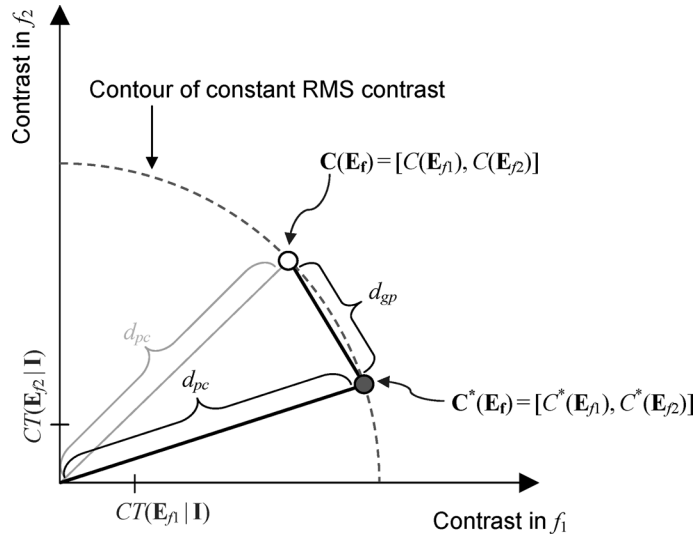

Fig. 3. Geometric illustration of the quantities d_{pc} and d_{gp} for $M = 2$ dimensions. The horizontal and vertical axes in this space correspond to the distortion contrasts in bands centered at f_1 and f_2, respectively. The open circle represents the actual distortion contrasts $\mathbf{C}(\mathbf{E_f})$ contained in a distorted image. The closed circle represents the global-precedence-preserving distortion contrasts $\mathbf{C}^*(\mathbf{E_f})$ computed via (12). The quantity d_{pc} corresponds to the distance from the origin, and the quantity d_{gp} corresponds to the distance between the two points. Whereas MSE is a pixel-value-based analogue of d_{pc}, the proposed measure of visual distortion is based on a linear combination of both d_{pc} and d_{gp}. Note that contrast detection thresholds, which are labeled as $CT(\mathbf{E}_{f_1}|\mathbf{I})$ and $CT(\mathbf{E}_{f_2}|\mathbf{I})$, are crucial for determining whether the distortions are visible; however, the utility of these thresholds for determining visual distortion for points in this space beyond the thresholds remains unclear.

We define *visual distortion*, VD, as the following linear combination of d_{pc} and d_{gp}

$$VD = \alpha d_{pc} + (1-\alpha)\frac{d_{gp}}{\sqrt{2}} \tag{15}$$

where the parameter $\alpha \in [0,1]$ determines the relative contribution of each distance (see Section V); and where d_{gp} is normalized by $\sqrt{2}$ so that $d_{gp}/\sqrt{2} \in [0, d_{pc}]$, and, thus, $VD \in [0, d_{pc}]$. Note that (15) includes both d_{pc} and d_{gp}. The quantity d_{pc} is needed to account for differences in perceived fidelity when two images have different total distortion contrasts, but both images have $d_{gp} = 0$. If two images have $d_{gp} = 0$, the image with the greater total distortion contrast (d_{pc}) will generally be rated lower in perceived fidelity (assuming the additional distortion contrast is visible). The quantity d_{pc} is needed to account for this condition.

The VSNR, in decibels, is accordingly given by

$$\text{VSNR} = 10\log_{10}\left(\frac{C^2(\mathbf{I})}{VD^2}\right) = 20\log_{10}\left(\frac{C(\mathbf{I})}{\alpha d_{pc} + (1-\alpha)\frac{d_{gp}}{\sqrt{2}}}\right) \tag{16}$$

where $C(\mathbf{I})$ denotes the RMS contrast of the original image $\mathbf{I}$ given by $C(\mathbf{I}) = \sigma_{\mathbf{L(I)}}/\mu_{\mathbf{L(I)}}$. Note that when global precedence is maximally disrupted for a given $C(\mathbf{E})$, at most $d_{gp} = \sqrt{2}d_{pc}$, $VD = d_{pc}$, and, thus, $\text{VSNR} = 20\log_{10}(C(\mathbf{I})/d_{pc}) = 20\log_{10}(C(\mathbf{I})/C(\mathbf{E}))$; in this case, the visual SNR is given by the (scaled, log) contrast SNR of $\mathbf{I}$ to $\mathbf{E}$.

D. Summary of the VSNR Metric

In summary, given an original image $\mathbf{I}$ and a distorted version of the image $\hat{\mathbf{I}}$, the VSNR metric is computed via the following steps.

1) *Preprocessing*.
   a) Compute the distortions $\mathbf{E} = \hat{\mathbf{I}} - \mathbf{I}$.
   b) Perform M-level DWTs of $\mathbf{I}$ and $\mathbf{E}$ to obtain subbands $\{\mathbf{s_I}\}$ and $\{\mathbf{s_E}\}$.
   c) Compute the vector of spatial frequencies $\mathbf{f}$ via (9).
2) *Stage 1: Assess the Detectability of the Distortions*
   a) For each f_m in $\mathbf{f}$, compute the contrast detection threshold $CT(\mathbf{E}_{f_m}|\mathbf{I})$ via (7).
   b) For each f_m in $\mathbf{f}$, measure the actual distortion contrast $C(\mathbf{E}_{f_m})$ via (10).
   c) If $C(\mathbf{E}_{f_m}) < CT(\mathbf{E}_{f_m}|\mathbf{I})$, $\forall m$, then $\text{VSNR} = \infty$, and then terminate.
3) *Stage 2: Compute the VSNR*
   a) Compute the perceived contrast of the distortions $d_{pc} = C(\mathbf{E})$ via (4).
   b) Compute the disruption of global precedence d_{gp} via (14).
   c) Compute $\text{VSNR} = 20\log_{10}(C(\mathbf{I})/(\alpha d_{pc} + (1-\alpha)(d_{gp}/\sqrt{2}))$.

The following section examines the performance of the VSNR metric on a variety of distortion types and natural images for which subjective fidelity ratings have been measured; the section also discusses the computational and memory requirements, and limitations and extensions of the metric.

V. Results and Analysis

In this section, the performance of the VSNR metric is analyzed in terms of its ability to predict fidelity in a manner that agrees with subjective ratings, and in terms of its computational and memory requirements. To assess its predictive performance, the VSNR metric was applied to images from the LIVE image database [66] for which subjective ratings are available. For comparison, these same sets of images were analyzed by using the following metrics: 1) *PSNR*; 2) *WSNR*, a weighted SNR in which the original and distorted images were filtered

by the CSF specified in [27], [67] (see [48] and [68] for additional details and code, respectively); 3) Universal Quality Index *(UQI)* of Wang *et al.* [69]; 4) Noise Quality Measure *(NQM)* of Damera-Venkata *et al.* [48]; 5) *SSIM* metric of Wang *et al.* [4]; 6) *VIF* metric of Sheikh *et al.* [43].

For all images, the VSNR metric was computed using $\alpha = 0.04$[4] and assuming sRGB display characteristics [70], a display resolution of 96 pixels/in, and a viewing distance of 19.1 in (approximately 3.5 picture heights); these parameters provide a reasonable approximation of typical viewing conditions. For an sRGB display monitor, the pixel-value-to-luminance parameters are $b = 0, k = 0.02874$, and $\gamma = 2.2$; and, a display resolution of 96 pixels/in and a viewing distance of 19.1 in yields spatial frequencies $\mathbf{f} = [1, 2, 4, 8, 16]$ cycles/degree for a five-level DWT. The WSNR and NQM metrics were computed by using this same display resolution and viewing distance. The PSNR was computed directly via (2). The SSIM metric was computed on filtered and downsampled versions of the images where the downsampling factor (typically four) was chosen based on the height of each image as described at [71]. The UQI and VIF metrics were applied using their default implementations provided at [72] and [73], respectively.

A. Performance in Predicting Visual Fidelity

The LIVE image database represents the largest collection of readily available distorted images for which subjective ratings of perceived distortion have been measured [66]. The database consists of 29 original 24-bits/pixel color images, and 779 distorted images. Five types of distortions were tested: 1) JPEG-2000 compression, 2) JPEG compression, 3) Gaussian white noise; 4) Gaussian blurring, and 5) Rayleigh-distributed bitstream errors of a JPEG-2000 compressed stream. Further details of the LIVE database are described in [43].

The seven metrics, PSNR, WSNR, UQI [69], NQM [48], SSIM [4], VIF [43], and VSNR, were applied to grayscale versions of the images which were obtained via a pixel-wise transformation of $I = 0.2989R + 0.5870G + 0.1140B$, where I, R, G, and B denote the 8-bit grayscale, red, green, and blue intensities, respectively. For each metric, a logistic function of the form [74]

$$f(x) = \frac{\tau_1 - \tau_2}{1 + e^{\frac{x-\tau_3}{\tau_4}}} + \tau_2 \tag{17}$$

was fitted to the data via a Nelder–Mead search [75] to obtain the parameters τ_1, τ_2, τ_3, and τ_4 which minimized the sum-squared error between the transformed metric outputs $\{f(x)\}$ and the corresponding subjective ratings.

The proposed VSNR metric is generally competitive with the other metrics in terms of prediction accuracy on the LIVE database. Fig. 4 depicts graphs of subjective ratings of perceived distortion plotted against the transformed metric outputs, and Fig. 5 depicts separate fits of the VSNR metric to results for each of the five distortion types. Clearly, for these images, the VIF metric outperforms the other metrics. However, because images containing different distortion types were tested in separate experiments, the subjective ratings for different distortion types are not directly comparable; the results listed for the LIVE database taken as a whole (all 779 images) are provided only for [76]. The correlation coefficients, RMS errors, and rank-order correlation coefficients are listed in Table I; these data were computed and are listed both for the database as a whole (779 images) and for the images containing each of the separate distortion types.

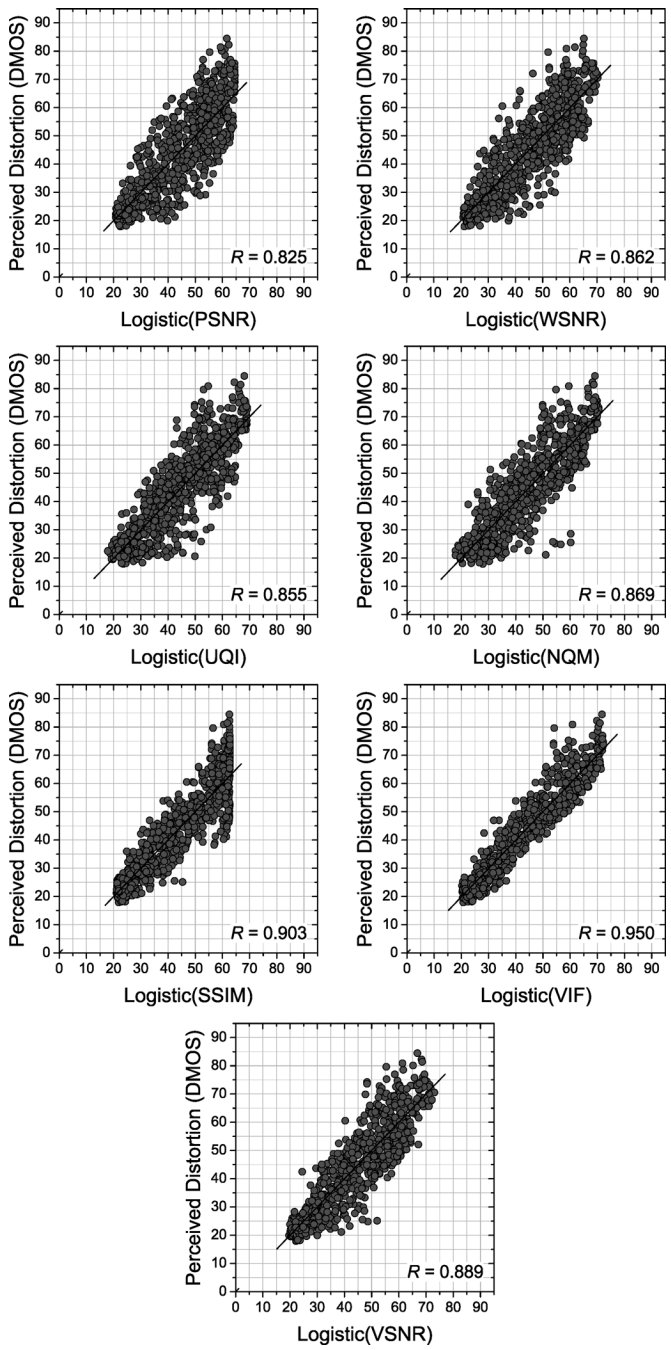

Fig. 4. Subjective ratings of perceived distortion for the 779 images of the LIVE [66] database plotted against predicted values from each of the seven metrics. In all graphs, the vertical axis denotes perceived distortion (difference mean opinion score) as reported by subjects. The horizontal axes correspond to metric outputs transformed via (17).

[4]Although the correct perceptual contributions of perceived contrast and disruption of global precedence is an area of further investigation (e.g., see [15] and [45]), we have found $\alpha = 0.04$ to provide reasonable fits to subjective rating data.

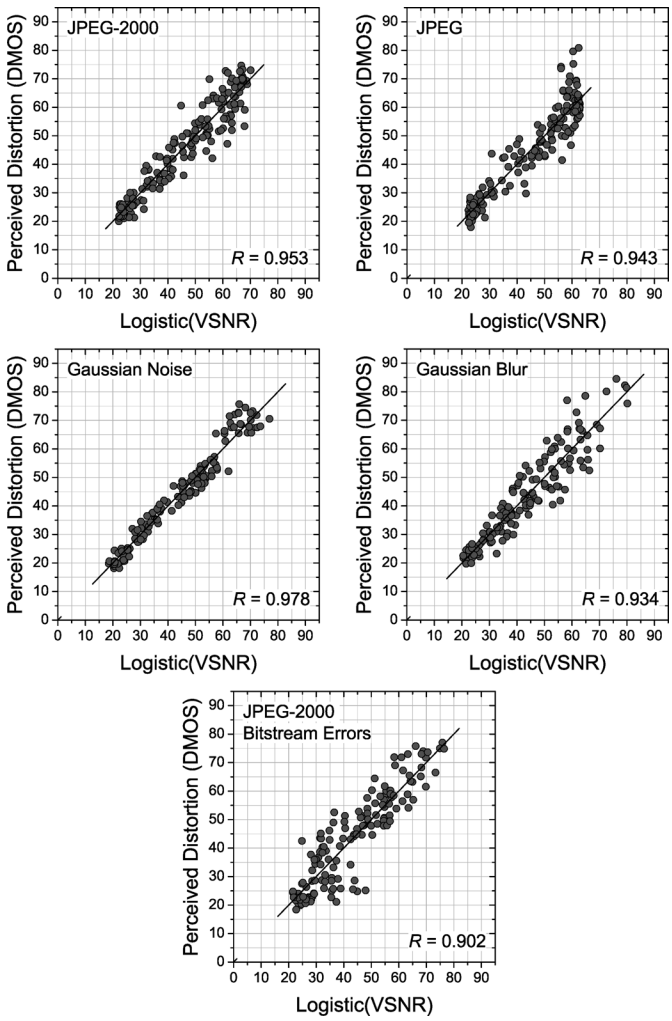

Fig. 5. Subjective ratings of perceived distortion for images of the LIVE [66] database plotted against transformed VSNR fitted separately for each distortion type. In all graphs, the vertical axis denotes perceived distortion (difference mean opinion score) as reported by subjects. The horizontal axes correspond to VSNRs transformed via (17) fitted separately for each distortion type.

To assess the statistical significance of each metric's performance relative to the other metrics, an F-test was performed on the prediction errors [74]. Specifically, if the prediction errors (residuals) are assumed to be distributed according to a Gaussian distribution, an F-test can be used to assess whether the residuals from two metrics correspond to the same population and can, thus, be used to determine if one metric has significantly larger residuals (greater prediction error) than another metric (see [74] and [77]). Let σ_A^2 and σ_B^2 denote the variance of the residuals from metrics A and B, respectively. The F statistic is given by $F = \sigma_A^2/\sigma_B^2$. Values of $F > F_{\text{critical}}$ (or $F < 1/F_{\text{critical}}$) signify that, at a given confidence level, metric A has significantly larger (or smaller) residuals than metric B, where F_{critical} is computed based on the number of residuals and the confidence level.

The lower half of Table I lists the F statistic resulting from an F-test performed on the residuals from each metric versus the residuals from VSNR. Also listed in Table I is the sample skewness and kurtosis of each metric's residuals, which can be used to gauge the Gaussianity of the residuals (the Gaussian distribution has a skewness of zero and a kurtosis of 3; commonly, kurtosis values between 2–4 are deemed Gaussian [74], though see [78] for a more formal test using the JB statistic). Values of $F > F_{\text{critical}}$ (or $F < 1/F_{\text{critical}}$), which are shown in boldface, signify that with 99% confidence the metric has significantly larger (or smaller) residuals than VSNR on the corresponding set of images. Thus, in terms of statistical significance, the VSNR metric generally outperforms PSNR, WSNR, UQI, and NQM, is competitive with SSIM for all but the Rayleigh-distributed bit-stream errors, and is competitive with VIF for images containing JPEG-2000, JPEG, and white noise distortions.

B. Computational Complexity and Memory Requirements

Two primary concerns which often dictate the use of a particular visual fidelity metric over another are the computational requirements and the memory requirements of the metric. In both of these regards, the proposed VSNR metric is competitive with previous approaches.

1) Computational Complexity: The VSNR metric has relatively low computational complexity. Using a basic C++ implementation, the metric requires for a 512×512 image approximately 0.7 s on a 2.5-GHz Intel Celeron machine. The two steps which require the bulk of computation are: 1) performing DWTs of the original image $\mathbf{I}$ and distortions $\mathbf{E}$ and 2) computing the variances of the subbands [for use with (10) and (13)] and the RMS contrasts of $\mathbf{I}$ and $\mathbf{E}$. However, these steps too are readily computed: As mentioned in Section IV-A, the subbands are obtained via a separable DWT using the 9/7 filters, for which efficient, lifting-based implementations are available; see also [79] (the source code for the VSNR metric, available at [83], provides a lifting-based DWT implementation). In addition, because the subbands are critically sampled, and, thus, the total number of DWT coefficients is equivalent to the number of pixels, computing the variances of each of the subbands requires approximately the same number of operations as computing the variance of the image's pixels. Furthermore, the RMS contrasts of $\mathbf{I}$ and $\mathbf{E}$ can be efficiently computed via a lookup table which maps pixel values to luminances (see [83]).

2) Memory Requirements: The VSNR metric also has low memory requirements. For N-pixel 8-bits/pixel images $\mathbf{I}$ and $\hat{\mathbf{I}}$, in the worst-case scenario in which all data are loaded concurrently, the metric requires approximately $12N$ bytes: $2N$ bytes to hold $\mathbf{I}$ and $\hat{\mathbf{I}}$, $2N$ bytes to hold a signed representation of $\mathbf{E}$, an additional $8N$ bytes to hold the (floating point) subbands of these images $\{\mathbf{s_I}\}$ and $\{\mathbf{s_E}\}$ (and a negligible amount of memory to hold the computed scalar quantities and the pixel-value-to-luminance lookup table). However, in practice, $\hat{\mathbf{I}}$ is required only long enough to compute $\mathbf{E}$; and $\mathbf{I}$ and $\mathbf{E}$ can be analyzed separately, which requires at most $2N$ bytes at any given time. In addition, only one subband of $\{\mathbf{s_I}\}$ or $\{\mathbf{s_E}\}$ is needed at any given time (to compute the variance of that subband), which requires at most $(1/4) \times 4N = N$ bytes (assuming the subband is from the first level of decomposition in which each band contains $(1/4)N$ coefficients). Thus, an efficient implementation would require at most approximately $2N$ bytes at any given time (e.g., 512 KB for a 512×512 image).

TABLE I
CORRELATION COEFFICIENT, RMSE, AND RANK-ORDER CORRELATION COEFFICIENT BETWEEN SUBJECTIVE RATINGS FOR THE LIVE DATABASE AND TRANSFORMED METRIC OUTPUTS. ALSO LISTED IS THE F STATISTIC FOR EACH METRIC'S RESIDUALS TESTED AGAINST VSNR'S RESIDUALS, AND THE SAMPLE SKEWNESS AND KURTOSIS OF EACH METRIC'S RESIDUALS (ITALICIZED VALUES DENOTE NON-GAUSSIAN RESIDUALS BASED ON THE JB STATISTIC [78]). VALUES OF $F > F_{\text{critical}}$ (OR $F < 1/F_{\text{critical}}$) SHOWN IN BOLDFACE SIGNIFY THAT WITH 99% CONFIDENCE THE METRIC HAS SIGNIFICANTLY LARGER (OR SMALLER) RESIDUALS THAN VSNR

Measure	Metric	All	JPEG-2000	JPEG	White Noise	Blurring	JP2 Bit Errors
Correlation Coefficient	PSNR	0.825	0.896	0.860	0.986	0.784	0.892
	WSNR	0.862	0.896	0.899	0.970	0.859	0.912
	UQI	0.855	0.842	0.844	0.936	0.945	0.944
	NQM	0.869	0.926	0.909	0.986	0.903	0.825
	SSIM	0.903	0.956	0.943	0.970	0.945	0.948
	VIF	0.950	0.957	0.923	0.982	0.975	0.961
	VSNR	0.889	0.953	0.943	0.978	0.934	0.902
RMSE	PSNR	9.127	7.281	8.253	2.729	9.907	7.539
	WSNR	8.192	7.264	7.071	3.933	8.153	6.832
	UQI	8.379	8.838	8.687	5.691	5.208	5.523
	NQM	7.999	6.179	6.749	2.709	6.867	9.434
	SSIM	6.946	4.797	5.399	3.954	5.232	5.325
	VIF	5.042	4.735	6.183	3.067	3.552	4.564
	VSNR	7.390	4.963	5.399	3.399	5.692	7.193
Rank Order Correlation Coefficient	PSNR	0.820	0.889	0.841	0.985	0.781	0.893
	WSNR	0.863	0.893	0.877	0.968	0.861	0.915
	UQI	0.854	0.840	0.821	0.909	0.938	0.932
	NQM	0.872	0.919	0.880	0.984	0.874	0.802
	SSIM	0.900	0.952	0.911	0.969	0.951	0.955
	VIF	0.953	0.953	0.913	0.985	0.973	0.965
	VSNR	0.889	0.946	0.908	0.979	0.941	0.906
Fcritical	—	1.182	1.434	1.425	1.476	1.476	1.476
1/Fcritical	—	0.846	0.697	0.702	0.677	0.677	0.677
F Statistic	PSNR	**1.525**	**2.153**	**2.337**	**0.645**	**3.030**	1.099
	WSNR	**1.229**	**1.710**	**1.470**	1.391	1.431	0.886
	UQI	**1.286**	**3.172**	**2.589**	**2.805**	0.837	**0.590**
	NQM	1.172	**1.551**	**1.563**	**0.635**	1.456	**1.720**
	SSIM	0.884	0.934	1.000	1.354	0.845	**0.548**
	VIF	**0.467**	0.922	1.327	0.826	**0.395**	**0.408**
	VSNR	1	1	1	1	1	1
Skewness / Kurtosis	PSNR	-0.10 / 2.66	0.30 / 3.22	0.15 / 3.41	-0.11 / 2.96	-0.30 / 2.84	-0.05 / 3.08
	WSNR	-0.21 / 3.36	*-0.16 / 2.99*	-0.70 / 3.51	*0.00 / 2.23*	*-0.33 / 3.04*	*0.39 / 3.46*
	UQI	*0.06 / 3.62*	*0.60 / 3.91*	0.01 / 3.64	0.10 / 2.70	0.13 / 3.09	*0.60 / 5.25*
	NQM	*0.04 / 4.26*	0.14 / 2.91	*-0.47 / 3.90*	0.34 / 3.25	-0.17 / 2.59	*0.73 / 3.64*
	SSIM	0.04 / 3.42	-0.12 / 2.97	*-0.62 / 4.65*	-0.06 / 2.14	0.10 / 2.92	0.00 / 2.86
	VIF	*-0.97 / 4.64*	-0.36 / 2.93	*-1.28 / 4.64*	-0.12 / 2.71	0.17 / 2.62	-0.48 / 3.24
	VSNR	*-0.21 / 3.45*	0.08 / 3.94	*-0.92 / 5.47*	-0.25 / 3.25	-0.19 / 3.36	0.22 / 3.46

C. Limitations and Extensions of the VSNR Metric

As with most algorithms which rely on models of human vision, there are limitations of the proposed VSNR metric. A primary shortcoming is that the metric is limited to grayscale images and is, therefore, blind to color-only errors. An extension of the VSNR metric might, therefore, include an additional stage which accounts for perceived distortion due to degradations in color (e.g., distances in S-CIELAB space [80]). Another limitation of the VSNR metric is its lack of spatial localization; rather, VSNR is measured for full-sized images and is, therefore, of limited utility for applications which require spatially localized measures of fidelity. Indeed, block-based or tree-structured variants of VSNR are certainly areas which warrant further investigation (see [56] for a spatially localized variant of the masking model). The VSNR metric is also sensitive to geometric distortions such as spatial shifting and rotations, transformations which are well known to have little effect on visual fidelity. A recent extension of the SSIM metric has shown complex wavelets to be useful in this regard [81]; a similar approach might also be used as an extension of VSNR.

Perhaps the most noteworthy limitation of the VSNR metric is that it is ultimately a measure of visual *fidelity*, and not a measure of visual *quality* [82]; thus, when $\mathbf{I}$ is visually indistinguishable from $\hat{\mathbf{I}}$, VSNR $= \infty$. However, it is well known that the visual quality of an image can be improved by various means, e.g., via sharpening or contrast enhancement. Moreover, we have recently shown that adding noise to a structurally distorted image tends to increase visual quality, despite the overall increase in distortion contrast [15]. Additional psychophysical research might facilitate future extensions of VSNR and other metrics to incorporate these types of percepts and, therefore, allow an absolute measure of visual quality.

VI. Conclusion

This paper presented a visual SNR for quantifying the visual fidelity of natural images. Via a two-stage approach, the proposed VSNR metric operates based on both low-level and mid-level properties of human vision. In the first stage, the visual detectability of the distortions is determined via wavelet-based models of visual masking and visual summation. If the distortions are below the threshold of detection, the distorted image is deemed to be of perfect visual fidelity (VSNR $= \infty$), and then the algorithm terminates. If the distortions are visible (suprathreshold), a second stage is applied in which the low-level property of perceived contrast and the mid-level property of global precedence are considered. These HVS properties are modeled as Euclidean distances in distortion-contrast space (in the context of a wavelet-based decomposition), and VSNR is determined based on the ratio of the original image's RMS contrast to the weighted sum of these two distances. The VSNR metric: 1) performs competitively with other visual fidelity metrics, 2) is efficient both in terms of computational complexity and in terms of memory requirements, and 3) operates based on physical luminances and visual angle and can therefore accommodate different viewing conditions.

The source code for the VSNR metric (in both C++ and MATLAB), along with preliminary comparisons with subjective ratings on other images, are available at http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html [83].

Appendix A
Computing $v(\mathbf{E})$ from $C(\mathbf{E})$

Let $\mathbf{I}$ and $\hat{\mathbf{I}}$ denote an original and distorted image, respectively; and let $\mathbf{E} \equiv \hat{\mathbf{I}} - \mathbf{I}$ denote the distortions contained within $\hat{\mathbf{I}}$. Let $C(\mathbf{E})$ denote the total RMS distortion contrast, and let $\mathbf{f} = [f_1, f_2, \ldots, f_M]$ denote a vector of octave-spaced frequencies in cycles/degree. The goal is to compute $v(\mathbf{E})$ such that the resulting $CSNR^*_{f_1}(\mathbf{E})$, $CSNR^*_{f_2}(\mathbf{E}), \ldots, CSNR^*_{f_M}(\mathbf{E})$ [computed via (8)] give rise to a total distortion contrast of $C(\mathbf{E})$. The following bisection search procedure is employed for this task.

1) Let $v_{lo} = 0$ and $v_{hi} = 1$.
2) Compute $v = (1/2)(v_{lo} + v_{hi})$.
3) Compute $CSNR^*_{f_1}, CSNR^*_{f_2}, \ldots, CSNR^*_{f_M}$ by using (8) and v from Step 2.
4) Compute $\hat{C} = \left(\sum_{m=1}^{M} [C(\mathbf{I}_{f_m})/CSNR^*_{f_m}]^2\right)^{1/2}$.
5) If $|\hat{C} - C(\mathbf{E})|$ is sufficiently small [e.g., within 1% of $C(\mathbf{E})$], then exit.
6) If $\hat{C} - C(\mathbf{E}) > 0$ (i.e., v is too large), then let $v_{hi} = v$, and then go to Step 2.
7) If $\hat{C} - C(\mathbf{E}) < 0$ (i.e., v is too small), then let $v_{lo} = v$, and then go to Step 2.

This procedure typically converges in less than 10 iterations. When using a tolerance in Step 4 of $|\hat{C} - C(\mathbf{E})| < 0.01C(\mathbf{E})$, as used to generate the results reported in Section V, the typical number of required iterations is 2–8.

Acknowledgment

The authors would like to thank H. Sheikh, Z. Wang, and the three anonymous reviewers for their helpful comments and suggestions on an earlier draft of this paper.

References

[1] P. C. Teo and D. J. Heeger, "Perceptual image distortion," *Proc. SPIE*, vol. 2179, pp. 127–141, 1994.
[2] Y. Lai and C. J. Kuo, "Image quality measurement using the haar wavelet," presented at the SPIE: Wavelet Applications in Signal and Image Processing V, 1997.
[3] S. Winkler, "Visual quality assessment using a contrast gain control model," in *Proc. IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, Sep. 1999, pp. 527–532.
[4] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[5] 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.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[6] G. Zhai, W. Zhang, X. Yang, and Y. Xu, "Image quality assessment metrics based on multi-scale edge presentation," in *Proc. IEEE Workshop on Signal Processing Systems Design and Implementation*, 2005, pp. 331–336.
[7] A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 422–429, Feb. 2006.
[8] B. Girod, "What's wrong with mean-squared error?," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 207–220.
[9] T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," in *Proc. Int. Conf. Image Processing*, 1996, pp. 237–240.
[10] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in jpeg 2000," *Signal Process.: Image Commun.*, vol. 17, pp. 85–104, 2001.
[11] D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 397–410, Apr. 2005.
[12] M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: Psychophysical evaluation and analysis," *J. Opt. Soc. Amer. A*, vol. 18, pp. 2385–2397, 2001.
[13] 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. Amer. A*, vol. 20, no. 7, Jul. 2003.
[14] D. M. Chandler and S. S. Hemami, "Suprathreshold image compression based on contrast allocation and global precedence," presented at the SPIE Human Vision and Electronic Imaging VIII, Santa Clara, CA, 2003.
[15] D. M. Chandler, K. H. S. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," presented at the SPIE Human Vision and Electronic Imaging XI, San Jose, CA, 2006.
[16] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," presented at the SPIE Human Vision and Electronic Imaging VII, San Jose, CA, 2002.
[17] C. Poynton, "The rehabilitation of gamma," in *Proc. SPIE Human Vision and Electronic Imaging III*, B. E. Rogowitz and T. N. Pappas, Eds., San Jose, CA, 1998, pp. 232–249.
[18] 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*, vol. 19, pp. 79–101, 1990.
[19] F. A. A. Kingdom, A. Hayes, and D. J. Field, "Sensitivity to contrast histogram differences in synthetic wavelet-textures," *Vis. Res.*, vol. 41, pp. 585–598, 1995.
[20] K. Tiippana, R. Näsänen, and J. Rovamo, "Contrast matching of two-dimensional compound gratings," *Vis. Res.*, vol. 34, pp. 1157–1163, 1994.
[21] P. J. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *J. Opt. Soc. Amer. A*, vol. 19, pp. 1096–1106, 2002.
[22] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, pp. 1458–1470, 1980.
[23] R. L. DeValois and K. K. DeValois, *Spatial Vision*. New York: Oxford Univ. Press, 1990.
[24] N. Graham, *Visual Pattern Analyzers*. New York: Oxford Univ. Press, 1989.
[25] 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.*, vol. 7, pp. 1–14, 1993.

[26] A. B. Watson, G. Y. Tangand, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.
[27] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of image," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 525–535, Jul. 1974.
[28] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. COM-30, no. 7, pp. 1679–1692, Jul. 1982.
[29] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. COM-33, no. 6, pp. 551–557, Jun. 1985.
[30] S. Daly, *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.
[31] S. J.P. Westen, R. L. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel HVS model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1995, vol. 4, pp. 2351–2354.
[32] J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, Ed. Singapore: World Scientific, 1995, pp. 245–283.
[33] 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 Int. Conf. Acoustics, Speech, Signal Processing*, May 1996, pp. 2291–2294.
[34] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective picture quality scale (PQS) for image coding," *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1215–1226, Sep. 1998.
[35] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," in *Proc. IEEE Int. Conf. Image Processing*, 1998, vol. 3, pp. 414–418.
[36] S. Winkler, "A perceptual distortion metric for digital color images," in *Proc. IEEE Int. Conf. Image Processing*, 1998, vol. 3, pp. 399–403.
[37] A. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Process.*, vol. 8, no. 5, pp. 717–730, May 1999.
[38] J. Lubin, "Method and apparatus for assessing the visibility of differences between two image sequences," U.S. Patent 5 974 159, 1999.
[39] Jndmetrix Technology [Online]. Available: http://www.sarnoff.com/ Sarnoff Corporation
[40] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, pp. 177–200, 1998.
[41] A. B. Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A*, vol. 14, pp. 2378–2390, 1997.
[42] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, no. 2, 2004.
[43] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," presented at the Asilomar Conf. Signals, Systems, and Computers, Nov. 2003.
[45] A. J. Ahumada and C. H. Null, "Image quality: A multidimensional problem," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 141–148.
[46] V. Kayargadde and J. Martens, "Perceptual characterization of images degraded by blur and noise: Experiments," *J. Opt. Soc. Amer. A*, vol. 13, no. 6, pp. 1166–1177, 1996.
[47] S. A. Karunasekera and N. G. Kingsbury, "Distortion measure for blocking artifacts in images based on human visual sensitivity," in *Proc. SPIE Visual Communications and Image Processing*, 1993, vol. 2094, pp. 474–486.
[48] 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.*, vol. 9, no. 4, pp. 636–650 , Apr. 2000.
[49] J. Nachmias and R. Sansbury, "Grating contrast discrimination may be better than detection," *Vis. Res.*, vol. 14, pp. 1039–1042, 1974.
[50] M. Carnec, P. L. Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. Int. Conf. Image Processing*, 2003, vol. 2, pp. 185–188.
[51] R. de Freitas Zampolo and R. Seara, "A comparison of image quality metric performances under practical conditions," in *Proc. Int. Conf. Image Processing*, 2005, vol. 3, pp. 1192–1195.
[52] J. Villasenor, B. Belzer, and J. Liao, "Wavelet filter evaluation for image compression," *IEEE Trans. Image Process.*, vol. 4, no. 8, pp. 1053–1060, Aug. 1995.
[53] N. Drasdo and C. W. Fowler, "Non-linear projection of the retinal image in a wide-angle schematic eye," *Brit. J. Ophthalmol.*, vol. 58, pp. 709–714, 1974.
[54] S. J. Daly, "Application of a noise-adaptive contrast sensitivity function to image data compression," *Opt. Eng.*, vol. 29, pp. 977–987, 1990.
[55] M. A. Masry, D. M. Chandler, and S. S. Hemami, "Digital watermarking using local contrast-based texture masking," in *Proc. Asilomar Conf. Signals, Systems, Computers*, Nov. 2003, pp. 1590–1595.
[56] M. D. Gaubatz, D. M. Chandler, and S. S. Hemami, "Spatial quantization via local texture masking," presented at the SPIE Human Vision Electronic Imaging X, San Jose, CA, 2005.
[57] D. M. Chandler and S. S. Hemami, "Visually lossless compression of digitized radiographs based on contrast sensitivity and visual masking," in *Proc. SPIE Medical Imaging*, 2005, vol. 5749, pp. 359–372.
[58] M. A. Georgeson and G. D. Sullivan, "Contrast constancy: Deblurring in human vision by spatial frequency channels," *J. Physiol.*, vol. 252, pp. 627–656, 1975.
[59] M. W. Cannon and S. C. Fullenkamp, "A transducer model for contrast perception," *Vis. Res.*, vol. 31, pp. 983–998.
[60] N. Brady and D. J. Field, "What's constant in contrast constancy? the effects of scaling on the perceived contrast of bandpass patterns," *Vis. Res.*, vol. 35, pp. 739–756, 1995.
[61] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cogn. Psych.*, vol. 9, pp. 353–383, 1977.
[62] P. G. Schyns and A. Oliva, "Dr. angry and mr. smile: When categorization flexibly modifies the perception of faces in rapid visual presentations," *Cognition*, vol. 69, pp. 243–265, 1999.
[63] A. Hayes, "Representation by images restricted in resolution and intensity range," Ph.D. dissertation, Univ. Western Australia, Perth, Australia, 1989.
[64] B. Willmore, R. J. Prenger, and J. L. Gallant, "Principles of neural shape coding in area V2," *J. Vis.*, vol. 5, no. 8, p. 82a.
[65] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Comput. Vis, Graph., Image Process.*, vol. 39, pp. 311–327, 1987.
[66] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, Image and Video Quality Assessment Research at LIVE [Online]. Available: http://live.ece.utexas.edu/research/quality/
[67] T. Mitsa and K. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1993, pp. 301–304.
[68] Refer to the wsnr_new.m MATLAB function included in the image quality 1.0 archive at: [Online]. Available: http://signal.ece.utexas.edu/software/ImageQuality/quality1.0/ImageQuality1_0.zip
[69] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Proces. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[70] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, A Standard Default Color Space for the Internet-sRGB [Online]. Available: http://www.w3.org/Graphics/Color/sRGB 1996
[71] SSIM website. [Online]. Available: http://www.cns.nyu.edu/\ zwang/files/research/ssim/index.html
[72] UQI website. [Online]. Available: http://www.cns.nyu.edu/\ zwang/files/research/quality_index/demo.html
[73] VIF website. [Online]. Available: http://live.ece.utexas.edu/research/quality/VIF.htm
[74] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[75] J. A. Nelder and R. Mead, "A simplex method for function minimization," *J. Comput.*, vol. 7, pp. 308–313, 1965.
[76] Z. Wang, May 1, 2006, personal communication.
[77] VQEG, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II August 2003 [Online]. Available: http://www.vqeg.org
[78] A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ. Lett.*, vol. 6, pp. 255–259, 1980.
[79] M. Martina and G. Masera, "Low-complexity, efficient 9/7 wavelet filters implementation," presented at the Int. Conf. Image Processing, 2005.
[80] X. Zhang and B. A. Wandell, "Color image fidelity metrics evaluated using image distortion maps," *Signal Process.*, vol. 70, no. 3, pp. 201–214, 1998.
[81] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in the complex wavelet domain," presented at the IEEE Int. Conf Acoustics Speech and Signal Processing, Mar. 2005.

[82] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics," presented at the SPIE Human Vision and Electronic Imaging VI, 2001.
[83] Online supplement. [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

Damon M. Chandler (S'03–M'06) received the B.S. degree in biomedical engineering from The Johns Hopkins University, Baltimore, MD, in 1998, and the M.Eng., M.S., and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000, 2003, and 2005, respectively.

From 2005 to 2006, he was with the Department of Psychology, Cornell University, working on topics in computational vision and image processing. In 2006, he joined the faculty of the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater. His research interests include image processing, data compression, computational vision, natural scene statistics, and visual psychophysics.

Sheila S. Hemami (S'89–M'95–SM'03) received the B.S. degree (summa cum laude) in electrical engineering from the University of Michigan, Ann Arbor, in 1990, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1992 and 1994, respectively.

She was with Hewlett-Packard Laboratories, Palo Alto, CA, in 1994. In 1995, she joined the faculty of the School of Electrical and Computer Engineering at Cornell University, Ithaca, NY, where she is currently an Associate Professor and Director of the Visual Communications Lab. She has held visiting positions at Princeton University, Princeton, NJ, and Rice University, Houston, TX (TI Distinguished Visiting Professor), and in 2001, she visited the Faculte de Sciences, Rabat, Morocco, as a Fulbright Distinguished Lecturer.

Dr. Hemami currently serves as Chair of the IEEE Image and Multidimensional Signal Processing Technical Committee, and has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She has served on various program committees and organizing committees. In 1997, she received a National Science Foundation CAREER Award. She held the Kodak Term Professorship of Electrical Engineering at Cornell University from 1996 to 1999. In 2000, she received the Eta Kappa Nu C. Holmes MacDonald Outstanding Teaching Award (a national award), and she has won numerous teaching awards at Cornell University. She was a finalist for the Eta Kappa Nu Outstanding Young Electrical Engineer in 2003. In 2005, she received the Alice H. Cook and Constance E. Cook Award at Cornell University for her leadership of the Women in Science and Engineering committee.