Quantifying the visual quality of wavelet-compressed images based on local contrast, visual masking, and global precedence

Damon M. Chandler, Mark A. Masry, and Sheila S. Hemami

Visual Communications Lab, School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853

Abstract—The paper presents a two-stage metric which quantifies the visual quality of images that have undergone wavelet-based compression. The first stage operates via a model of visual pattern masking, which takes as input original and distorted images, and which outputs masked contrast detection thresholds. For distortions beyond the threshold of detection, the images and the thresholds are fed into a second stage which estimates visual quality based on the distance between the distribution of assumed ideal and actual contrast signal-to-noise ratios across scale-space. Results indicate that the proposed metric yields a higher correlation with subjective-rating data than other visual quality metrics when applied to a sample of wavelet-coded images (with rates ranging from approximately 0.08–0.85 bits/pixel) for which peak signal-to-noise ratio correlates poorly with subjective quality.

I. INTRODUCTION

Classical approaches to assessing the visual quality of a distorted image have focused on quantifying the detectability of the distortions based on well-established properties of low-level vision (e.g., contrast sensitivity, visual masking) [1][2]. Methods of this type are useful at determining whether the distortions are below or beyond the threshold of visual detection; however, for suprathreshold distortions, it has proved difficult to relate the detectability of the distortions to the quality of the distorted image.

Other approaches to quality assessment have been based primarily on the intensity of the distortions (e.g., mean-squared error), oftentimes with adjustments for how these distortions are perceived at near-threshold and suprathreshold contrasts [3][4]. However, regardless of whether intensity is defined as physical or perceived, several researchers have suggested that the distortions alone cannot account for how a human observer perceives the quality of the distorted image [5][6]. In particular, the perceived contrast of suprathreshold wavelet subband quantization distortions was shown to be a poor indicator of the visual quality of wavelet-compressed images [5], suggesting that different visual processes may be involved during detection and recognition [7][8][9].

In this paper, a metric which quantifies the visual quality of images that have undergone wavelet-based compression is presented. The proposed metric, which operates for both near-threshold and suprathreshold distortions, estimates the visual quality of a reconstructed image relative to the original via a two-stage approach: In the first stage, the original image and the distortions are transformed into a set of discrete-wavelet-transform (DWT) subbands, which are then input into a model of visual pattern masking to determine whether the distortions are below or beyond the threshold of detection. If the distortions are suprathreshold, the subbands are input into a second stage which estimates visual quality based on how differently the contrast signal-to-noise ratios (CSNRs) of the distorted image are distributed across scale-space as compared to an assumed ideal distribution of CSNRs chosen to preserve the semblance of the image.

This paper is organized as follows: Section II provides definitions of contrast and contrast signal-to-noise ratio, and it describes how threshold and recognition-preserving CSNRs are modeled for wavelet subband quantization distortions. Section III describes how these data are used as part of a visual quality metric. Section IV compares the results of the proposed metric with subjective-rating data. General conclusions are presented in Section V.

II. BACKGROUND

Quantifying the visual quality of a distorted image requires understanding human visual responses to both the distortions and the image upon which these distortions are displayed. Given an original image \( I \) and a distorted image \( I^\prime \), the (mean-offset) distortions \( E \) are given by \( E = I - I^\prime \), where \( I^\prime \) denotes the mean pixel value of \( I \). Because the human visual system (HVS) responds to ratios of luminances (i.e., contrasts) rather than to absolute luminances, it is necessary to quantify the contrasts of \( E \) and \( I \). The root-mean-squared (RMS) contrast \( C_S \) of a stimulus \( S \), which has proved useful for compound [8], noise [9], wavelet [11], and natural-image [12] stimuli, is given by

\[
C_S = \frac{1}{\mu_L} \left( \frac{1}{N \times M} \sum_{y=0}^{N} \sum_{x=0}^{M} [L_S(x,y) - \mu_L]^2 \right)^{1/2}
\]

where \( \mu_L \) denotes the average luminance of \( S \), \( L_S(x,y) \) denotes the luminance of the pixel at location \((x,y)\), and \( N \times M \) denotes the total number of pixels. Physical luminance \( L_S(x,y) \) can be computed from digital pixel value \( S(x,y) \) as \( L_S(x,y) = \gamma [b + k \cdot S(x,y)]^{\gamma} \), where \( b \) represents the black-level offset, \( k \) the pixel-value-to-voltage scaling factor, and \( \gamma \) the gamma of the display monitor [13].
Contrast signal-to-noise ratio at a spatial scale centered at frequency \( f \) is defined as

\[
CSNR_f = \frac{C(I,f)}{E(I)}
\]  
(2)

where \( CSNR_f \) denotes contrast signal-to-noise ratio; where \( I(f) \) and \( E(f) \) denote, respectively, the image and distortion content within the spatial scales centered at frequency \( f \); and where \( C(I,f) \) and \( C(E,f) \) denote the respective RMS contrasts of these components.

A. CSNRs for Distortions at Threshold

When \( E(f) \) is at the threshold of detection when presented against an image-masker \( I \), the contrast of \( E(f) \) is given by \( CT_{E(I)} \), where \( CT_{E(I)} \) denotes the masked contrast detection threshold of \( E(f) \). The corresponding threshold contrast signal-to-noise ratio is given by \( CSNR_{f,thr} = \frac{C(I,f)}{CT_{E(I)}} \).

In a previous study [14], we have shown that for targets consisting of wavelet subband quantization distortions, and for typical values of \( f \), the threshold contrast signal-to-noise ratios are modeled as

\[
CSNR_{f,thr} = a_0(I) \cdot f^{a_1(I)} + a_2(I),
\]  
(3)

where \( CSNR_{f,thr} \) denotes the threshold contrast signal-to-noise ratio for distortions centered at spatial frequency \( f \) and presented against \( I \). This functional represents a parabola in log-frequency log-contrast coordinates whose offset and shape are defined by the image-dependent parameters \( a_0(I), a_1(I) \) and \( a_2(I) \).

Figure 1(a) depicts \( CSNR_{f,thr}(I) \) computed at \( f = 1.15, 2.3, 4.6, 9.2 \), and 18.4 cycles/degree for image horse (see Figure 3) using the 9-7 biorthogonal filters [15] and assuming sRGB [16] display characteristics. In this case, \( a_0 = 36.0, a_1 = 0.54, \) and \( a_2 = -0.47 \).

To predict the parameters \( a_0(I), a_1(I) \) and \( a_2(I) \) for any given natural image, we employ a model of visual pattern masking. This model, which is described in detail in [17], computes a spatially localized measure of visual masking via a gain-control mechanism in which the combined response of channels on the orientation dimension is used to regulate the excitatory response of the detecting channel [18]. To estimate \( a_0(I), a_1(I) \) and \( a_2(I) \), these masking data are then pooled across space and spatial frequency using a Minkowski metric with exponents of 2 and 1.5, respectively.

B. Recognition-Preserving CSNRs for Suprathreshold Distortions

When the distortions are visible within the image (i.e., when \( C(E,f) > CT_{E(I)} \)), frequency-dependent detection thresholds fail to correlate with actual values of perceived contrast; rather, psychophysical studies using suprathreshold stimuli have traditionally revealed an invariance of perceived contrast with spatial frequency (contrast constancy [19][20]).

We have advocated in [21] and [5] that the contrasts of suprathreshold distortions should be proportioned across spatial frequency so as to preserve the visual system’s ability to integrate image-features across scale-space. Specifically, because features are visually integrated in a coarse-to-fine scale order (global precedence [22][23]), the recognizability of an image can be maintained by preserving coarse-scale structure at the expense of fine-scale structure. Results from a previous experiment [21] suggest that the contrast signal-to-noise ratios required to effect this criterion are modeled as

\[
CSNR_f(I,C_E) = b_0(I,C_E) \cdot f^{b_1(I,C_E)} + b_2(I,C_E),
\]  
(4)

where \( C_E \) denotes the total RMS contrast of the distortions and where \( CSNR_f(I,C_E) \) denotes the CSNR that is believed to preserve the semblance of the image for a given \( C_E \). The parameters \( b_0(I,C_E), b_1(I,C_E) \), and \( b_2(I,C_E) \) serve to adapt the distribution of CSNRs across spatial frequency based on \( C_E \) and \( I \); these parameters were fit to the experimental data of [21] and are defined as follows:

\[
b_0(I,C_E) = \frac{1.5 - a_0(I)}{10} \cdot \log\left(\frac{C_E}{CT_{E(I)}}\right) + a_0(I),
\]

\[
b_1(I,C_E) = \frac{1.0 - a_1(I)}{10} \cdot \log\left(\frac{C_E}{CT_{E(I)}}\right) + a_1(I),
\]

\[
b_2(I,C_E) = \frac{-2.0 - a_2(I)}{10} \cdot \log\left(\frac{C_E}{CT_{E(I)}}\right) + a_2(I),
\]

where \( CT_{E(I)} \) denotes the total RMS contrast of the distortions at the masked threshold of detection. Note that Equation (3) is a special case of Equation (4) for \( C_E = CT_{E(I)} \).

Figures 1(b)–(i) depict \( CSNR_f(I,C_E) \) computed at five spatial scales of image horse (see Figure 3) for increasing values of \( C_E > CT_{E(I)} \). For distortions induced via quantization of wavelet subbands, CSNRs \( \leq 1.0 \) represent subbands that have been discarded (i.e., quantized to all zeros). Notice from the lower curves of Figure 1, which correspond to greater values of \( C_E \), that subbands are discarded in a fine-to-coarse scale progression.

The following section describes an algorithm which uses the threshold and recognition-preserving CSNRs to estimate the visual quality of a distorted image.

III. ALGORITHM

Let \( f \) denote a vector of spatial frequencies given by \( f = [f_1, f_2, \ldots, f_M] \), where \( f_n = 2^n f_{base} \), \( n = 1, 2, \ldots, M \) for some baseline spatial frequency \( f_{base} \). Let \( CSNR_f^*(I,C_E) \) denote a vector of CSNRs given by \( CSNR_f^*(I,C_E) = [CSNR_{f_1}^*(I,C_E), CSNR_{f_2}^*(I,C_E), \ldots, CSNR_{f_M}^*(I,C_E)] \).

Assuming that, for a given image \( I \), \( CSNR_f^*(I,C_E) \) defines a \( C_E \)-parameterized path of minimum visual distortion in an \( M \)-dimensional space in which the \( n \)th axis represents \( CSNR_{f_n} \), the visual quality between the original image \( I \) and the distorted image \( I' \) can be estimated based on a sum of two distances:

1) The distance along the path of minimum visual distortion;
2) The Euclidean distance between the (ideal) CSNRs and the (actual) CSNRs measured from the distorted image.

The first of these two distances is designed to measure the visual distortion between the original image \( I \) and the image that
would have been obtained had the contrasts of the distortions been allocated as prescribed by \( \text{CSNR}_I(I, C_E) \). The second distance is designed to measure the visual distortion between this assumed ideal image and the actual distorted image \( I \). The following sections describe how these distances are computed and then mapped to a measure of visual quality.

A. Preprocessing

Given the original and distorted images \( I \) and \( \hat{I} \) and parameters \((h, k, \gamma)\) of the display device, the distortions \( E \) are computed as described in Section II. An \( M \)-level discrete wavelet transform (DWT) is then performed on both \( I \) and \( \hat{I} \) to obtain two sets of \( 3M + 1 \) subbands: \( \{ s_I \} \) and \( \{ s_E \} \). The vector \( f \) is computed as described in Section III, where \( f_{\text{base}} \) is given by the intended viewing distance and the resolution of the display (see [24]). Visual quality, \( V_Q \in [0, 1] \), is then computed via the following steps.

B. Compute Reference and Actual CSNRs

The subbands \( \{ s_I \} \) and \( \{ s_E \} \) are input into a model of visual pattern masking [17] to determine whether the distortions \( E \) are below or beyond the threshold of detection when presented against the image-masker \( I \). If the distortions are subthreshold, then \( V_Q = 1 \) (denoting maximum quality), and then the algorithm terminates.

If the distortions are suprathreshold, then the model is queried for the parameters \( a_0, a_1, \alpha_1, \) and \( \beta_1 \). These data are then used in Equation (4) to compute a vector of reference CSNRs, \( \text{CSNR}_I(I, C_E) \).

Next, a vector of actual CSNRs, \( \text{CSNR}_I^\text{act} \), is either measured directly via Equation (2) (which requires a series of inverse DWTs) or is estimated based on the subbands via

\[
\text{CSNR}_I(I) = \left( \frac{\sigma_{s_I, h, n}^2 + \sigma_{s_I, k, n}^2 + \sigma_{s_I, \gamma, n}^2}{\sigma_{s_E, h, n}^2 + \sigma_{s_E, k, n}^2 + \sigma_{s_E, \gamma, n}^2} \right)^{1/2}
\]

where \( \sigma_{s_I, h, n} \) and \( \sigma_{s_I, k, n} \) denote, respectively, the standard deviations of the subbands of \( I \) and \( E \) at the \((M - n - 1)^{th}\) level of decomposition with orientation \( \theta = LH, HL, \) or \( HH \).

C. Compute Visual Quality

Given \( \text{CSNR}_I(I, C_E) \) and \( \text{CSNR}_I^\text{act} \), visual distortion \( V_D \in [0, 1] \) is computed as

\[
V_D = 1 - \frac{1}{1 + \left( a_0(d + d^*) + a_1 \frac{N_{\text{disc}}}{C_E} \right)^{\beta}}
\]

where \( N_{\text{disc}} \) denotes the count of the number of subbands for which \( \text{CSNR} < 1 \) (thus, \( N_{\text{disc}} / C_E \) serves as a rough measure of blur), and where \( d^* \) and \( d \) correspond to the previously-described distances in CSNR space.

The quantity \( d^* \) is the distance between \( \text{CSNR}_I(I, C_TK) \) and \( \text{CSNR}_I(I, C_E) \) computed as a sum of line-segment distances along the \( \text{CSNR}_I \) path:

\[
d^* = \sum_{C = C_TK} \| \text{CSNR}_I(I, C) - \text{CSNR}_I(I, C + \Delta C) \|_2.
\]

The quantity \( d \) is a weighted Euclidean distance between \( \text{CSNR}_I(I, C_E) \) and \( \text{CSNR}_I^\text{act} \):

\[
d = \left( \sum_{n=1}^{M} w_n \cdot \left| \text{CSNR}_I^\text{act}(I, C_E) - \text{CSNR}_I^\text{act}(I) \right|^2 \right)^{1/2}
\]

where \( w_n = 1 \) for \( \text{CSNR}_I^\text{act}(I) < \text{CSNR}_I(I, C_E) \), and 0 otherwise; \( d \) therefore includes contributions only at scales for which \( I \) has a lower CSNR than that specified by \( C_E \).

As described in the following section, the free parameters \( a_0, a_1, \alpha_1, \alpha_2, \) and \( \beta \) in Equation (5) are determined based on subjective-rating data.

Using Equation (5), visual quality is given by \( V_Q = 1 - V_D \), where a value of \( V_Q = 1 \) denotes maximum visual quality. Note that the limits on \( V_Q \) and \( V_D \) imply that visual quality cannot be increased by the addition of \( E \) to \( I \).

IV. EXPERIMENT AND RESULTS

A psychophysical experiment was performed in which subjects rated visual distortion in images containing distortions at fixed values of \( C_E \) allocated across scale-space in four different ways: (1) equal distortion contrast at each scale; (2) Embedded DCQ quantization [25]; (3) JPEG-2000 compression using the default Post-Compression Rate-Distortion optimization; and (4) baseline JPEG compression. Three natural images, "horse," "harbour," and "baby," were used in the experiment; the distorted versions of "horse" and "harbour" contained distortions at \( C_E \) values of 0.11, 0.145, and 0.21; the distorted versions of "baby" contained distortions at \( C_E \) values of 0.053, 0.065, and 0.082. Four adult subjects participated in the experiment; all subjects had normal or corrected-to-normal visual acuity. The raw scores for each subject were converted to Z-scores and the average Z-scores over all subjects were adjusted to span the range \([0, 0.54]\), where the upper limit was chosen based on previous subjective ratings of the images reconstructed using only their respective lowest-frequency \(( LL )\) subbands.

For these images, the proposed metric yields a higher correlation with subjective ratings than other visual quality metrics. Figure 2 depicts the subjective-rating scores plotted against various visual quality metrics. In Figure 2(a), the subjective-rating scores are plotted against peak signal-to-noise ratio; the best-fitting logistic function to these data yield a coefficient of determination of \( R^2 = 0.37 \). In Figure 2(b), the subjective-rating scores are plotted against SSIM (Structural SIMilarity) indices [6]; the best-fitting logistic function to these data yield \( R^2 = 0.78 \). In Figure 2(c), the subjective-rating scores are plotted against \( V_Q \) values of the proposed metric using \( a_0 = 0.012, a_1 = 0.015, \) and \( \beta = 3.1 \); these data yield \( R^2 = 0.90 \).

Figure 3 depicts the results of the proposed metric on image "horse" containing wavelet subband quantization distortions at a total RMS contrast of \( C_E = 0.21 \). The original image is provided in Figure 3(a). Figure 3(b) was generated using embedded DCQ quantization [25]. Figure 3(c) was generated using the default JPEG-2000 Post-Compression Rate-Distortion optimization [26]. The image in Figure 3(d) was generated by adding spatially correlated wavelet distortions in
which the contrasts of the distortions were set to equal values for each spatial scale.

The $V_2$ values of Figures 2 and 3 were generated using five levels of decomposition with the 9-7 filters [15] and assuming an SRGB [16] display and a viewing distance of three picture heights. Also listed in Figure 3 are subjective-rating scores and indices of UIQ (Universal Image Quality) [27] and SSIM [6]. For the images depicted in Figures 3(b) and (c), UIQ and SSIM indices correlate well with the proposed metric ($V_2$) and with the subjective ratings. For the image in Figure 3(d), the UIQ and SSIM metrics fail to capture the visual distortion induced by the disruption of global precedence; both of these metrics rate Figure 3(d) greatest in visual quality.

V. CONCLUSIONS

This paper presented a visual quality metric for wavelet-compressed images. The proposed metric operates by using: (1) a model of visual pattern masking to determine if the distortions are visible; and (2) a suprathreshold stage which estimates the visual quality of the distolted image based on actual contrast signal-to-noise ratios across scale-space. When applied to a sample of wavelet-coded images for which peak signal-to-noise ratio correlates poorly with subjective quality, the proposed metric yields a higher correlation with subjective- compressed images. The proposed metric operates by using:

The relationship between $d$, $d'$, $N_{\text{discard}}$ and $V_2$.

REFERENCES


Fig. 1. Contrast signal-to-noise ratios of image horse at five spatial scales centered at spatial frequencies 1.15-18.4 cycles/degree. (a) CSNRs at the threshold of detection; (b)-(i) CSNRs that are assumed to minimize visual distortion as the contrasts of the distortions become increasingly suprathreshold. Note that when $V_2 \leq 1.0$ (denoted by the shaded region), the corresponding subbands have been quantized to all zeros.


Fig. 2. Scatterplots depicting subjective-rating scores (vertical axis) plotted against three quality metrics: (a) PSNR, (b) SSIM, and (c) the proposed metric. Solid lines denote best-fitting logistic or linear functions.

Fig. 3. (a) Original image horse. (b)-(d) Image horse distorted via: (b) Embedded DCQ quantization [5]; (c) JPEG-2000 compression [28]; and (d) addition of spatially correlated wavelet noise. All distorted images contain distortions at a total RMS contrast of $\xi = 0.21$ (assuming sRGB display characteristics). UIQ indices [27] range from $-1$ to $1$; subjective ratings, SSIM indices [6], and $V_e$ values range from 0 to 1; in all cases, a value of 1 denotes maximum visual quality.