Analyzing the Role of Visual Structure in the Recognition of Natural Image Content with Multi-Scale SSIM

David M. Rouse and Sheila S. Hemami
Visual Communications Lab, School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853

ABSTRACT
Natural images are meaningful to humans — the physical world exhibits statistical regularities that permit the human visual system (HVS) to infer useful interpretations. These regularities communicate the visual structure of the physical world and govern the statistics of images (image structure). A signal processing framework is sought to analyze image characteristics for a relationship with human interpretation. This work investigates the first step toward an objective visual information evaluation: predicting the recognition threshold of different image representations. Given a image sequence, whose images begin as unrecognizable and are gradually refined to include more information according to some measure, the recognition threshold corresponds to first the image in the sequence in which an observer accurately identifies the content. Sequences are produced using two types of image representations: signal-based and visual structure preserving. Signal-based representations add information as dictated by conventional mathematical characterizations of images based on models of low-level HVS processing and use basis functions as the basic image components. Visual structure preserving representations add information to images attributed to visual structure and attempt to mimic higher-level HVS processing by considering the scene’s objects as the basic image components. An experiment is conducted to identify the recognition threshold image. Several full-reference perceptual quality assessment algorithms are evaluated in terms of their ability to predict the recognition threshold of different image representations. The cross-correlation component of a modified version of the multi-scale structural similarity (MS-SSIM) metric, denoted MS-SSIM*, exhibits a better overall correlation with the signal-based and visual structure preserving representations’ average recognition thresholds than the standard MS-SSIM cross-correlation component. These findings underscore the significance of visual structure in recognition and advocate a multi-scale image structure analysis for a rudimentary evaluation of visual information.

Keywords: Visual Information, Human Visual System Modeling, Cognition, Perceptual Quality Assessment

1. INTRODUCTION
Natural images convey meaningful information to human observers. A prevailing theory of the human visual system (HVS) asserts that the physical world exhibits statistical regularities that permit the HVS to infer useful interpretations.1–3 These regularities communicate the visual structure of the physical world to human observers and govern image characteristics4 (image structure). Marr attributes the following basic inferences from natural images to visual structure: 1) smooth image regions correspond to the same surface and 2) image discontinuities indicate either a change in surface orientation or depth disparities.5 These basic inferences parse an image in terms of meaningful objects and boundaries. Thus, the visual structure conveys visual information. We use the term visual information to quantify the extent of a natural image’s scene content in a manner consistent with subjective ratings in response to the query “How much content is in this image?” For example, an image of a desktop cluttered with books and writing utensils would possess more visual information than an image of an empty desktop.

Useful transmission and storage of natural images demands that image processing algorithms minimize the loss of visual information. Modern algorithms analyze natural images strictly as signals possessing mathematical characteristics such as correlation and frequency content, using processing designed to exploit or manipulate these characteristics but overlooking the meaning or usefulness of the images to human observers. Several processing
models consistent with the low-level, signal analysis portion of the HVS decompose the input according to frequency, orientation, and contrast. Such a decomposition aims to simulate neural responses to these image components. The HVS, when viewed as an information processing system, derives an interpretation from the “signal” generated from the excitation of photoreceptors elicited by a visual stimulus. This continuum of processing from stimulus to interpretation, and namely recognition, inspires extensions to current signal processing models that integrate properties of higher-level visual processing. In particular, a signal processing framework is sought to analyze image features for a relationship with human interpretation.

This work investigates the first step toward an objective visual information evaluation: predicting the recognition threshold of different image representations. Explicitly quantifying an image’s visual information is currently intractable, so this work investigates the recognition threshold of different image representations. The recognition threshold indicates the lowest quality of visual information retained by a distorted image to be a useful surrogate for the reference image. For a specified representation, the recognition threshold distinguishes images based on their ability to convey salient visual information from the reference image. For a human observer, images whose visual information exceeds that of the recognition threshold are useful, whereas those with visual information of lower quality than the recognition threshold have little or no use. Detecting that an image is distorted such that it lies beneath the recognition threshold (i.e., the content of the reference image is unrecognizable) avoids devoting resources to a useless image. Understanding the image characteristics that make an image recognizable will yield algorithms that efficiently maximize the available visual information.

Our approach to predict natural image recognition thresholds consists of two parts. In the first part, image representations are defined, and an experiment is conducted to identify each representation’s recognition threshold for several natural images. Two types of natural image representations are investigated in this work: signal-based and visual structure preserving. Signal-based (SB) representations (cf. Figure 1(d)) add information as dictated by conventional mathematical characterizations of images based on models of low-level HVS processing and use basis functions as the basic image components. Visual structure preserving (VSP) representations (cf. Figure 1(b)) add information to images attributed to visual structure and attempt to mimic higher-level HVS processing by considering the scene’s objects as the basic image components. These image representations specify image sequences that are ordered according to the information available from the reference natural image. In the experiment, observers view the sequences in order of increasing information and provide written responses that describe the content of each image in the sequence. Recognition thresholds are identified from the observer responses. Since observers recognize the same content in each image representation, the SB recognition threshold image and the VSP recognition threshold image must retain common characteristics from the reference natural image attributed to visual structure.

In the second part, the ability of perceptual quality assessment (QA) algorithms to predict recognition thresholds is evaluated and compared. Such algorithms attempt to compute objective evaluations of visual quality consistent with subjective human evaluations and provide a means to distinguish characteristics that facilitate recognition. Based on conclusions from previous studies by one of the authors, we hypothesize that subjective quality evaluations rely upon the observers’ expectations of the phenomenal appearance of the physical world. Hence, QA algorithms consistent with subjective quality evaluations for a variety of image artifacts are also likely to predict image recognition thresholds.

Among the QA algorithms evaluated, the cross-correlation component of a modified version of the multi-scale structural similarity (MS-SSIM) metric, denoted MS-SSIM*, exhibits a better overall correlation with the SB and VSP representations’ average recognition thresholds than the standard MS-SSIM cross-correlation component. Thus, the image structure assessment by the MS-SSIM* cross-correlation component provides a assessment of visual structure that predicts natural image recognition thresholds. Furthermore, this finding underscores the significance of visual structure in recognition and advocate a multi-scale image structure analysis for a rudimentary evaluation of visual information.

This paper has the following organization: Section 2 mathematically specifies the two types of image representations investigated. The methods and stimuli used in the experiment are described in Section 3. Experiment results are presented in Section 4. Section 5 reviews the QA algorithms evaluated in this work. An analysis and with regard to QA algorithms is provided in Section 6. Conclusions are presented in Section 7.
2. IMAGE REPRESENTATION SPECIFICATIONS

This section describes the two image representations investigated: signal-based and visual structure preserving. Each representation specifies an image sequence that evolves from an unrecognizable version to a recognizable version of the reference natural image. That is, subsequent images in a sequence contain additional detail or information relative to the previous images.

2.1 Signal-based Representations

Signal-based (SB) (cf. Figure 1(d)) representations add information as dictated by conventional mathematical characterizations of images. This representation mimics the initial, low-level signal analysis portion of the HVS and uses a discrete wavelet transform (DWT) to decompose an image into subbands tuned to a band of spatial frequencies and orientations. Although the DWT is not necessarily an accurate model of the initial stages of the HVS,\textsuperscript{6,7} many lossy image compression standards, including JPEG-2000, adopt this strategy.\textsuperscript{16}

Two types of SB representations were examined in a previous experiment,\textsuperscript{17} and both representations encode an image using a JPEG-2000 (J2K) encoder, which uses the DWT. The two types differed in quantization strategy. For the first type, denoted DCQ-J2K, the quantization step-sizes are assigned according to the dynamic contrast-based quantization (DCQ) algorithm.\textsuperscript{18} Given a specified bitrate, a measure of visual distortion is calculated based on the image, the subband, and display characteristics,\textsuperscript{19} which the DCQ algorithm uses to assign subband quantization step-sizes. In addition, the DCQ strategy incorporates the property of global precedence\textsuperscript{20} and discards subband coefficients in a fine-to-coarse order. The DCQ algorithm produces lossy images generally rated superior in visual quality to those produced by other visually lossy compression algorithms.\textsuperscript{18}

For the second type, denoted MSE-J2K, the quantization step-sizes are assigned to minimize the mean-squared error (MSE) for a particular bitrate. This allows a comparison of MSE-based and perceptually-based quantization in the context of signal-based representations.

The encoding bitrate parameterizes the image sequences for both the DCQ-J2K and the MSE-J2K SB representations. Smaller values of bitrate correspond with low-quality images that appear near the beginning of the sequence, and larger values of bitrate correspond with high-quality images that appear near the end of the sequence. The SB representation’s recognition thresholds are reported in terms of the encoding bitrate of the recognition threshold image.

2.2 Visual Structure Preserving Representations

Visual structure preserving (VSP) representations (cf. Figure 1(b)) add information to images attributed to visual structure. Visual structure manifests itself in images as structural features (e.g., object boundaries and edges) that are hypothesized to convey salient visual information, whereas finer object details such as textures carry negligible visual information. This representation is designed to mimic higher-level HVS processing, since
objects serve as the basic components of these representations. VSP representations employ processing strategies that discard finer textures with limited disruption to visual structure.

Piecewise smooth images emphasize object boundaries and edges over textures and coincide with the properties desired for VSP representations. Total variation (TV) regularization traditionally has been applied to the problem of image denoising and generates piecewise smooth images. Let \( g(t) \) be a continuous signal obtained by adding noise to a reference signal for \( t \in [a, b] \). TV regularization finds a restored signal \( f \) from \( g \) by solving an optimization problem of the form:

\[
\min_{f} \int_{a}^{b} \left( (f(t) - g(t))^2 + \lambda \left| \frac{d}{dt} f(t) \right| \right) dt,
\]

where the first term maintains the similarity between the \( f \) and \( g \), the second term penalizes deviations from smoothness, and \( \lambda \) is a regularization parameter to control the amount of smoothing.

An alternative and equivalent approach to finding \( f \) is via soft thresholding of undecimated Haar wavelet coefficients in all subbands except the coarsest LL subband. Soft thresholding with thresholding parameter \( \tau \) is given by

\[
S_{\tau}(x) = \begin{cases} 
  x - \tau \text{sgn}(x) & |x| > \tau \\
  0 & |x| \leq \tau,
\end{cases}
\]

where \( \text{sgn}(x) \) is the signum function. Given a thresholding parameter \( \tau \), thresholds are appropriately scaled for wavelet coefficients at each scale. Adjusting the parameter \( \tau \) varies the level of detail removed from the reference image, where smaller values of \( \tau \) result in the removal of very few details. On the other hand, larger values of \( \tau \) induce more aggressive smoothing which may simultaneously compromise image structures (e.g., edges) important for interpretation.

Presented at SPIE Human Vision and Electronic Imaging XIII, San Jose, CA 2008
Select images from sequences corresponding to two types of VSP representations of the natural image *airplane* are shown in Figure 2. The two types differ with respect to the inclusion of the low spatial frequency coefficients from the coarsest LL subband. We hypothesized that the signal information contained within the LL subband contributes negligible visual information. Moreover, this hypothesis contends that excluding these low spatial frequency coefficients will not significantly alter the recognition thresholds.

The soft-thresholding parameter $\tau$ parameterizes the image sequence for the VSP representations. Larger values of $\tau$ correspond with low-quality images that appear near the beginning of the sequence, and smaller values of $\tau$ corresponding with high-quality images that appear near the end of the sequence. The recognition thresholds for the VSP representation are reported in terms of the soft thresholding parameter $\tau$ of the recognition threshold image.

3. EXPERIMENT AND METHODS: OBSERVER RECOGNITION THRESHOLDS

The experiment conducted in our previous study\(^\text{17}\) identified recognition threshold images (and their bitrates) for both types of signal-based (SB) representations: DCQ-J2K and MSE-J2K. The experiment conducted for this work identifies recognition threshold images for two types of visual structure preserving (VSP) representations. The remainder of this section reviews the methods for the experiment designed in the authors’ previous study\(^\text{17}\) to identify recognition threshold images for the VSP representations.

**Stimuli:** Nine grayscale natural images of size $512 \times 512$ pixels were cropped from reference natural images. The images selected contained simple events that could be briefly described with a few words. The nine natural images used in these experiments are shown in a previous publication by the authors.\(^\text{17}\) The brief descriptions of the image content used to identify an observer’s point of recognition are provided in the caption for Figure 3.

Image sequences corresponding to VSP representations for each natural image were generated by soft thresholding undecimated Haar wavelet coefficients from a 5-level decomposition for 24 values of the parameter $\tau$, which were logarithmically equally spaced from 2048 to 1. The sequence evolves from a low-detail cartoon, capturing the basic shapes of the image contents, toward a high-quality version of the reference natural image.

**Procedure:** For each of the nine images, sequences of VSP representations were generated as described. Observers viewed image sequences corresponding to one of the two the VSP representations (i.e., including verses excluding the coefficients of the coarsest LL subband) for each of the nine natural images. For each image in a sequence, the observer was tasked to provide a written description of the image content. The next image in the sequence was shown upon submission of a description; a time limit was not imposed. Participants typically completed the experiment in about 30 minutes.

An observer’s point of recognition was identified when the description contained both adequate and accurate information to briefly describe the reference image content. The soft threshold parameter $\tau$ corresponding to the image for each observer’s recognition threshold of each natural image was noted for both versions of the VSP representations.

**Participants:** Twenty-four observers with normal or corrected-to-normal acuity participated in this experiment. The age of the observers ranged from 18 to 56 where the average and median age is 28.2 and 24, respectively. The observers were compensated for their participation. Prior to their participation in the experiment, observers read a summary of the study’s objectives. In addition, sample images from a sequence corresponding to a natural image not tested were shown to observers before the experiment to illustrate the visual effects of both versions of the VSP representations. Sequences for each version of the VSP representations were viewed by 12 observers.

4. RESULTS: OBSERVER RECOGNITION THRESHOLDS

This section reports the average observer recognition threshold for both versions of the visual structure preserving (VSP) representations and each natural image. In addition, the average recognition thresholds for the signal-based (SB) representations from the previous study are provided for comparison.\(^\text{17}\) For each image representation and each natural image, the sample average value of the representation parameter (i.e., bitrate or $\tau$) corresponding to the observer recognition threshold images is computed.
Figure 3. Average recognition thresholds for each natural image are presented according to the logarithm of the parameter varied for the corresponding image representations. The ordinate axis of the VSP representations has been reversed such that larger values of $\tau$ appropriately correspond with lower values of bitrate for the SB representations. The ordinate axes are ordered such that recognition thresholds near the top of the figure reflect the need for more information from the reference natural image to recognize the content. Acronyms based on the image descriptions identify the nine natural images: 
- airplane (A), 
- boy and cat (B&C), 
- backhoe (B), 
- train (T), 
- pianist (P), 
- skier (S), 
- jack-o-lanterns (J-L), 
- caged birds (CB), and 
- guitarist (G).

The average recognition threshold parameters for each natural image and each version of the VSP representations are graphed in Figure 3(a). The average recognition threshold parameters determined by our previous experiment\(^\text{17}\) for the SB representations have been provided in Figure 3(b) for comparison. The recognition thresholds are presented with respect to the logarithm of the parameter varied for the corresponding representation. For the VSP representation results, the ordinate axis for the VSP representations has been reversed such that larger values of $\tau$ appropriately correspond with lower values of bitrate for the SB representations. Furthermore, both ordinate axes are ordered such that values closer to the top of the figure coincide with average observer recognition occurring later in the sequence. In the case of the VSP representations, lower threshold values in terms of the soft thresholding parameter $\tau$ reflect the need for more information from the reference natural image for recognition, while for the SB representations the need more information coincides with higher encoding bitrates.

Comparisons of the average recognition thresholds for the VSP representations with the SB representations reveal similar trends among the nine images tested. Among both image representations, the images boy and cat, backhoe, and caged birds have average recognition threshold parameters that coincide with the inclusion of more information from the reference natural image than the remaining six images tested. This reflects the difficulty encountered by the observers to recognize the basic image content. The similarities between the two image representations suggest a common characteristic from the reference natural image retained by both image representations that facilitates recognition.

A two-sample $t$ test was performed for each natural image to determine if the average recognition threshold parameter values for the two versions of the VSP representation differ when the LL subband coefficients are discarded. At the 5% significance level, the test only rejected the null hypothesis (equal thresholds) for the image guitarist ($p = 0.0276$). This analysis emphasizes the importance of the higher-frequency subbands for the recognition task, which corroborates results from psychological studies emphasizing the value of higher spatial frequency content (e.g., object boundaries and edges) in recognition tasks.\(^\text{24}\)

5. FULL-REFERENCE QUALITY ASSESSMENT ALGORITHMS

The incompatibility of the image representation parameters, $\tau$ and bitrate, limits comparisons across representations to a qualitative level and offers little insight as to the image characteristics retained from the reference.
image that permit recognition. Quality assessment (QA) algorithms provide a mechanism to analyze recognition threshold images to distinguish characteristics that facilitate or permit recognition. This section reviews the full-reference (FR) QA algorithms investigated for their ability to predict the recognition threshold of both image representations. FR QA algorithms evaluate a processed (test) image \( \hat{X} \) using the corresponding reference image \( X \) to quantify the perceptual “distance” between the two images. These FR QA algorithms can be categorized as 1) conventional distortion measures, 2) algorithms based on psychophysical properties of the HVS, and 3) algorithms derived from hypothetical high-level HVS objectives. While many QA algorithms based on hypothetical high-level HVS objectives demonstrate evaluations of visual quality consistent with subjective evaluations, this work specifically focuses on the structural similarity (SSIM)\(^{10}\) metric and its multi-scale extension, MS-SSIM.\(^9\) A mathematical description of these two QA algorithms is included in the last part of this section.

A challenge for QA algorithms is to generate consistent evaluations with subjective quality evaluations for a variety of image artifacts.\(^{25,26}\) We hypothesize that subjective quality evaluations exhibit a preference for distorted images that are compatible with observers’ expectations of the phenomenal appearance of the physical world. Indeed, images corrupted with correlated distortions (e.g., those caused by quantizing wavelet coefficients) receive lower subjective quality evaluations than images corrupted with uncorrelated distortions (e.g., additive Gaussian noise).\(^{15}\) That study and another study by one of the authors support our claim and conclude that the principle of global precedence accounts for this phenomena.\(^{14,15,20}\) Hence, QA algorithms consistent with subjective evaluations of visual quality for a variety of image artifacts are also likely to predict image recognition thresholds.

### 5.1 Conventional Distortion Measures

Mean-square error (MSE), which is used to compute peak signal-to-noise ratio (PSNR), and root-mean squared (RMS) distortion contrast provide computationally simple evaluations of signal quality. These measures evaluate quality solely in terms of the energy of the distortions. Root-mean-squared (RMS) distortion contrast evaluates the visibility of the distortions \( E = \hat{X} - X \) when comparing the images on a particular display device\(^{19}\) and is given by

\[
C_{\text{rms}}(E) = \frac{1}{\mu_L(X)} \sqrt{\frac{1}{M} \sum_{i=1}^{M} (L(E_i + \mu_X) - \mu_L(E+\mu_X))^2},
\]

where \( \mu_L(X) \) denotes the average luminance of the reference image \( X \), \( L(E_i + \mu_X) \) denotes the luminance of the \( i^{th} \) pixel of \( E + \mu_X \), \( \mu_L(E+\mu_X) \) denotes the average luminance of the mean shifted distortions \( E + \mu_X \), and \( M \) is the total number of pixels. Eq. (3) normalizes the standard deviation of the luminance values \( E + \mu_X \) according to the mean luminance of \( X \). This normalization accounts for Weber’s Law, which asserts that distortions of equal energy are more difficult to detect in brighter regions of an image than in darker image regions.

### 5.2 Algorithms Based on Psychophysical Properties of the HVS

Several quality assessment algorithms capitalize on models and principles characterizing low-level HVS properties such as contrast sensitivity,\(^{27}\) contrast masking,\(^{27-29}\) and perceived contrast.\(^{30,31}\) These properties model the detection of a visual target (e.g., the distortions in a distorted image) under a variety of conditions based on the contrast of the distortions.

The weighted SNR (WSNR) and noise quality measure (NQM) QA algorithms evaluate visual quality by incorporating HVS properties to first simulate the appearance of the reference and test images to a human, and then, the SNR is computed based on the difference of the simulated images.\(^{8}\) WSNR generates the simulated images through filtering with the contrast sensitivity function (CSF).\(^{27}\) NQM produces the simulated images through nonlinear processing based on Peli’s contrast pyramid.\(^{32}\) NQM’s processing model accounts for the HVS properties of contrast sensitivity, contrast masking, and suprathreshold contrast perception.

The visual signal-to-noise ratio (VSNR)\(^{13}\) quality assessment algorithm evaluates visual quality according to a contrast model accounting for low-level HVS properties and the mid-level HVS property of global precedence.\(^{20}\) VSNR incorporates models\(^{29}\) for low-level HVS properties based on experiments investigating the contrast of

---

Presented at SPIE Human Vision and Electronic Imaging XIII, San Jose, CA 2008
wavelet subband quantization distortions in natural images. To evaluate visual quality, VSNR first assesses the visibility of the distortions. For subthreshold distortions, the algorithm evaluates the test image as having perfect visual quality. For suprathreshold distortions, the VSNR visual quality evaluation accounts for the HVS properties of perceived contrast and global precedence.\textsuperscript{14,20}

5.3 Algorithms Derived from Hypothetical High-Level HVS Objectives

A family of QA algorithms has been developed based on the premise that the HVS has evolved in response to the statistical regularities exhibited by the physical world, and, thus, evaluates visual quality according to the resemblance of a test image’s structural information to that of the reference image. This section first summarizes each of these QA algorithms. Second, this section presents the mathematical specification of SSIM and MS-SSIM. This section concludes with a discussion and specification of the proposed modifications to MS-SSIM, denoted as MS-SSIM*. 

5.3.1 Summary of SSIM, MS-SSIM, IFC, and VIF

The structural similarity (SSIM)\textsuperscript{10} quality assessment algorithm and its multi-scale extension (MS-SSIM)\textsuperscript{9} evaluate visual quality based on the premise that the HVS has evolved to process structural information from natural images, and, hence, a high-quality image is one whose structure closely matches that of the reference image. To this end, SSIM employs a modified measure of spatial correlation between the pixels of the reference and test images to quantify the degradation of an image’s structure. MS-SSIM extends SSIM by evaluating this modified spatial correlation measure across several image scales.

Sophisticated models characterizing natural scene statistics (NSS)\textsuperscript{3} inspired two closely related quality assessment algorithms: information fidelity criterion (IFC)\textsuperscript{11} and visual information fidelity (VIF).\textsuperscript{12} Both use an information-theoretic framework that models the test image as the consequence of passing the reference image through distortion channels. These algorithms quantify the visual quality based on a measurement of the mutual information between the test image and the reference image.

For this class of QA algorithms, our analysis will be restricted to the structural similarity (SSIM) metric and its multi-scale extension (MS-SSIM), leaving the analysis of the information fidelity criterion (IFC) and visual information fidelity (VIF) for future work.

5.3.2 Mathematical Specification of SSIM and MS-SSIM

SSIM quantifies visual quality with a similarity measure between two patches $x$ and $y$ as the product of three components: mean $m(x, y)$, variance $v(x, y)$, and cross-correlation $r(x, y)$. The two patches, $x$ and $y$, correspond to the same spatial window of the images $X$ and $Y$, respectively. The SSIM value for the patches $x$ and $y$ is given as

$$\text{SSIM}(x, y) = m(x, y)^\alpha \times v(x, y)^\beta \times r(x, y)^\gamma = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right)^\alpha \times \left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right)^\beta \times \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right)^\gamma, \quad (4)$$

where $\mu_x$ denotes the mean of $x$, $\sigma_x$ denotes the standard deviation of $x$, $\sigma_{xy}$ is the cross-correlation (inner product) of the mean shifted images $x - \mu_x$ and $y - \mu_y$, and the $C_i$ for $i = 1, 2, 3$ are small positive constants. The positive exponents $\alpha, \beta,$ and $\gamma$ allow adjustments to the respective component’s contribution to the overall SSIM value. The overall SSIM image quality index, $\text{SSIM}(X, Y)$, for the images $X$ and $Y$ is computed by pooling the SSIM values computed for small patches of the two images. Typically, the SSIM value is computed with $\alpha = \beta = \gamma = 1$ and after downsampling the images $X$ and $Y$ by two in both spatial directions.\textsuperscript{10}

MS-SSIM extends SSIM by computing the variance and cross-correlation components at $K$ image scales, where the $k^{th}$ scale image corresponds to low-pass filtering and subsampling, by a factor of two in both spatial directions, the original image $(k - 1)$ times. The mean component is only computed at the coarsest scale, $K$. The MS-SSIM index is given by

$$\text{MS-SSIM}(X, Y) = m_K(X, Y)^\alpha_K \prod_{k=1}^K v_k(X, Y)^\beta_k r_k(X, Y)^\gamma_k, \quad (5)$$

Presented at SPIE Human Vision and Electronic Imaging XIII, San Jose, CA 2008
where \( m_k(X, Y) \), \( v_k(X, Y) \), and \( r_k(X, Y) \) respectively correspond to the mean, variance, and cross-correlation component computed and pooled over the image patches from scale \( k \) with \( k = 1 \) as the full-resolution image. The exponents \( \alpha_k \), \( \{\beta_k\}_{k=1}^{K} \), and \( \{\gamma_k\}_{k=1}^{K} \) vary according to \( k \) and adjust the contribution of the components based on experimental results by Wang et al.\(^9\) that examined perceptual quality across image scales for distortions with equal MSE. Thus, the exponents weight the contribution of the objective quality evaluations at each image scale to the overall MS-SSIM index. The exponents are nonnegative and normalized to sum-to-one across scale (i.e. \( \sum_{k=1}^{K} \beta_k = 1 \)). The exponents obtained from the experiment by Wang et al.\(^9\) are \( \beta_1 = 0.0448, \beta_2 = 0.2856, \beta_3 = 0.3001, \beta_4 = 0.2363, \) and \( \beta_5 = 0.1333 \) with \( \beta_k = \gamma_k = \alpha_k \) for \( k = 1, 2, \cdots, K \).

### 5.3.3 MS-SSIM* Specification

The constants \( C_i \), for \( i = 1, 2, 3 \), in Eq. (4) were introduced to counteract stability issues when either \( (\mu_x^2 + \mu_y^2) \), \( (\sigma_x^2 + \sigma_y^2) \), or \( (\sigma_x \sigma_y) \) are very close to zero.\(^\oplus\) Our analysis of SSIM and MS-SSIM evaluates and compares the ability of the individual components (i.e., mean, variance, and cross-correlation) to predict recognition thresholds. As a result, when \( (\sigma_{xy} \ll C_3) \) and \( (\sigma_x \sigma_y \ll C_3) \), both the SSIM cross-correlation component and MS-SSIM cross-correlation component are approximately one. This behavior occurs in our present study, for example, when computing the SSIM cross-correlation component of the MS-SSIM cross-correlation component between the VSP representation for \( \tau = 2048 \) of the airplane image, \( Y \), (cf. Figure 2(e)) and the original, \( X \), (cf. Figure 1(a)), since the VSP representation of the airplane image for \( \tau = 2048 \) is a constant valued image (i.e., \( \sigma_y = 0 \)). This casts doubt upon the significance of the SSIM and MS-SSIM cross-correlation component values to accurately assess the structure of images. Thus, an alternative version of MS-SSIM, henceforth identified as MS-SSIM*, is proposed where the positive constants \( C_i \) in each component have been set to zero.

The component definitions for MS-SSIM* follow from straightforward consideration of the scenarios leading to the stability concerns address in the preceding paragraph. Suppose the constants \( C_i \) have been set to zero. When both patches \( x \) and \( y \) have average pixel values of zero, the mean component is set to one, since the patches have identical mean values. Thus, the alternative mean component definition is given by

\[
m^*(x, y) = \begin{cases} \frac{1}{m(x, y)} & \mu_x^2 + \mu_y^2 = 0, \\
\text{else} & \end{cases}
\]

for \( m(x, y) \) as defined in Eq. (4) with \( C_1 = 0 \). Similarly, when both patches have variance zero, the variance component is set to one, since the patches have identical variances. The alternative variance component is given by

\[
v^*(x, y) = \begin{cases} \frac{1}{v(x, y)} & \sigma_x^2 + \sigma_y^2 = 0, \\
\text{else} & \end{cases}
\]

for \( v(x, y) \) as defined in Eq. (4) with \( C_2 = 0 \). Now, suppose that \( \sigma_x > 0 \), and the patch \( y \) is constant. Then, the variance of the patch \( y \) is zero. Under this scenario, \( y \) does not correlate with \( x \), so the cross-correlation component must be set to zero. When both patches have equal variance and \( C_3 = 0 \), the cross-correlation component must be set to one. The alternative cross-correlation component is given as

\[
r^*(x, y) = \begin{cases} 0 & \sigma_x > \sigma_y = 0 \text{ or } \sigma_y > \sigma_x = 0, \\
1 & \sigma_x = \sigma_y = 0, \\
\text{else} & \end{cases}
\]

for \( r(x, y) \) as defined in Eq. (4) with \( C_3 = 0 \).

Combining Eqs. (6)-(8) and following the extension from SSIM to MS-SSIM, MS-SSIM* is given as

\[
\text{MS-SSIM}^*(X, Y) = m_k^*(X, Y)^{\alpha_K} \prod_{k=1}^{K} v_k^*(X, Y)^{\beta_k} r_k^*(X, Y)^{\gamma_k},
\]

where \( m_k^*(X, Y) \), \( v_k^*(X, Y) \), and \( r_k^*(X, Y) \) respectively correspond to the alternative mean, variance, and cross-correlation components each computed and pooled over the image patches from scale \( k \) with \( k = 1 \) as the full-resolution image. MS-SSIM* inherits the MS-SSIM values for the exponents: \( \alpha_K, \{\beta_k\}_{k=1}^{K}, \) and \( \{\gamma_k\}_{k=1}^{K} \).
Table 1. Calculated p-values for a two-sample t test for each natural image to determine if the average value from the quality assessment algorithms corresponding to the recognition threshold images differ between the two image representations tested. Values in boldface lie in the acceptance region for the null hypothesis (equal means) at the 5% significance level. 

An algorithm’s frequent acceptance the null hypothesis endorses the algorithm’s ability to predict image recognition thresholds. Acronyms based on the image descriptions in Figure 3 identify the nine natural images.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>A</th>
<th>B&amp;C</th>
<th>B</th>
<th>T</th>
<th>S</th>
<th>P</th>
<th>J-L</th>
<th>CB</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.002</td>
<td><strong>0.052</strong></td>
<td>0.014</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.032</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Crms(E)</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.038</td>
<td>0.024</td>
<td>0.014</td>
<td>0.002</td>
<td>&lt; 0.001</td>
<td>0.011</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>WSNR</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.002</td>
<td><strong>0.054</strong></td>
<td>0.013</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.029</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>NQM</td>
<td>0.003</td>
<td>&lt; 0.001</td>
<td>0.001</td>
<td>0.031</td>
<td>0.013</td>
<td>0.005</td>
<td>&lt; 0.001</td>
<td>0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>VSNR</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.003</td>
<td>0.025</td>
<td>0.002</td>
<td>0.004</td>
<td>&lt; 0.001</td>
<td>0.005</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SSIM cross-corr.</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.028</td>
<td>0.021</td>
<td>0.002</td>
<td>&lt; 0.001</td>
<td><strong>0.080</strong></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>MS-SSIM cross-corr.</td>
<td>0.001</td>
<td><strong>0.305</strong></td>
<td>&lt; 0.001</td>
<td><strong>0.548</strong></td>
<td><strong>0.208</strong></td>
<td><strong>0.341</strong></td>
<td>0.049</td>
<td><strong>0.835</strong></td>
<td><strong>0.088</strong></td>
</tr>
<tr>
<td>MS-SSIM* cross-corr.</td>
<td><strong>0.079</strong></td>
<td><strong>0.108</strong></td>
<td><strong>0.570</strong></td>
<td><strong>0.840</strong></td>
<td>0.025</td>
<td><strong>0.976</strong></td>
<td>0.017</td>
<td><strong>0.257</strong></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>IFC</td>
<td><strong>0.054</strong></td>
<td><strong>0.239</strong></td>
<td>0.029</td>
<td><strong>0.215</strong></td>
<td><strong>0.193</strong></td>
<td><strong>0.187</strong></td>
<td><strong>0.822</strong></td>
<td><strong>0.408</strong></td>
<td><strong>0.100</strong></td>
</tr>
<tr>
<td>VIF</td>
<td><strong>0.841</strong></td>
<td><strong>0.086</strong></td>
<td><strong>0.901</strong></td>
<td><strong>0.696</strong></td>
<td><strong>0.705</strong></td>
<td><strong>0.685</strong></td>
<td>0.004</td>
<td><strong>0.659</strong></td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

6. ANALYSIS AND DISCUSSION

This section first evaluates the capability of the QA algorithms discussed in Section 5 to predict the recognition thresholds of the signal-based (SB) and visual structure preserving (VSP) image representations. Second, this section discusses the relationships between the recognition threshold, the MS-SSIM cross-correlation component, the MS-SSIM* cross-correlation component, and image structure.

6.1 Recognition Threshold Prediction Capability of Quality Assessment Algorithms

For each natural image and image representation, the image corresponding to each observer’s recognition threshold was evaluated with respect to the reference natural image using all of the FR QA algorithms discussed in Section 5. The evaluations corresponding to both types of SB representations were analyzed using a two-sample t test to test the null hypothesis that the samples belong to distributions with equal means. The test concluded that the null hypothesis could not be rejected at the 5% significance level for all natural images tested except the image pianist. Based on this result, the observer responses for the DCQ-J2K and MSE-J2K versions of the SB representations were pooled together for the following analysis.

A second hypothesis test identifies QA algorithms whose objective evaluations yield equal values for the average recognition threshold images for the SB and VSP representations. Specifically, a two-sample t test was performed for each natural image to test the null hypothesis that the QA values corresponding to the recognition threshold images for each image representation belong to distributions with equal means. For a particular QA algorithm, acceptance of the null hypothesis for most of the images tested endorses the ability of the QA algorithm to predict a natural image’s recognition threshold. The test is conducted at the 5% significance level. The p-value corresponding to a test statistic indicates the minimum value for which the null hypothesis is accepted and is provided in Table 1 for each combination of QA algorithm and natural image. Boldface values in Table 1 exceed 0.05 and lie within the acceptance region of the null hypothesis at the 5% significance level. Table 1 has been arranged according to the categorization of the QA algorithms established in Section 5.

The hypothesis tests indicate that the two conventional distortion measures, PSNR and RMS distortion contrast, fail to predict the recognition thresholds across the two image representations tested. For only the train image, the null hypothesis cannot be rejected with PSNR, yet for all of the images, the null hypothesis was rejected with RMS distortion contrast. Thus, Euclidean distance measures in both pixel space (PSNR) and luminance space (RMS distortion contrast) are inconsistent among the recognition threshold images for the SB and VSP image representations.

The QA algorithms based on the psychophysical properties of the HVS also fail to predict recognition thresholds for both image representations. The hypothesis test results for WSNR resemble those for PSNR despite
the incorporation of the CSF in the quality assessment. The null hypothesis is rejected for all nine natural images when using both NQM and VSNR. Thus, even the inclusion of both low-level and mid-level (i.e. global precedence) HVS properties in VSNR provide an insufficient analysis of characteristics from the reference image retained in recognition threshold images for both image representations.

Each of these QA algorithms based on HVS psychophysical properties adopts a contrast-based analysis of the distortions to evaluate quality. Their demonstrated failure to predict the recognition thresholds of both image representations devalues the significance of contrast in image interpretation. A previous study conducted by the authors identified recognition thresholds for a line drawing image representation.\textsuperscript{17} Line drawings were generated from the natural images with the Canny edge detector.\textsuperscript{33} Although this produced binary images, people clearly recognized the content from the line drawings.\textsuperscript{17} Traditional measures of contrast are meaningless for these binary images and cannot account for the image characteristics facilitating recognition for line drawings.

The multi-scale QA algorithms based on hypothetical high-level HVS objectives, perhaps indirectly incorporating high-level HVS processing, predict the recognition threshold of each image representation for most of the natural images tested. Specifically, the hypothesis tests conclude that the MS-SSIM cross-correlation component (specified by $R(X,Y) = \prod_{k=1}^{K} r_k(X,Y)$), the MS-SSIM* cross-correlation component (specified by $R^*(X,Y) = \prod_{k=1}^{K} r_k^*(X,Y)$), IFC, and VIF analyze characteristics that predict the recognition thresholds across both image representations. The insignificance of the coarsest LL subband (cf. Figure 2) demonstrated by the recognition task experiment results (cf. Section 4) dismisses the mean component of SSIM, MS-SSIM, and MS-SSIM* as a feasible predictor of the recognition threshold of both image representations. The hypothesis test also rejected the null hypothesis across eight natural images when only using the SSIM variance component. For the MS-SSIM cross-correlation component, the MS-SSIM* cross-correlation component, IFC, and VIF, the null hypothesis is accepted for most of the natural images tested. The consistent rejection of the null hypothesis with SSIM but frequent acceptance with MS-SSIM and MS-SSIM* emphasizes importance of an analysis across multiple image scales to predict a natural image’s recognition threshold.
6.2 Recognition Thresholds, Image Structure, and MS-SSIM* Cross-Correlation

Inspection of the MS-SSIM* cross-correlation component values emphasizes the relationship between an analysis of image structure and the recognition thresholds. The MS-SSIM and MS-SSIM* cross-correlation component values for each natural image are graphed in Figure 4. The range of the ordinate axes have been adjusted for both MS-SSIM and MS-SSIM* to emphasize their qualitative differences.

The MS-SSIM* cross-correlation component definition does not undermine the underlying principle of analyzing image structure for quality assessment, yet only the values of the MS-SSIM* cross-correlation component (cf. Figure 4(b)) exhibit similarities with the recognition threshold values corresponding to the image representation parameters (cf. Figure 3). Linear regression is performed to fit the MS-SSIM cross-correlation component values (i.e., Figure 4(a)) and the MS-SSIM* cross-correlation component values (i.e., Figure 4(b)) to the logarithm of the average recognition threshold parameters (i.e., Figure 3). The linear correlation between the MS-SSIM cross-correlation component values and the SB representation and VSP representation average recognition threshold parameters are 0.62 and 0.82, respectively. The linear correlation between the MS-SSIM* cross-correlation component values and the SB representation and VSP representation average recognition threshold parameters are 0.86 and 0.76, respectively. Overall, the evaluation of image structure by the MS-SSIM* cross-correlation component predicts the recognition thresholds of each image representation better than the MS-SSIM cross-correlation component, especially for the SB representations. This suggests that the recognition threshold image of each image representation retains common characteristics of the reference natural image attributed to visual structure (e.g., edges and object boundaries).

The hypothesis test results presented in Table 1 highlight the inadequacies of the single-scale analysis employed by the SSIM cross-correlation, and a closer inspection of the MS-SSIM* cross-correlation values at each scale stresses the relevance of a multi-scale image analysis. For example, the recognition thresholds for caged birds and boy & cat yield larger values with the MS-SSIM* cross-correlation component relative to the other natural images due to larger MS-SSIM* cross-correlation component values in finer scales. These images require additional, finer scale information for recognition than the other natural images and, again, illustrate the benefit of an multi-scale feature analysis.

Evidence from the experimental results (cf. Figure 3) and analysis indicates that the MS-SSIM* cross-correlation component provides a rudimentary measure of visual information. The linear relationship between the logarithm of the recognition threshold parameters (bitrate and $\tau$) and the MS-SSIM* cross-correlation component demonstrates that larger values of the MS-SSIM* cross-correlation component predict the presence of more visual information in the recognition threshold images. For example, the images caged birds and boy & cat had higher recognition thresholds than the remaining natural images, since adequate visual information for observer recognition was only available in images appearing later in the image sequences viewed. Furthermore, among the nine images tested, the written descriptions chosen to indicate recognition for these images are more specific: the phases caged birds and boy & cat conveys more explicit meaning than the single word pianist. The MS-SSIM* cross-correlation component values exhibit this variation in the amount of visual information necessary for recognition. Similar relationships were noted for both the IFC and VIF QA algorithms. However, an experiment should be conducted to study how human observers rate the relative information content of these natural images. Comparing the results of such an experiment would determine if the evaluations by the QA algorithms actually reflect the evaluation of visual information by human observers.

7. CONCLUSIONS

This work investigates the first step toward an objective visual information evaluation: predicting recognition thresholds for different image representations. Recognition threshold images of different image representations carry equivalent visual information. Two types of image representations were examined: signal-based and visual structure preserving. Signal-based (SB) representations add information as dictated by conventional mathematical characterizations (e.g., frequency content) of images. Visual structure preserving (VSP) representations add information to images features attributed to visual structure (e.g., object boundaries and edges). In addition to identifying recognition threshold images for the VSP, the results from the experiment demonstrate the insignificance of low-frequency image information in the recognition task.
Several full-reference perceptual quality assessment are evaluated in terms of their ability to predict the recognition threshold of different image representations. The proposed MS-SSIM* cross-correlation component exhibits a better overall correlation with the signal-based and visual structure preserving representations’ average recognition thresholds than the standard MS-SSIM cross-correlation component. This result endorses the image structure evaluation of the MS-SSIM* cross-correlation component as a predictor image recognition thresholds. These findings underscore the significance of visual structure in recognition and advocate a multi-scale image structure analysis for a rudimentary evaluation of visual information.

REFERENCES


Presented at SPIE Human Vision and Electronic Imaging XIII, San Jose, CA 2008


